



# Ontology Repository and Ontology-Based Services – Challenges, contributions and applications to biomedicine & agronomy

Clement Jonquet

## ► To cite this version:

Clement Jonquet. Ontology Repository and Ontology-Based Services – Challenges, contributions and applications to biomedicine & agronomy. Web. Université de Montpellier, 2019. tel-02133335

**HAL Id: tel-02133335**

**<https://theses.hal.science/tel-02133335>**

Submitted on 17 May 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# HABILITATION A DIRIGER DES RECHERCHES (HDR)

Spécialité Informatique  
École Doctorale Information, Structures, Systèmes

Université de Montpellier

## ONTOLOGY REPOSITORY AND ONTOLOGY-BASED SERVICES

Challenges, contributions and applications to  
biomedicine & agronomy

Manuscript v4.0 – May 2019

Clement Jonquet

(ORCID: 0000-0002-2404-1582)

Jury

(defense May 28th 2019)

Michel Dumontier	(professor),	Maastricht University	(reviewer)
Nathalie Aussenac-Gilles	(DR CNRS),	CNRS, Toulouse	(reviewer)
Mathieu D'Aquin	(professor),	National University of Ireland, Galway	(reviewer)
Fabien Gandon	(DR INRIA),	INRIA Sophia Antipolis	(examiner)
Juliette Dibie-Barthélemy	(professor),	AgroParisTech, Paris	(examiner)
Pascal Poncelet	(professor),	University of Montpellier	(examiner)
Mark A. Musen	(professor),	Stanford University	(invited)
Stefano A. Cerri	(prof. emeritus),	University of Montpellier	(invited)

Laboratory of Informatics, Robotics, and Microelectronics of Montpellier (LIRMM),  
University of Montpellier & CNRS, France





# Abstracts

## Abstract

With the explosion of the number of ontologies and vocabularies available in the semantic web, ontology libraries and repositories are mandatory to find and use them. Their functionalities span from simple ontology listing with metadata description to rich platforms offering various advanced ontology-based services: browse, search, visualization, metrics, annotation, recommendation, data access, etc. Studying ontology repositories opens then a wide spectrum of informatics research questions in areas such as knowledge representation, semantic web, data integration, natural language processing, ontology alignment and more. Ontology repositories are usually developed to address certain needs and communities. BioPortal, the ontology repository built by the US National Center for Biomedical Ontologies is the most important resource in biomedicine. It relies on a domain independent open technology that we have contributed to build (at Stanford) and extensively reused and extended for our research (at University of Montpellier) and applications to biomedicine and agronomy.

In this manuscript, we present and discuss six high level challenges for ontology repositories and services: (i) standardize and extend metadata used to describe ontologies and use these metadata to facilitate ontology evaluation, identification and selection; (ii) multilingualism, which requires rethinking ontology repositories to embrace (and encourage) the multilingual semantic web; (iii) all issues related to ontology alignment, not just the automatic generation of mappings, but also their extraction, storage, validation, etc., (iv) the design of better and new generic ontology-based methods especially for processing free text data, (v) the use of ontologies for semantic annotations & linked data; and finally, (vi) scalability & interoperability of the different semantic resources management platforms. For each challenge, we describe and point to results obtained in the context of our ontology repository projects over the last 12-years, especially the NCBO, SIFR, PractiKPharma and AgroPortal projects. We believe our results illustrate potential solutions to move forward in this domain of research.

## Keywords

Ontologies, ontology libraries & repositories, semantic web, ontology metadata, ontology services, ontology-based services, ontology selection, ontology alignment, ontology enrichment, terminology extraction, semantic annotation, semantic indexing, linked data.

## Résumé

L'explosion du nombre d'ontologies et de vocabulaires disponibles dans le web sémantique rend les portails d'ontologies obligatoires pour trouver et utiliser ces ressources. Leurs fonctionnalités vont de la simple liste d'ontologies décrites avec quelques métadonnées, à des plateformes riches et offrant divers services : navigation, recherche, visualisation, métriques, annotation, recommandation, accès aux données, etc. L'étude des portails d'ontologies ouvre ainsi un large spectre de questions de recherche en informatique dans des domaines tels que la représentation des connaissances, le web sémantique, l'intégration de données, le traitement du langage naturel, l'alignement d'ontologies, etc. Les portails d'ontologies sont généralement développés pour répondre à certains besoins et communautés. BioPortal, le portail d'ontologies construit par le *US National Center for Biomedical Ontology* est la ressource la plus importante en biomédecine. Il s'appuie sur une technologie ouverte indépendante du domaine que nous avons contribué à créer (à Stanford) et largement réutilisé et étendu pour nos recherches (à l'Université de Montpellier) dans nos applications en biomédecine et agronomie.

Dans ce manuscrit, nous présentons et discutons six grands défis pour les portails d'ontologies et les services qui y sont liés: (i) normaliser et étendre les métadonnées qui décrivent les ontologies et utiliser ces métadonnées pour faciliter l'évaluation, l'identification et la sélection des ontologies; (ii) le multilinguisme, qui nécessite de repenser les portails d'ontologies pour adopter (et encourager) le web sémantique multilingue; (iii) toutes les

questions liées à l'alignement des ontologies, pas seulement la génération automatique de *mappings*, mais aussi leur extraction, leur stockage, leur validation, etc. ; (iv) l'amélioration et la conception de nouvelles méthodes génériques basées sur des ontologies, en particulier pour le traitement des données textuelles ; (v) l'utilisation d'ontologies pour l'annotation sémantique et les données liées; et enfin (vi) le passage à l'échelle et l'interopérabilité des différentes plateformes de gestion de ressources sémantiques. Pour chaque défi, nous décrivons les résultats obtenus dans le cadre de nos projets sur les portails d'ontologies au cours des 12 dernières années, en particulier les projets NCBO, SIFR, PractiKPharma et AgroPortal. Ces résultats illustrent des solutions possibles pour ce vaste domaine de recherche.

### Mots clés

Ontologies, portail d'ontologies, web sémantique, métadonnées d'ontologies, services basés sur les ontologies et pour les ontologies, sélection d'ontologies, alignement d'ontologies, enrichissement d'ontologies, extraction terminologique, annotation sémantique, indexation sémantique, données liées.

# Contents

<b>Abstracts</b>	<b>3</b>
<b>Contents</b>	<b>5</b>
<b>Acknowledgments</b>	<b>7</b>
<b>Chapter I. Prelude</b>	<b>9</b>
I.1 Organization of the manuscript	9
I.2 Short biography	10
I.3 Research support and people involved	11
I.4 Collaborations	13
<b>Chapter II. Introduction</b>	<b>15</b>
<b>Chapter III. Background</b>	<b>19</b>
III.1 Ontology libraries and repositories	19
III.2 Focus on the NCBO BioPortal	21
III.2.1 BioPortal, a “one stop shop” for biomedical ontologies	21
III.2.2 Reuse of the NCBO technology	22
III.3 Two collaborative ontology repository projects	23
III.3.1 Semantic Indexing of French Biomedical Data Resources (SIFR)	23
III.3.2 AgroPortal: a vocabulary and ontology repository for agronomy	25
<b>Chapter IV. Challenges, propositions and results</b>	<b>29</b>
IV.1 Challenge 1: Metadata, evaluation and selection	29
IV.1.1 Harnessing the power of unified metadata in an ontology repository	30
IV.1.2 Metadata vocabulary for Ontology Description and Publication (MOD)	34
IV.1.3 NCBO Recommender: a biomedical ontology recommender web service	35
IV.2 Challenge 2: Multilingualism	37
IV.2.1 A roadmap for making BioPortal multilingual	38
IV.2.2 Multilingual mapping reconciliation between English-French biomedical terminologies	40
IV.3 Challenge 3: Ontology alignment	41
IV.3.1 What four million mappings can tell you about two hundred ontologies	43
IV.3.2 Enhancing ontology matching with background knowledge (A. Annane’s PhD project)	45
IV.3.3 Building a “Lingua Franca” in agri-food and biodiversity	49
IV.4 Challenge 4: Generic ontology-based services (especially for free text data)	50
IV.4.1 Semantic annotation of biomedical text with the NCBO Annotator	52
IV.4.2 SIFR Annotator: a publicly accessible ontology-based annotation tool to process French biomedical text	53
IV.4.3 Detecting negation, temporality and experienter in French clinical notes	56
IV.4.4 Annotating and indexing English clinical text with the NCBO Annotator+	56
IV.4.5 Terminology extraction and ontology enrichment (J-A. Lossio’s PhD project)	57

IV.4.6	MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums	61
<b>IV.5</b>	<b>Challenge 5: Annotations and linked data</b>	<b>62</b>
IV.5.1	NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources	63
IV.5.2	AgroLD, an RDF knowledge base for agronomy	66
IV.5.3	Pharmacogenomics Linked Open Data (PGxLOD)	66
IV.5.4	ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph (G. Surroca's PhD project)	68
<b>IV.6</b>	<b>Challenge 6: Scalability and interoperability</b>	<b>69</b>
<b>Chapter V.</b>	<b>Conclusion and Perspectives</b>	<b>71</b>
<b>V.1</b>	<b>Conclusion</b>	<b>71</b>
<b>V.2</b>	<b>Perspectives and research project</b>	<b>72</b>
<b>Chapter VI.</b>	<b>Curriculum Vitae</b>	<b>79</b>
	Contact & Professional Situation	79
	CV Strengths	79
	Work Experience	79
	Education	80
	Research Activity	80
	Professionnal Responsibilities	83
	Detailed Seminars & Invited Presentations	84
	Publications Summary	85
	Summary of Teaching Activities	86
	Technical Skills	86
	Personnal topics	87
	Languages	87
	Referees	87
<b>List of Publications</b>		<b>87</b>
	Journal	87
	International Conference	89
	Serie	90
	Workshop	91
	National (French) Conference	92
	Editor	93
	Dissertation	93
	Poster & Demonstration	93
	Report	95
	Under Review or in Progress	96
<b>(French) Details des activités d'enseignement</b>		<b>96</b>
	Expérience	96
	Récapitulatif	96
	Enseignements effectués	97
	Interventions diverses & encadrements	98
<b>Chapter VII.</b>	<b>Selected Publications</b>	<b>101</b>
	<b>Journal</b>	<b>101</b>
	<b>Conference</b>	<b>101</b>
<b>References</b>		<b>199</b>

# Acknowledgments

The work presented in this manuscript has mostly been realized in small groups. Therefore, I hereby deeply acknowledge all my collaborators, colleagues, students and partners for they help and willing to investigate these interesting scientific questions with me. This include of course my collaborators at LIRMM or in the Montpellier area, but also the ones at Stanford and in different research organizations in France and abroad.

I have some special thoughts for the people that I was honored to work closely with and (co)-supervise at LIRMM: Juan-Antonio Lossio Ventura, Vincent Emonet, Guillaume Surroca, Amina Annane, Anne Toulet, Andon Tchechmedjiev, Amine Abdaoui and Elcio Abrahao.

I especially kindly thank Pr. Stefano A. Cerri and Pr. Mark A. Musen for their guidance and advising roles respectively in Montpellier and Stanford. Both have been very supportive and encouraging since the very beginning of my career of researcher. They have always trusted me, supported me and encouraged me to take my decisions and fly with my own wings.

I would also like to express my deepest gratitude to all the members of the jury a who have done me the honor to review and evaluate the manuscript and participate in the defense.

Miscellaneous warm thanks also go to:

- Pr. Nigam H. Shah for his guidance, example and close partnership during my postdoc at Stanford. I learn a lot from him and I am still impressed by his ability to make things reachable.
- John Graybeal and the NCBO team at Stanford for their assistance and BioPortal related discussions (and more) during my mobility at Stanford.
- LIRMM and BMIR administrative staff who were always here when I would need them with the logistics of the researcher's life. Thank you, we could not proceed without your help.
- My colleagues at Polytech Montpellier with who I really enjoyed each part of the teaching activities and also whom allowed me the opportunities of focus more on research for a few years.
- My hierarchy at LIRMM and Polytech who trusted me and supported me in my research initiatives.
- Juan-Antonio Lossio Ventura who has invited me to give a keynote at SIMBig 2017 and then pushed me to write a short note that would eventually provide the pitch for this manuscript.
- Fabien Gandon for having me in his team at INRIA Sophia Antipolis during my detachment.
- University of Montpellier and CNRS for concrete miscellaneous support at LIRMM.
- The US National Park Service for preserving such great environments and landscapes (and campgrounds!) which refreshed my brain during the writing of this manuscript.

An acknowledgment section will not be complete without explicit references to the organizations and programs that have supported the research synthetized in this manuscript:



- My work during my postdoc at Stanford was supported by the *National Center for Biomedical Ontology*, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health.
- In Montpellier, my work was partly achieved within *the Semantic Indexing of French biomedical Resources* project that received funding mainly from the French National Research Agency's Young Researcher (JCJC) program 2012 (grant ANR-12-JS02-01001), the European Union's Horizon 2020 research and innovation

programme under the Marie Skłodowska-Curie grant agreement No 701771 as well as by the University of Montpellier and the CNRS.

- The *PractiKPharma* project was supported by French National Research Agency's generic research project program 2015 (grant ANR-15-CE23-0028).
- The *AgroPortal* project was explicitly supported in part by the NUMEV Labex (grant ANR-10-LABX-20), the Computational Biology Institute of Montpellier (grant ANR-11-BINF-0002), the Agro Labex (grant ANR-10-LABX-0001-01).

Finally, I would like to deeply thank my family for their trust and support along the way. To my wife, Isabelle, who is here at my side from the very beginning and without who I could not get the passion and energy. To my children, Antoine and Mathilde who bring so much light in our life.

# Chapter I.

## Prelude



Mount Rainier National Park

### I.1 Organization of the manuscript

This document is a synthesis of my research activities started during my postdoc at Stanford University in September 2007 and pursued in my current assistant professor position at University of Montpellier, since 2010. For 12 years, I have been working in the semantic web area [1–3], **designing, implementing, experimenting and evaluating scientific methods and technologies for ontologies and their use.**

This manuscript is inspired from a communication produced for a keynote presented during the *4<sup>th</sup> Symposium on Information Management and Big Data (SIMBig) 2017* [CJ47]. I hereby present my work taking as a common denominator “ontology repositories” as they represent the framework in which I have designed and experimented **ontology services (i.e., for ontologies) and ontology-based services (i.e., using ontologies)** within applications to biomedicine and agronomy. I will elicitate six different research challenges in this area and then present my work concerned with each of these challenges. Consequently, I do not present my work chronologically, but sub-domain by sub-domain.

This work has been accomplished in the context of several collaborative projects, described later, first as Stanford University then at University of Montpellier in **partnership with several teams, colleagues and students** that I explicitly acknowledge here.

The rest of this manuscript is organized as follows:

- **Chapter I. Prelude** (page 9). In the rest of this chapter, I provide a short biography for the reader to visualize my background and professional path. I also list here the different funding schemes and projects that have supported my research as well as the collaborations and people supervised.
- **Chapter II. Introduction** (page 15) announces the scientific content of this manuscript. I quickly introduce the overarching subject of this manuscript –ontology repositories and ontology-based services– then I propose six challenges for ontology repositories that will be the prism for presenting my research results of the last 12 years.
- **Chapter III. Background** (page 19) provides the necessary background information for appreciating the rest of the manuscript and the contributions. I define ontology libraries and repositories and survey the ones available today. I focus on the NCBO BioPortal ontology repository, a platform at the center of my research when I was postdoc at Stanford. Then, I will explain how we have reused this technology and continued our work in this area within the SIFR and AgroPortal projects which are the main contexts of my research at Montpellier. For both projects, I list the main results obtained (and provide references); each will be more extensively described in the rest of the manuscript.
- **Chapter IV. Challenges, propositions and results** (page 29) is the most significant chapter of this manuscript because it contains the scientific contributions. I will present how my work contributed (and still contributes) to address the challenges previously introduced. In each subsection, I quickly introduce my work done in this area, summarize the results, and point to the relevant publications detailing the contributions.
- **Chapter V. Conclusion and Perspectives** (page 71) concludes the manuscript and reviews my vision for ontology repositories and ontology-based services and future work in this domain of research.
- **Chapter VI. Curriculum Vitae** (page 79) is the first annex of this manuscript. It contains a detailed CV covering all my research activities, training, team & projects and a complete list and analysis of my publications. It also covers my teaching activities.

- **Chapter VII. Selected Publications** (page 101) is the second annex of this manuscript. It is a reproduction of a selection of important publications which content is partially reported in this manuscript. In the rest of the document, I cite my publications and communications using [CJ#] and list them in a specific section of the CV (page 79). Those must be distinguished from other references cited only by number and available in the **References section**, page 199. When a reference will be in bold e.g., **[CJ10]**, it means the publication is included in O. I also include work currently under preparation or review and in that case, it will be referred as [CJ-UR#] and listed at the end of the main list of publications.

## I.2 Short biography

I obtained a BSc, MSc, and **PhD in Informatics from University of Montpellier in Nov. 2006**. I had a French government PhD grant and was supervised by Pr. Stefano A. Cerri while working on “Dynamic Service Generation” with multi-agent systems, grid and service-oriented computing in the context of the European FP6-IST ELeGI project (002205). Then I served as a **postdoc for 3 years (2007-2010) at the Stanford Center for BioMedical Informatics Research (BMIR)** within Pr. Mark A. Musen’s group where I was working closely with Pr. Nigam H. Shah, on semantic annotations of biomedical data using biomedical ontologies in the context of the National Center for Biomedical Ontology (NCBO) project supported by the National Institutes of Health (U54-HG004028). I contributed actively to the design, evolution and development of the NCBO BioPortal, an ontology repository widely used in the biomedical informatics community. With the NCBO team, we won the 1<sup>st</sup> prize at the *International Semantic Web Conference*, Semantic Web Challenge 2010 with our ontology-based web application for searching and mining biomedical data.

Since September 2010, I am **Assistant Professor at University of Montpellier**, researcher at the Laboratory of Informatics, Robotics, and Microelectronics of Montpellier (LIRMM) and computer science teacher to the students of Ecole Polytechnique Universitaire de Montpellier. I have taught programming, computer architecture, web applications, semantic web to engineer students. From 2015 to 2018, I was back at Stanford BMIR as visiting scholar.

Since 2013, I am the **principal investigator of the SIFR project** (Semantic Indexing of French Biomedical Data Resources – [www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) interested in designing ontology-based services first for French biomedicine but also in agronomy. The project was mainly funded by the French National Research Agency (ANR) Young Researcher program (ANR-12-JS02-01001) and the H2020 Marie Skłodowska-Curie program (701771) which both supported my mobility to Stanford. I am also **co-PI of PractiKPharma project** (Practice-based evidences for actioning Knowledge in Pharmacogenomics – <http://practikpharma.loria.fr>) supported by ANR (ANR-15-CE23-0028). I lead the AgroPortal project (<http://agroportal.lirmm.fr>), a repository of ontologies and vocabularies for agronomy and related domains (agriculture, plant science, food and biodiversity); this project gathers several national research institutions (INRA, IRSTEA, CIRAD, IRD) but also involve important international organizations (FAO, CGIAR, H2020 eRosa, RDA Agrisemantics). From June 2019 I will be the **principal investigator of D2KAB project** (Data to Knowledge in Agronomy and Biodiversity – [www.d2kab.org](http://www.d2kab.org)) supported by ANR (ANR-18-CE23-0017). D2KAB’s objective is to create a framework to turn agronomy and biodiversity data into knowledge – semantically described, interoperable, actionable, open– and investigate scientific methods and tools to exploit this knowledge for applications in science & agriculture. D2KAB project brings together a unique multidisciplinary consortium of 12 partners to achieve this objective.

I am interested and have experience in several domains such as biomedical/agronomical informatics, ontologies and the semantic web, knowledge representation, data integration, semantic annotation, information retrieval and text mining, distributed systems, agents, service-oriented computing, collaborative systems, and web science. I am the (co)author of +80 publications cumulating more than 2200 citations, including 23 international journals in multiple domains (biomedical informatics, semantic web, distributed systems & AI), 6 as first author, 3 as last. Since 2010, I co-supervised 3 PhD students (J-A Lossio, G. Surroca, A. Annane) each time at 33% (with two other supervisors) and since 2007, 12 MSc students.

I was local chair of the 10<sup>th</sup> *European Semantic Web Conference* (ESWC) 2013 (350-person conference) and organizer of the *Semantics for Biodiversity Workshops* (S4BIODIV) 2013 & 2017. I am a member of several workshop and conference program committees related to life sciences and informatics including: ISWC (2017-2018), WWW 2012&2018, ESWC 2017, SWAT4LS (2015-2018), BioOntologies (2010-2017), WebSci 2012, ITS (2012&2014) MedEx (2010-2011). I reviewed articles for several journals such as Bioinformatics (Oxford Journals), BMC Bioinformatics (BioMed Central), Web Semantics (Elsevier), Biomedical Informatics (Elsevier), Biomedical Semantics (BioMed Central). I also chair the Web Science Montpellier and AgroHackathon Meetups.

As a teacher, I was also interested in the use of ICT in education, especially the use of iPads in the classroom. In 2014, I was program chair of a track at the French ICT in Education Conference and for 3 years I coordinated a

pedagogical innovation with ICT group at Polytech Montpellier. I am recipient of the French ministry distinction (PES) since 2013.

Figure 1 presents a summary of my professional timeline. Chapter VI contains my complete curriculum vitae and shall be consulted for complete enumeration and listings.

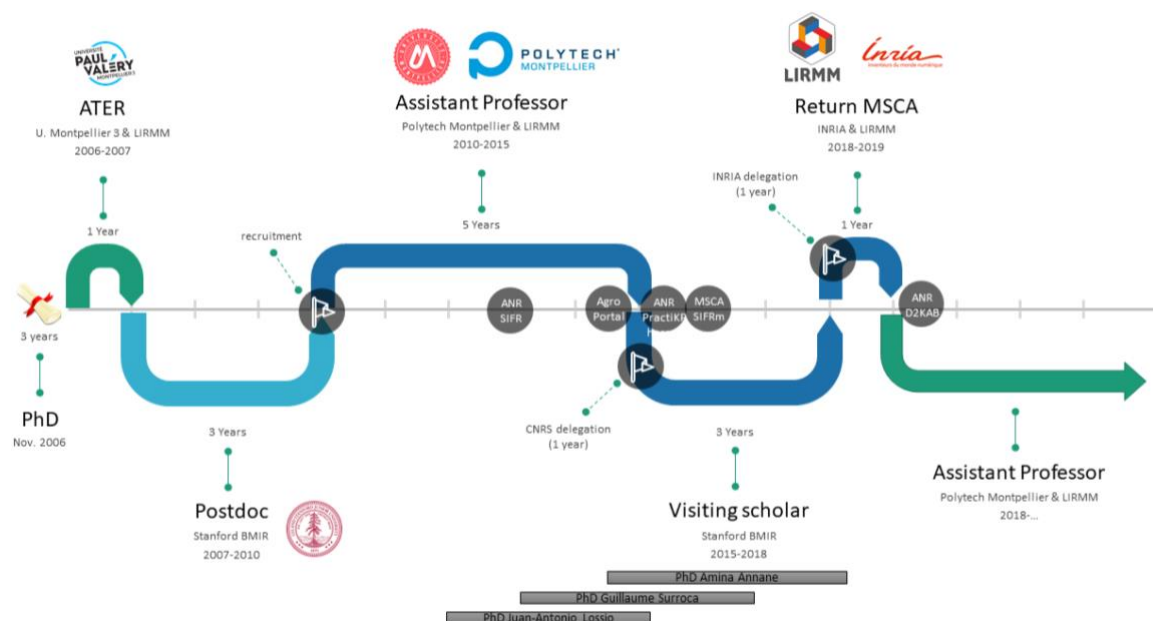


Figure 1. Professional timeline.

### 1.3 Research support and people involved

Since 2011, I have been independently supporting my research activities with different grants and funding sources; I obtained approximatively 2M€ of research funding as summarized in Table 1.

**Table 1. Funded grants (as leader).** The amounts are the support explicitly allocated to LIRMM; when relevant the total amount allocated (not total budget) of the project is detailed in parenthesis.

Project (#)	Program	Date	Amount (Total)	Type of support	Collaboration	Topic
<b>TUBO</b>	CNRS PICS	2011-2013	18K€	operating costs	Stanford BMIR, CHU Rouen	Ontology repository interoperability
<b>French GDR STIC-Santé collaborative actions</b>		2012	1K€	operating costs	CHU Rouen	
<b>Univ. Montpellier 2 scientific council PhD student grant</b>		2012-2015	90K€	PhD fellowship	UMR TETIS	Terminology extraction and ontology enrichment
<b>SIFR (ANR-12-JS02-01001)</b>	ANR JCJC call 2012	2013-2017	277K€	project	Stanford BMIR, CHU Rouen, UMR TETIS	Semantic indexing, ontology repositories, knowledge representation
<b>French CNRS, support for H2020 project preparation</b>		2014	3K€	operating costs	CIRAD	Application for EU project related to ViewpointS
<b>ANR IBC of Montpellier young researcher grant (ANR-11-BINF-0002)</b>		2014	10K€	operating costs		Complement for SIFR. Kick off of AgroPortal
<b>AgroPortal (ANR-10-LABX-20)</b>	Labex NUMEV call	2014-2015	46K€	1-year engineer	Multiple	Building first AgroPortal prototype
<b>PractiKPharma (ANR-15-CE23-0028)</b>	ANR generic call 2015	2015-2019	137K€ (677K€)	project	LORIA (Nancy), HEGP (Paris), CHU St Etienne	Electronic health records text mining and pharmacogenomics
<b>SIFR mobility (701771)</b>	H2020-MSCA-IF-2015	2016-2019	265K€	project	Stanford BMIR, INRIA (Zenith & Wimmics)	Support for mobility in the context of SIFR and AgroPortal
<b>e-Tera (partner of H2020 eRosa)</b>	ANR MRSEI	2016-2017	5K€ (24K€)	operating costs	INRA-DIST, IRD	Roadmap for e-infrastructure in agri-food

<b>AgroPortal (ANR-11-BINF-0002)</b>	IBC of Montpellier WP5	2016-2018	100K€	postdoc	Multiple	Community support and outreach for AgroPortal
<b>VisaTM / AgroPortal</b>	BSN-10	2017-2018	15K€ (160K€)	postdoc	INRA Versailles, CNRS-INIST	Text & data mining, semantic resources
<b>Lingua / AgroPortal (ANR-10-LABX-0001-01)</b>	NUMEV-Agro-CEMEB Interlabex	2017	90K€	postdoc	INRA Montpellier, CNRS-CEFE	Ontology mapping lifecycle in AgroPortal. Collaboration with GACS
<b>EUDAT Semantic Working group</b>		2018	6K€ (30K€)	postdoc	H2020 EUDAT, eScience Factory	Ontology portal interoperability
<b>H2020 OpenMinTed (OMTD) call for tender</b>		2018	15K€	postdoc	H2020 OpenMinTed	Text & data mining, semantic resources
<b>D2KAB (ANR-18-CE23-0017)</b>	ANR generic call 2018	2019-2023	(300K€) 950K€	project	INRIA (Wim-mics), INRA, IRSTEA, CNRS-CEFE, ACTA	Data to knowledge in agronomy and bio-diversity. AgroPortal, and linked data
<b>Joint Montpellier-Stanford Laboratory</b>	CNRS LIA	2019-2023	NA	operating cost	LIRMM, Stanford (3 teams)	Medical robotics, underwater robotics and Semantic Web

The work presented in this manuscript, when realized after my postdoc, has been done either by me directly or with someone (intern, PhD student, postdoc, engineer) under my supervision or co-supervision with other permanent colleagues as synthesized in Table 2. Since 2015, we work as a subgroup with regular meetings, exchanges and reporting. The skills and profiles of the subgroup members are always very different as our projects. We have investigated several research questions with students and postdocs and concretely experimented and transferred them into applications and prototypes with our engineers. Six team members came to Stanford during my mobility: V. Emonet, A. Annane, A. Tchechmedjiev, A. Abdaoui, C. Goehrs, S. Zevio.

**Table 2. Team supervision from 2012 to 2018.**

Person	Position	Date	Project	Support	Co-supervisors	Next situation
<b>Juan-Antonio Lossio Ventura</b>	PhD student	2012-2015	SIFR	UM2	M. Roche & M. Teisseire	Postdoc Univ. of Florida, USA
<b>Khedidja Bouarech</b>	MSc student	2013	SIFR	ANR		
<b>Guillaume Surroca</b>	PhD student	2013-2017	SIFR /ViewpointS	ANR	P. Lemoisson & S. Cerri	Industry (Agixis)
<b>Awa Dia</b>	Student eng.	2014	ViewpointS	Cirad	P. Lemoisson & G. Surroca	
<b>Soumia Melzi</b>	MSc student	2014	SIFR	ANR		
<b>Luc-Henri Méric</b>	Student eng.	2014	ViewpointS	Cirad	P. Lemoisson & G. Surroca	
<b>Vincent Emonet</b>	Engineer	2015-2017	SIFR	ANR		Ing. Univ of Maastricht, NL
<b>Anne Toulet</b>	Engineer	2016-2018	AgroPortal /VisaTM	NUMEV, IBC, BSN, OMTD		Ing. CIRAD (Montpellier)
<b>Amina Annane</b>	PhD student	2015-2018	SIFR /PractiKPharma	ESI Algeria & Eiffel prog.	Z. Bellashene & F. Azouaou	Postdoc IRIT (Toulouse)
<b>Solène Eholié</b>	MSc student	2016	SIFR	ANR	S. Bringay & M. Tapi-Nzali	Industry (Amadeus IT)
<b>Andon Tchechmedjiev</b>	Postdoc	2016-2018	PractiKPharma	ANR	S. Bringay	Adjunct Assit. Professor (Ales)
<b>Amine Abdaoui</b>	Postdoc	2016-2017	PractiKPharma	ANR	S. Bringay	Industry (Stack Labs)
<b>Clement Goehrs</b>	MD & MSc student	2017	PractiKPharma	personal		Industry (Synapse Med)
<b>Stella Zevio</b>	MSc student	2017	PractiKPharma	ANR	S. Bringay & A. Tchechmedjiev	Doctorat LIPN (Paris)
<b>Elcio Abrahao</b>	Postdoc	2018-2020	AgroPortal	Labex Agro	K. Todorov & P. Neveu	Industry (Brazil)

## I.4 Collaborations

I have always considered my research activities as collaborative ones. Either in the lab, locally in the Montpellier research ecosystem, nationally or internationally, I have always tried to reach out to other colleagues interested in similar topics and with who we can join our forces for a better scientific impact.

**At LIRMM:** A part from the contributions and exchanges related to the multi-agent systems & interaction research group “SMILE” (S. Cerri), I have strongly interacted and worked with members of other teams, especially the data and text mining “ADVANCE” team (S. Bringay, M. Roche, M. Teisseire); the ontology alignment & linked data “FADO” team (Z. Bellahsene, K. Todorov, F. Scharffe); the big data and scientific workflow “ZENITH” INRIA team (P. Valduriez, P. Larmande).

**Within the Montpellier research ecosystem:** During SIFR, I co-supervised 2 PhD students with researchers from UMR TETIS (M. Roche, M. Teisseire, P. Lemoisson). The AgroPortal initiative rapidly found an echo locally that have encouraged us to concretize the project in Montpellier. Since 2014, it brought me to collaborate with IRD (P. Larmande) on agronomic linked data (AgroLD project) and AgroPortal; CIRAD (M. Ruiz), Bioversity International (E. Arnaud), INRA (P. Neveu, P. Buche), CNRS-CEFE (E. Garnier) on several use cases for AgroPortal. More recently, I have started a partnership with ANR network #DigitAg and IRSTEA (V. Bellon-Maurel) for data interoperability projects on digital agriculture.

**Nationally:** The collaborations mainly happened during formal projects (cf. Table 1). Within SIFR, I started the project with the CISMeF group of CHU Rouen (S. Darmoni), and then had multiple interactions with other organizations such as INSERM LIMICS (J. Charlet), related to French biomedical ontologies, CHU Nancy (N. Girerd) on knowledge extraction from electronic health records. Within the ANR PractiKPharma consortium, led by LORIA (A. Coulet), I also collaborate with HEGP hospital (B. Rance & A. Burgun) and CHU St Etienne (C. Bousquet). AgroPortal found interests with other INRA research groups (C. Pichot, C. Nédellec, C. Pommier) but also with INRA Scientific and Technical Information department (S. Aubin, O. Hologne) with who I closely collaborate since 2016 on agri-food data interoperability within AgroPortal, VisaTM, eRosa/eTera projects and several working groups of the Research Data Alliance. Since 2017, both within VisaTM and the GDR SemanDiv, I exchange with CNRS-INIST (C. Francois, D. Vachez) on accessing and sharing semantic resources. At the national level, current or past industrial exchanges include: Sanofi (T. Pages), Ontologos (C. Million), Logixys (P. Dugénie), Mondeca (F. Amardeilh), eScience Data Factory (Y. Le Franc).

**Internationally:** See CV (Chapter VI) for before 2011. After being recruited at LIRMM, I have kept and extended the collaboration with Stanford BMIR (M. Musen), with the Protégé & NCBO groups. This collaboration has allowed us to develop an expertise in Montpellier on ontology services and repositories, the core of the work presented in this manuscript. In the last three years, the context of AgroPortal brought me to join and work with several international working groups such as RDA *Vocabulary and Semantic Services Interest Group* where I co-lead the ‘ontology metadata’ task group ([www.rd-alliance.org/groups/vocabulary-services-interest-group.html](http://www.rd-alliance.org/groups/vocabulary-services-interest-group.html)); the H2020 eRosa (e-infrastructure Roadmap for Open Science in Agriculture – [www.erosa.aginfra.eu](http://www.erosa.aginfra.eu)) project community (especially INRA, WUR, AgroKnow, FAO); the GACS project working group (J. Keyser) to design the Global Agricultural Concept Scheme; the AgBioData group ([www.agbiodata.org](http://www.agbiodata.org)) which gathers model organism databases in agriculture in the US; the RDA Agrisemantics working group (S. Aubin, C. Caracciolo – <http://agrisemantics.org>) and RDA *Wheat Data Interoperability* working group (E. Dzalé – [www.rd-alliance.org/groups/wheat-data-interoperability-wg.html](http://www.rd-alliance.org/groups/wheat-data-interoperability-wg.html)) as a use case for AgroPortal. We have also worked with the Food Agriculture Organization (V. Pesce) in the GODAN Action project on the design of the Agrisemantics Map of Data Standards (<http://vest.agrisemantics.org>). Since 2016, I have started a collaboration with Indian Statistical Institute (B. Dutta), on ontology metadata.



# Chapter II.

## Introduction



North Cascades National Park

A key aspect in addressing semantic interoperability is using ontologies as a common denominator to structure data, make them interoperable and turn them into structured and formalized knowledge. Ontologies formalize the knowledge of a domain by means of concepts, relations and rules that apply to that domain [4, 5]. When properly built, **ontologies allow representing data with clear semantics that can be leveraged by computing algorithms to search, query or reason on the data**. One way of using ontologies is by means of creating semantic annotations. An annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the concept [6, 7]. These annotations can then be used to build semantic indexes to leverage the knowledge inside the ontologies for better information mining and retrieval [8].

The semantic web produces many vocabularies and ontologies to represent and annotate any kind of data. In 2007 Swoogle's homepage [9] announced searching over 10000 ontologies. Today, a simple Google Search for "filetype:owl" returns around 34K results. How much ontologies are available online now? The big data deluge and the adoption of the semantic web technologies to describe and link these data [10] have made the **number of ontologies grow to numbers for which machines are mandatory to index, search and select them**. It has become cumbersome for domain experts to identify the ontologies to use so that automatic recommender systems have been designed to help them with this task, as for instance in the biomedical domain [CJ13, CJ8]. In addition of being spread out, ontologies are in different formats, of different size, with different structures and from overlapping domains. Therefore, with big number of ontologies new problems have raised such as describing, selecting, evaluating, trusting, and interconnecting them.

The scientific community has always been interested in designing common platforms to list and sometime host and serve ontologies, align them one another, and enable their (re)use [11–14]. These platforms range from simple ontology *listings*, rich *libraries* with structured metadata, to advanced **repositories (or portals) which feature a variety of services for multiple types of semantic resources** (ontologies, vocabularies, terminologies, taxonomies, thesaurus) such as browse/search, visualization, metrics, recommendation, or annotation.

In this manuscript, we will focus on ontology repositories, as they are the framework in which we have designed and experimented ontology-services (i.e., for ontologies) and ontology-based services (i.e., using ontologies). Ontology repositories allow to address important questions:

- If you have built an ontology, how do you let the world know and share it?
- How do you connect your ontology to the rest of the semantic world?
- If you need an ontology, where do you go to get it?
- How do you know whether an ontology is any good?
- If you have data to index or represent, how do you find the most appropriate ontology for your data?
- If you look for data, how may the semantics of ontologies help you locate them?

More generally, ontology repositories help "ontology users" to deal with ontologies without managing them or engaging in the complex and long process of developing them. As any other data, repositories help making ontologies FAIR (Findable, Accessible, Interoperable, and Re-usable) [15] as we will explained in Section IV.1.

From our experience working first on: (i) the US National Center for Biomedical Ontologies (NBCO) BioPortal, the most widely adopted biomedical ontology repository [CJ19][16]; (ii) the SIFR BioPortal, a specific local instance of BioPortal to address the French speaking biomedical community [CJ65]; and (iii) AgroPortal, an ontology repository for agronomy [CJ10]; we review and discuss six challenges in designing such platforms:

1. **Metadata, evaluation and selection.** Ultimately, ontology repositories are made to share and reuse ontologies. But which ontology should we reuse? With too many different and overlapping ontologies, properly describing them with metadata and facilitate their evaluation, identification and selection becomes an important issue [17]. We believe, as any other data, ontologies must be FAIR and although there are multiple dimensions to make ontologies FAIR, one will agree that developing open ontology repositories is one of them. Repositories are the best environment in which the metadata about ontologies can be described and valued. However, can we say that ontology developers describe their ontologies with relevant metadata properties that will facilitate manual or automatic search, identification and selection of ontologies? There exists a significant number of metadata vocabularies that could be used for ontologies but none of the existing ones can completely meet this need if taken independently. In this section, we will present our work on ontology metadata adopting first the perspective of designers of an ontology repository and report on our effort to develop a unified ontology metadata model for this repository; and second by presenting our effort in generalizing this work by collaboratively building a new shared specification for describing ontologies and semantic resources in general. In this section, we will also present our work in building automatic ontology recommendation algorithms and tools.
2. **Multilingualism.** We live in a multilingual world, so are the concepts and entities from this world. The semantic web offers now tools and standards to develop multilingual and lexically rich ontologies [18]. Recently, ontology localization, i.e., “the process of adapting an ontology to a concrete language and culture community” [19], has become very important in the ontology development lifecycle, but when efforts are made to properly represent lexical or multilingual information, it is rarely leveraged by ontology repositories. Repositories must be able to deal with multiple languages which means being able to deal with interface and content internationalization. While interface is easy, content internationalization is complex as semantic resources can be monolingual or multilingual and a repository must incorporate multilingual features at every level (search, mappings, annotation, etc.). To the best of our knowledge, there exist today no ontology repository fully multilingual. In this section, we will describe our choices and propositions made in 2014 to internationalize the NCBO BioPortal. Then, we will present some of our work in the context of the SIFR project, where we are building the SIFR BioPortal to host ontologies and terminologies with French labels. We will explain how we have addressed some of the requirements for multilingual ontology repositories – but not built one yet – especially with a new metadata model and by reconciliating multilingual ontology mappings between French medical terminologies in the SIFR BioPortal and their English counterparts in the NCBO BioPortal.
3. **Ontology alignment.** No conceptualization is an island. It is now commonly agreed data interoperability cannot be achieved by means of a single common ontology for a domain, and interlinking ontologies is the way forward. But the more ontologies are being produced, the more the need to identify mappings (correspondences) between different ontologies of the same domain becomes important. This process is known as ontology matching or ontology alignment. Building algorithms to identify these mappings, is itself a scientific challenge [20], but when dealing with ontology repositories, we also must address all the issues related to storing, retrieving, merging, scoring and evaluating these mappings so to create a valuable resource for the community addressed by the repository. In this section, we will present several of our researches in this area ranging from: (i) Analyzing, in 2009, what 4 million mappings in the NCBO BioPortal tell us about the ontologies themselves, the structure of the ontology repository, and the ways in which the mappings can help in the process of ontology design and evaluation; (ii) Creating new automatic ontology alignment algorithms and methods especially using existing ontology alignments as background knowledge; (iii) Making AgroPortal the reference platform for mapping extraction, generation, validation, evaluation, storage and retrieval by adopting a complete semantic web and linked open data approach and engaging the community.
4. **Generic ontology-based services (especially for free text data).** One reason to adopt semantic web standards and use ontology repositories is to benefit from multiple services for –and based on– ontologies. No one likes to reimplement something already existing and that can be generalized to another ontology just by dropping it in a repository. The portfolio of ontology-based services available in repositories should then grow. These services are ‘generic’ if they are domain independent i.e., not specific to a domain, group of ontologies, specific format or design principles. One important use of ontologies is for processing natural language: they can support multiple applications such as semantic annotation of free text, automatic translation or sentence generation, semantic search, terminology extraction and more. In this section, we will show how *ontologies can be used for text data* and present various results we have obtained in working

on semantic annotation of free text, first building in 2009 the NCBO Annotator, one of the most used ontology-based annotation web service in biomedicine and second by investigating similar questions but for French biomedical data within the SIFR project and exploring the challenges of dealing with clinical text. Then, we will show how *text data can be used for ontologies* by presenting our work in automatic terminology extraction and ontology enrichment.

5. **Annotations and linked data.** Ontologies and vocabularies are the backbone of semantically rich data (linked open data, knowledge bases, semantic indexes, etc.) as they are used to semantically annotate and interlink datasets[5, 7]. Besides the scientific interest of capturing and formalizing the knowledge of a domain, the main reason why ontologies are developed is for representing and semantically accessing data. The use of ontologies is established as one of the requirements for FAIR data [15]. Therefore, there exists a challenge in interconnecting ontology repositories with semantically rich data repositories and enable semantic search and data access directly from the repositories. In this section, we will present our efforts in building semantic knowledge bases and connecting them with an ontology repository. We will first present the NCBO Resource Index, a large-scale ontology-based index of more than 50 heterogeneous biomedical resources, integrated within the NCBO BioPortal. Then we will present other related work on using ontologies to build knowledge bases in agronomy or pharmacogenomics. Finally, as an alternative to ontologies, we will introduce an exploratory research designing a brain-inspired knowledge representation approach where semantic and social web contributions are merged into an adaptive knowledge graph which is then topologically, rather than logically, explored and assessed.
6. **Scalability & interoperability.** The community of ontology developers and users is growing both horizontally (i.e., new domains) and vertically (i.e., new adopters inside a domain). More and more ontologies are being developed and therefore, more and more ontology libraries and repositories are built. There is a challenge for ontology repositories to scale to high number of ontologies, while still addressing the five previously described challenges. In addition, when multiple repositories are created, they must be interoperable to ensure their users that they will not have to work with multiple web applications and programming interfaces if their ontologies of interest are not all hosted by the same repositories. In this section, we will present our vision to achieve this challenge and illustrate some steps forward. We strongly believe that sharing a common technology is the best way to make ontology repositories interoperable. This is the main motivation behind our reuse of the NCBO technology to build SIFR BioPortal and AgroPortal.

In the rest of the manuscript, we will detail these challenges and describe or point to results obtained in the context of our multiple ontology repository projects. In some sense, this manuscript is an index of 12-years of published research in the domain of ontology repositories and ontology-based services. For each challenge, we will not report or cover all related work published in the literature, but we will provide sufficient references to previous published work to appreciate and evaluate our contributions on each topic. We do not claim to have solved all the problems identified by these challenges, rather to have significantly contributed to potential solutions and progress in that domain of research.



# Chapter III.

## Background



Yellowstone National Park

### III.1 Ontology libraries and repositories

With the growing number of ontologies developed, ontology libraries and repositories have always been of interest in the semantic web community. Ding and Fensel (2001) introduced the notion of *ontology library* and presented a review of libraries at that time:

“A library system that offers various functions for managing, adapting and standardizing groups of ontologies. It should fulfill the needs for re-use of ontologies. In this sense, an ontology library system should be easily accessible and offer efficient support for re-using existing relevant ontologies and standardizing them based on upper-level ontologies and ontology representation languages.”

The terms “collection”, “listing” or “registry” are also used to describe ontology libraries. All correspond to systems that help reuse or find ontologies by simply listing them (e.g., DAML or DERI listings) or by offering structured metadata to describe them (e.g., FAIRSharing, BARTOC, Agrisemantics Map). But those systems do not support any services beyond description, especially based on the content of the ontologies.

Hartmann et al., (2009) introduced the **concept of ontology repository, with advanced features** such as search, browsing, metadata management, visualization, personalization, mappings and an application programming interface to query their content/services:

“A structured collection of ontologies (...) by using an Ontology Metadata Vocabulary. References and relations between ontologies and their modules build the semantic model of an ontology repository. Access to resources is realized through semantically-enabled interfaces applicable for humans and machines. Therefore, a repository provides a formal query language.”

By the end of the 2000's, the topic was of high interest as illustrated by the 2010 ORES workshop [21] or the 2008 Ontology Summit.<sup>1</sup> The Open Ontology Repository Initiative [22] was a collaborative effort to develop a federated infrastructure of ontology repositories. At that time, the effort already reused the NCBO BioPortal technology [23] that was the most advanced open source technology for managing ontologies but not yet packaged in an “virtual appliance” as it is today (cf. Section III.2.2). More recently the initiative studied OntoHub [24] technology for generalization but the Open Ontology Repository Initiative is now discontinued.

In parallel, there have been effort do index any semantic web data online (including ontologies) and offer search engines such as Swoogle and Watson [9, 25–27]. We cannot talk about ontology library or repositories for those “**semantic web indexes**”, even if they support some features of ontology repositories (e.g., search). Other similar products are terminology servers which are usually developed to host one or a few terminologies for a specific community (e.g., SNOMED-CT terminology server); they are usually not semantic web compliant and do not handle the complexity of ontologies.

In the biomedical or agronomic domains there are several standards and/or **ontology libraries** such as FAIRSharing (<http://fairsharing.org>) [28], the FAO & GODAN led Agrisemantics Map of Data Standards (<http://vest.agrisemantics.org>), and the agINFRA linked data vocabularies (<http://vocabularies.aginfra.eu>). They usually register ontologies and provide various metadata attributes about them. However, because they are registries not especially focused on vocabularies and ontologies, **they do not support the level of features that an ontology repository offers**. In the biomedical domain, the OBO Foundry [29] is a reference community effort to help the biomedical and biological communities build their ontologies with an enforcement of design and

<sup>1</sup> <http://ontolog.cim3.net/wiki/OntologySummit2008.html>

reuse principles that have made the effort very successful. The OBO Foundry web application (<http://obofoundry.org>) is not an ontology repository per se, but relies on other applications that pull their data from the foundry, such as the NCBO BioPortal [CJ19], OntoBee [30], the EBI Ontology Lookup Service [31] and more recently AberOWL [32]. In France, we can also mention HeTOP, the Health Terminology Ontology Portal [33] which supports several multilingual features (but mostly focus on French).<sup>2</sup>

Not specific to life sciences, there exist **other ontology libraries and repository efforts** such as the Linked Open Vocabularies [34], OntoHub [24], and the Marine Metadata Initiative's Ontology Registry and Repository [35]. More recently, we created the SIFR BioPortal [CJ65] prototype to build a French Annotator and experiment multilingual issues in BioPortal [CJ19]; and we developed AgroPortal, a vocabulary and ontology repository for agronomy and neighboring domains such as food, plant sciences and biodiversity [CJ10].

D'Aquin and Noy, (2012) and Naskar and Dutta, (2016) provided the latest reviews of ontology repositories. In Table 3, we provide a non-exhaustive –but quite rich– list of ontology libraries, repositories and web indexes available today. We only very partially address here the so called “**terminology servers**” which are similar platforms but deal with semantic resources less complex than ontologies. For examples, ANDS vocabulary service, NERC vocabulary server, SNOMED terminology server, Ortolang, Terraref, CLARIN vocabulary services, etc. These ad-hoc platforms mostly built independently of the semantic web effort are recently evolving to adopt SKOS (Simple Knowledge Organization System) [37], the W3C Recommendation for terminology and vocabulary. For instances, Finto or Loterre have adopted SKOSMOS [38] as backend technology; ANDS or NERC are SKOS compliant.

**Table 3. Non-exhaustive list of ontology libraries, repositories and web indexes available today.** We also included some known “technology” that can be “reused” to setup an ontology repository. Blue cells are projects in life sciences. The symbol \* identifies ontology repositories which reuse(d) NCBO technology.

Ontology libraries	Ontology repositories
OBO Foundry	NCBO BioPortal*
WebProtégé	Ontobee
Romulus	EBI Ontology Lookup Service
DAML ontology library	AberOWL
Colore	CISMEF HeTOP
Agrisemantics Map of Data Standards	SIFR BioPortal*
FAIRsharing	OKFN Linked Open Vocabularies
DERI Vocabularies	ONKI Ontology Library Service
OntologyDesignPatterns	MMI Ontology Registry and Repository*
SemanticWeb.org	ESIPportal*
W3C Good ontologies	AgroPortal*
TaxoBank	OntoHub
BARTOC	Finto
GFBio Terminology Service	EcoPortal (proposition end 2017)*
agINFRA Linked Data Vocabularies	Ontoserver
oeGOV	Loterre
Semantic Web indexes	Technology
Swoogle	NCBO Virtual Appliance (Stanford)
Watson	OLS technology (EBI)
Sindice	LexEVS (Mayo Clinic)
Falcons	Intelligent Topic Manager (Mondeca)
	SKOSMOS
Abandoned projects include: Cupboard, Knoodl, Schemapedia, SchemaWeb, OntoSelect, OntoSearch, OntoSearch2, TONES, SchemaCache, Soboleo	

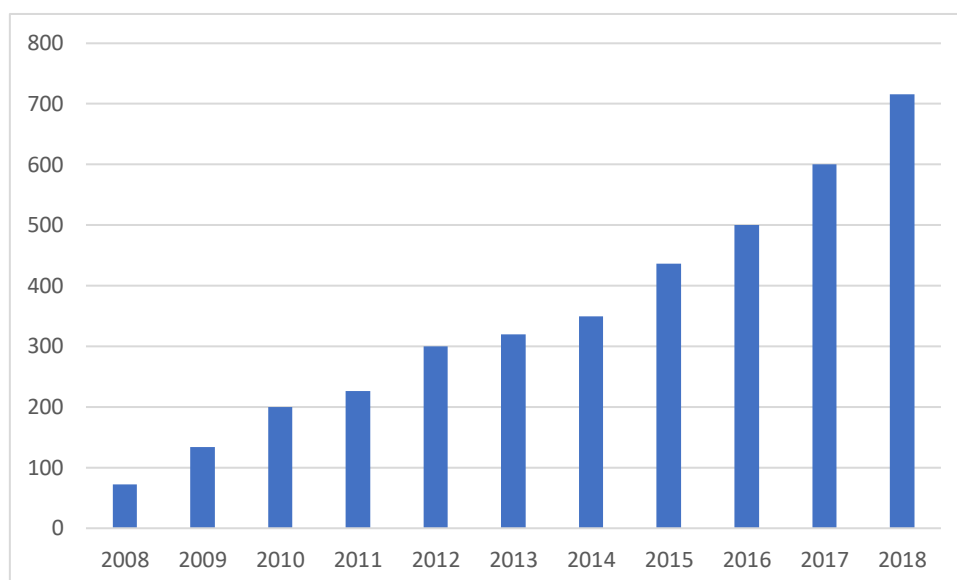
<sup>2</sup> For a comparison of the NCBO BioPortal and CISMeF HeTop, see our work done at the beginning of the SIFR project [CJ53].

## III.2 Focus on the NCBO BioPortal

### III.2.1 BioPortal, a “one stop shop” for biomedical ontologies

In the biomedical domain, BioPortal (<http://bioportal.bioontology.org>) [CJ19][16], developed by the National Center for Biomedical Ontologies (NCBO) at Stanford, is considered now as the reference open repository for biomedical ontologies originally spread out over the web and in different formats. There are around 770 public ontologies in this collection as of end 2018. **By using the portal’s features, users can browse, search, visualize and comment on ontologies both interactively through a web interface, and programmatically via web services.** The majority of BioPortal ontologies were contributed by their developers directly to the portal. A number of ontologies come from OBO Foundry [29], and BioPortal also includes publicly available terminologies from the Unified Medical Language System (UMLS) [39], a set of terminologies which are manually integrated and distributed by the US National Library of Medicine. BioPortal also includes ontologies that are developed in a variety of formats, including OBO file format, UMLS’s RRF format and of course the semantic web standards OWL, RDF(S), and more recently SKOS. Such ontologies include the NCI Thesaurus, Human Disease Ontology, the Medical Subject Headings (MeSH), the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), the Gene Ontology, the Foundational Model of Anatomy, and more.

NCBO BioPortal **has adopted from scratch semantic web technologies** e.g., ontologies, mappings, metadata, notes, and projects are stored in an RDF<sup>3</sup> triple store [40] and the functionalities have been progressively extended in the last 12 years. NCBO technology is now mature and quite robust considering the constant augmentation of the number of ontologies in the portal as illustrated by Figure 2.

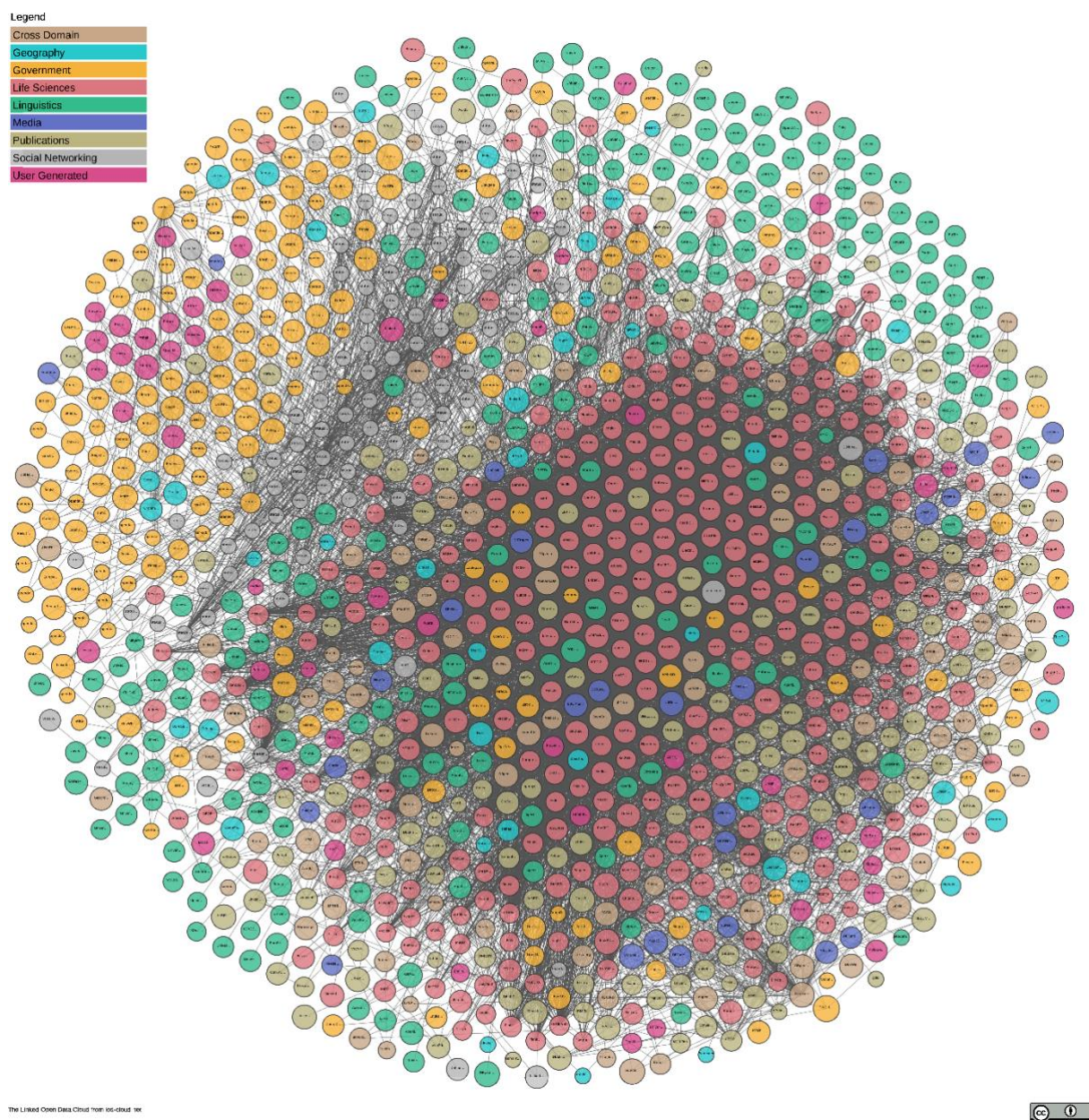


**Figure 2. Evolution of the number of ontologies in NCBO BioPortal over the last 10 years (source [41]).**

In the semantic web and linked open data world, the impact of BioPortal is easily illustrated by the famous linked open data cloud diagram (Figure 3, <http://lod-cloud.net>) that since 2017 includes ontologies imported from the NCBO BioPortal (most of the Life Sciences section).

Within BioPortal, ontologies are used to develop an annotation workflow, the NCBO Annotator [CJ41], that has been used inside the portal to build the NCBO Resource Index, a database of several biomedical data resources indexed using the knowledge formalized in ontologies to provide semantic search features and enhance information retrieval experience [CJ15]. Both applications will be later described respectively in Section IV.4 and IV.5.

<sup>3</sup> The Resource Description Framework (RDF) is the W3C language to described data. It is the backbone of the semantic web. SPARQL is the corresponding query language. By adopting RDF as the underlying format, an ontology repository based on NCBO technology can easily make its data available as linked open data and queryable through a public SPARQL endpoint.



**Figure 3.** The Linked Open Data cloud in early 2018 (source: <http://lod-cloud.net>). Many of the Life Sciences section in pink are from resources harvested and integrated inside NCBO BioPortal. Note that the diagram only contains NCBO BioPortal data as of 2013. Since then BioPortal has doubled in size.

### III.2.2 Reuse of the NCBO technology

An important aspect is that NCBO technology [23] is **domain-independent and open source**. A BioPortal virtual appliance<sup>4</sup> is available as a server machine embedding the complete code and deployment environment, allowing anyone to set up a local ontology repository and customize it. It is important to note that the NCBO virtual appliance has been quite regularly reused by organizations which needed to use services like the NCBO Annotator but, for privacy reason, had to process the data in house. Via the virtual appliance, NCBO technology has already been adopted for different ontology repositories in related domains and was also originally chosen as foundational software of the Open Ontology Repository Initiative [22]. The Marine Metadata Interoperability Ontology Registry and Repository [35] used it as its backend storage system for over 10 years, and the Earth Sciences Information Partnership earth and environmental semantic portal [42] was deployed several years ago. We are also currently working on the SIFR BioPortal and AgroPortal projects described next section.

<sup>4</sup> [www.bioontology.org/wiki/index.php/Category:NCBO\\_Virtual\\_Appliance](http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance)

In the context of our projects, to avoid building new ontology repositories from scratch, we have considered which of the technologies cited Section III.1 were reusable. While **most of them are “open source,” only the NCBO BioPortal<sup>5</sup> and OLS<sup>6</sup> are really meant for reuse**, both in their construction, and with their documentation provided. SKOSMOS is another alternative, but only support SKOS vocabularies. Although we cannot know all the applications of other ontology repository technologies, the visibly frequent reuse of the NCBO technology definitively confirmed it was the best candidate for our SIFR and AgroPortal projects. Also, of the two candidate technologies, we believe NCBO technology implements the highest number of required features in our projects (cf. [CJ10] for details). There are two other major motivations for reusing this technology: (i) to avoid re-developing tools that have already been designed and extensively used and instead contribute to long term support of a shared technology; and (ii) to offer the same tools, services and formats to biomedical (French & English) and agri-food communities, to facilitate the interface and interaction between their domains.

### III.3 Two collaborative ontology repository projects

In this section, we briefly introduce the SIFR and AgroPortal projects within which we design and develop two ontology repositories. We have worked during each project on various subjects (annotation, metadata, alignment, term extraction, etc.) but in most of the cases our research can be connected to the repositories. For each project, we list the main results obtained (and provide references); each of these results will be more extensively described in Chapter IV.

#### III.3.1 Semantic Indexing of French Biomedical Data Resources (SIFR)

##### III.3.1.1 *Scientific context, objective and partnership*

The volume of data in biomedicine is constantly increasing. Despite a large adoption of English in science, a significant quantity of these data uses the French language. Biomedical data integration and semantic interoperability are necessary to enable new scientific discoveries that could be made by merging different available data [43]. A key aspect to address those issues is the use of terminologies and ontologies to structure biomedical data and make them interoperable [44–46]. The community has turned toward ontologies to design semantic indexes of data that leverage the medical knowledge for better information mining and retrieval. However, besides the existence of various English tools, **there are considerably less ontologies available in French and there is a strong lack of related tools and services to exploit them** [47]. This lack does not match the huge amount of biomedical data produced in French, especially in the clinical world (e.g., electronic health records).

*SIFR  
project*

The *Semantic Indexing of French Biomedical Data Resources* (SIFR – [www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) project investigates the scientific and technical challenges in building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data. Our main goal is **to enable straightforward use of ontologies** freeing health researchers to deal with knowledge engineering issues and to concentrate on the biological and medical challenges; especially when exploiting ontologies for free text data. Indeed, researchers have called for the need of automated annotation methods and for leveraging natural language processing tools in the curation process. Still, even if the issue is being currently addressed for English, French is not in the same situation: there is little readily available technology (i.e., “off-the-shelf” technology) that allows the use of ontologies uniformly in various annotation and curation pipelines with minimal effort.

The SIFR project (ANR 2013-2017) originally brought together several young researchers at LIRMM to achieve this objective. Dr. Clement Jonquet coordinates the project. He was accompanied by two young researchers (HDR): Pr. Sandra Bringay and Dr. Mathieu Roche both expert in biomedical data/text mining. In addition, highly qualified and experienced partners are associated to the project: (i)°Stanford BMIR, a worldwide leader providing (English-)ontology-based services to assist health professionals and researchers in the use of ontologies to design biomedical knowledge-based systems; (ii)°The TETIS group, a joint applied research unit (AgroParisTech, IRSTEA, CIRAD) specialized in geographic information, environment and agriculture. ANR support for SIFR ended in August 2017, but the project continues until end 2019, supported by the European H2020 Marie Skłodowska-Curie program and the PraktikPharma project (ANR 2015-2019) in which we continue our work.

<sup>5</sup> The technology has always been open source, and the virtual appliance has been made available since 2011. However, the product became concretely and easily reusable after BioPortal v4.0 end of 2013.

<sup>6</sup> The technology has always been open source but some significant changes (e.g., the parsing of OWL) facilitating the reuse of the technology for other portals were done with OLS 3.0 released in december 2015.

### III.3.1.2 Methods

Within SIFR, we have developed the SIFR BioPortal (<http://bioportal.lirmm.fr>), **an open platform to host French biomedical ontologies and terminologies** based on the technology developed by the NCBO. The portal facilitates use and fostering of terminologies and ontologies which were only developed in French or translated from English resources and are not well served in the English-focused NCBO BioPortal. As of end 2018, the portal contains 28 public ontologies and terminologies (+ 6 private ones) that cover multiple areas of biomedicine, such as the French versions of standards terminologies (e.g., MeSH, MedDRA, ATC, ICD-10) but also multilingual ontologies such as Rare Human Disease Ontology, OntoPneumo or Ontology of Nuclear Toxicity. Ontologies have been offered by the CISMeF group from Rouen University Hospital, or taken from the UMLS, or directly uploaded by users. When ontologies are multilingual, we directly connect to the main NCBO BioPortal and only parse the French content –so users do not have to upload their multilingual ontologies twice. The SIFR BioPortal has been released in June 2015 and actively used and improved since then.<sup>7</sup>

The original motivation in building the SIFR BioPortal was to design an ontology-based indexing workflow and develop the SIFR Annotator (<http://bioportal.lirmm.fr/annotator>) to address the lack of out-of-the-shelf openly and easily accessible **semantic annotation system for French**. The service is originally based on the NCBO Annotator, a web service allowing scientists to utilize available biomedical ontologies for annotating their datasets automatically but was significantly enhanced and customized for French. The SIFR Annotator service processes raw textual descriptions, tags them with relevant biomedical ontology concepts and returns the annotations to the users in several formats such as JSON-LD, RDF or BRAT. In building the SIFR BioPortal and Annotator our vision was to embrace semantic web standards and promote openness and easy access.

Within the SIFR project, we have worked on several research questions from semantic indexing, text mining, terminology extraction, ontology enrichment, disambiguation, multilingualism in ontologies and semantic annotation in order to offer the community with services and applications capable of leveraging the use of biomedical ontologies in their data workflows. For instance, in order to **extract specialized terminology from free texts** in French, our approaches (J-A. Lossio's PhD) are based on new ranking functions that combine statistical and linguistic methods for highlighting relevant terms. Then we offer a complete methodology to identify (non)polysemic terms and choose the appropriate attachment in an already existing ontology. As another example, we developed a **new agent-centered graph-based knowledge representation approach**, called ViewpointS (G. Surroca's PhD), that enables to merge formal data representation (e.g., from the semantic web) with informal users' contributions (e.g., from the social web) and reveals relevant semantic paths between resources.

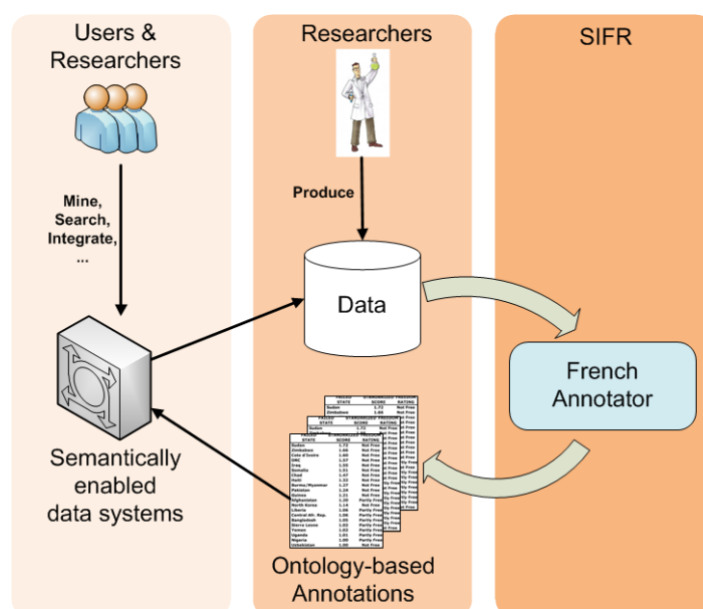


Figure 4. Building semantic data systems using annotated data within SIFR.

<sup>7</sup> [https://github.com/sifrproject/bioportal\\_web\\_ui/wiki/Release-notes](https://github.com/sifrproject/bioportal_web_ui/wiki/Release-notes)

### III.3.1.3 Main results

We provide a listing of SIFR project's main results. Each of them will be described in more detail in Chapter IV:

- 65 scientific publications and communications including: 13 international articles journal such as in *Information Retrieval*, *Bioinformatics*, *Web Semantics*. 29 international conferences (such as ISWC, IDEAS, MIE, KEOD, MEDINFO, EKAW). 3 PhD thesis. Full listing available here: <http://bit.ly/194ImnR>.
- We achieved an exhaustive comparison of CISMef HeTOP and NCBO BioPortal [CJ49], including the comparison of the annotation workflow and made HeTOP terminologies exportable in OWL.
- We deployed, customized and maintain an ontology repository for French biomedical ontologies/terminologies, the SIFR BioPortal (<http://biportal.lirmm.fr>) that hosts 28 terminologies and ontologies and offer multiple ontology-related services to the community [CJ65].
- We developed a French biomedical ontology repository including the SIFR/French Annotator (<http://biportal.lirmm.fr/annotator>). A service that for a given piece of text returns biomedical ontology concepts directly mentioned in the text or semantically expanded [CJ65][CJ2].
- We developed the BioTex methodology and tool (<http://tubo.lirmm.fr/biotex>) for automatic extraction of biomedical terms from plain text using existing extraction methods (e.g., C-Value) as well as keyword-based indexing methods (e.g., Okapi, Tf-Idf) [CJ13][CJ93].
- We developed a proxy web service for the NCBO Annotator ([http://biportal.lirmm.fr/ncbo\\_annotatorplus](http://biportal.lirmm.fr/ncbo_annotatorplus)) that offers, for English, access to new features that has been investigated and implemented within SIFR [CJ8].
- We reconciled 228K multilingual mappings between French and English biomedical ontologies/terminologies and stored them as linked data in the SIFR BioPortal [CJ29].
- We worked on automatic detection of emotion on public health forums using text mining techniques and we have built a patient vocabulary out of public patient-written online resources (<http://biportal.lirmm.fr/ontologies/MUEVO>) [CJ26].
- We conceived a semantic indexing and knowledge representation approach, called ViewpointS that captures formal data and informal contributions into an evolutionary knowledge graph [CJ11].
- Within PractiKPharma project (<http://pratikpharma.loria.fr>), we are enhancing the annotation workflow to capture clinical narrative and semantically annotate electronic health records from the G. Pompidou Hospital to extract pharmacogenomics knowledge [CJ2][CJ-UR2].

SIFR enabled the emergence of new research domain at LIRMM and materialized an important international collaboration with Stanford BMIR. SIFR offered the French speaking biomedical community (e.g., clinicians, health professionals, researchers) highly valuable ontology-based services that will enhance their data production and consumption workflows. In addition, the results of the project are not limited to French (also include English, Spanish) and we are also transferring our results in the agronomic domain by kicking-off the new AgroPortal project (<http://agroportal.lirmm.fr>) described hereafter.

SIFR's developments source code is open source and available on <https://github.com/sifrproject>.

## III.3.2 AgroPortal: a vocabulary and ontology repository for agronomy

### III.3.2.1 Scientific context, objective

Agronomy, food, plant sciences, and biodiversity are complementary scientific disciplines that benefit from integrating the data they generate into meaningful information and interoperable knowledge. Many **vocabularies and ontologies are produced to represent and annotate agronomic data**. For instances, the Plant Ontology [48], Crop Ontology [49], Environment Ontology [50], and more recently, the Agronomy Ontology [51], TOP Thesaurus [52], Food Ontology [53], Process and Observation Ontology [54], the IC-FOODS initiative's ontologies [55], and the animal traits ontology [56]. Semantic interoperability is a key issue for agronomy, and the use of ontologies a way to address it [57]. Similarly, resolving semantic heterogeneity has been identified as a key aspect to data integration, sharing and reuse for biodiversity and ecological sciences [58, 59]. Ontologies have opened the space to various types of semantic applications [60], to data integration [61], to process and transformation description [62] or decision support [63]. However, those ontologies are spread out over the web (or even unshared), in many different formats and types, of different size, with different structures and from overlapping domains. Therefore, there is **need for a common platform to receive and host them, align them, and enabling their use in agro-informatics applications**. There exists a need of a one-stop-shop for ontologies in the agronomy, food and biodiversity domains enabling to identify and select an ontology for a specific task as well as offering generic services to exploit them in search, annotation or other scientific data management



processes. For instance, plant genomics produces a large quantity of data (annotated genomes), and ontologies are used to build databases to facilitate cross-species comparisons e.g., [64]. Recently, it has been established that the scientific challenges in plant breeding have switched from genetics to phenotyping and that standard traits/phenotypes vocabularies are necessary to facilitate breeder's data integration and comparison [65]. The need is also for a community-oriented platform that will enable ontology developers and users to meet and discuss their respective opinions and wishes.

The *AgroPortal project*, is a community effort started by the Montpellier scientific community (LIRMM, IRD, CIRAD, INRA, Bioversity International) to build an ontology repository for agronomy and related domains. Our goal is to facilitate the adoption of **metadata and semantics to facilitate open science and the production of FAIR data**. By enabling straightforward use of ontologies, we expect data managers and researchers to focus on their tasks, without requiring them to deal with the complex engineering work needed for ontology management.

### III.3.2.2 Driving agronomic use cases

The AgroPortal project has been led from scratch by five driving agronomic use cases that participated in the design and orientation of the project to anchor it in the community:

- *Agronomic Linked Data* (AgroLD – <http://agrold.org>) project [CJ3] which develops methods for agronomic data integration and knowledge management within agronomic sciences to improve information accessibility and interoperability using semantic web and linked open data technologies. AgroLD is more extensively described in Section IV.5.2.
- *INRA Linked Open Vocabularies* (LovInra – <http://lovinra.inra.fr>) an initiative of the INRA's Scientific and Technical Information department to publish vocabularies produced or co-produced by INRA scientists and foster their reuse beyond the original researchers.
- *RDA Wheat Data Interoperability* (WDI) working group of the *Research Data Alliance* (RDA – <https://rd-alliance.org>) which goal is to provide a common framework for describing, representing, linking and publishing wheat data with respect to open standards [CJ9].
- The *Crop Ontology project* ([www.cropontology.org](http://www.cropontology.org)) of the Consultative Group on International Agricultural Research (CGIAR) and Biodiversity International have goals to publish online fully documented lists of breeding traits used for producing standard field books[49]; and to support data analysis and integration of genetic and phenotypic data through harmonized breeders' data annotation. The Crop Ontology contains 18 species-specific ontologies in addition to ontologies related to the crop germplasm domain.
- The *Agrisemantics Map of Data Standards* (<http://vest.agrisemantics.org>), that has been recently kicked off under the umbrella of the GODAN Action and Food and Agriculture Organization of the UN [CJ111].

### III.3.2.3 Methods and main results

Mid-2015, by reusing the NCBO BioPortal technology [CJ68][CJ88], we have designed AgroPortal (<http://agroportal.lirmm.fr>), an ontology repository for the agronomy domain but also food, plant, and biodiversity sciences (illustrated in Figure 5). AgroPortal [CJ10] offers **a robust and reliable advanced prototype that features ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation, stores and exploits ontology alignments**, and enables interoperation with the semantic web. The AgroPortal specifically satisfies requirements of the agronomy community in terms of ontology formats (e.g., SKOS vocabularies and trait dictionaries) and supported features (offering detailed metadata and advanced annotation capabilities).

AgroPortal version v1.4 was released in July 2017.<sup>8</sup> **The platform currently hosts 106 ontologies**, with more than 2/3 of them not present in any similar ontology repository (like NCBO BioPortal), and 7 private ontologies. We have identified 90 other candidate ontologies and we work daily to import new ones while involving/informing the original ontology developers. The platform already has more than 100 registered users and some vocabularies are visited more than 100 times per month.

In addition to its core repository of ontology mission, AgroPortal also offers many applicable tools, including a mapping repository, an annotator, an ontology recommender, and community support features. Our vision was to adopt, as the NCBO did, an **open and generic approach where users can easily participate** to the platform, upload content, and comment on others' content (ontologies, concepts, mappings, and projects).

---

<sup>8</sup> <https://github.com/agroportal/documentation/wiki/Release-notes>

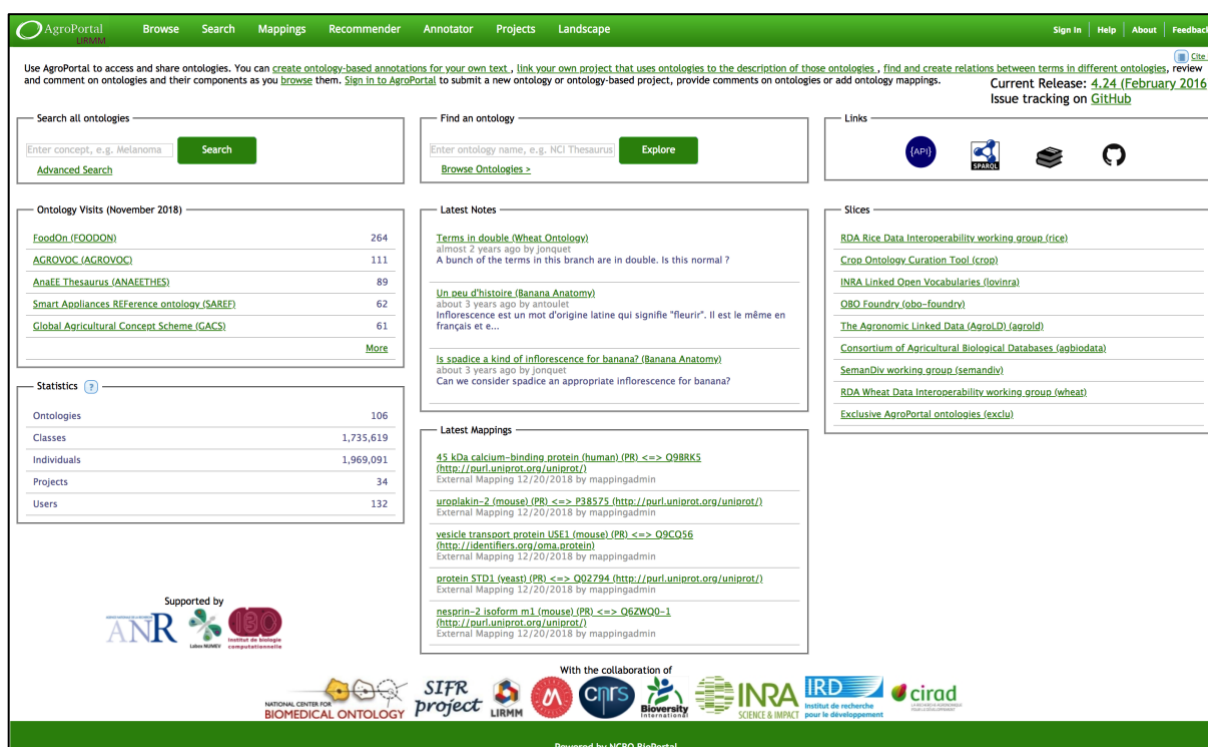


Figure 5. AgroPortal home page.

While working on the AgroPortal project we have implemented several new features and worked on guidelines and recommendations with several community groups, both described hereafter. Each item will be described in more detail in Chapter IV:

- New features implemented within the SIFR BioPortal have been made available also inside AgroPortal, including: multilingual ontologies practices, mapping related functionalities, semantic annotation scoring and contextualizing, new ontology formats. They are described in detail in [CJ10].
- We have implemented a new metadata model to better support descriptions of ontologies and their relations, respecting recent metadata specifications, vocabularies, and practices used in the semantic web community [CJ5]. The new model supports new functionalities for ontology identification and selection including facilitating the comprehension of the agronomical ontology landscape by displaying diagrams and charts about all the ontologies in the repository.
- We partnered with the AgroLD project to produce an RDF knowledge base of 100M triples created by annotating and integrating more than 50 datasets coming from 10 data sources with 10 ontologies [CJ3]. The knowledge base help solve complex biological and agronomical questions related to the implication of genes/proteins in, for instances, plant disease resistance or high yield traits (Section IV.5.2).
- We have participated to the RDA Wheat Data Interoperability working group recommendations [CJ9] which promote standards for most important data types –identified by the wheat research community (nucleotide sequence variants, genome annotations, phenotypes, germplasm data, gene expression experiments, and physical maps). For each of these data types, the guidelines recommend best practices in terms of use of data formats, metadata standards and ontologies.
- We partnered with the GODAN Action and FAO to build a broadly scoped global map of standards (i.e., library) in agri-food [CJ111]. To achieve this, we built on top of the existing VEST Registry, and added bidirectional mechanisms linking the new map with AgroPortal mainly to import ontologies and vocabularies from AgroPortal to the map.
- We have participated to the AgBioData consortium ([www.agbiodata.org](http://www.agbiodata.org)) recommendations for sustainable genomics and genetics databases for agriculture [CJ4]. AgBioData groups several agricultural biological databases, data archives and knowledge bases who strives to identify common issues in database development, curation, and management.
- We have joined the Global Agricultural Concept Scheme (GACS) project working group led by CABI, FAO, NAL and GODAN to participate in the establishment of a common hub of concepts for agri-food.

- Within the VisaTM project, we partnered with the H2020 OpenMinTed project which builds a shared infrastructure for text and data mining, to import any AgroPortal ontologies (or any NCBO-like ontology repository) in the OpenMinTed platform for use in text and data mining workflows [CJ62].
- We are developing ontology mapping capabilities to align AgroPortal ontologies in order to set up the bricks of a lingua franca for agronomy and biodiversity. We are investigating issues related to mapping extraction, generation, validation, evaluation, storage and retrieval focusing on some targeted ontology first and using GACS as a common hub. This work is still in progress.

With the experience acquired in the biomedical domain and building atop of an already existing technology, we think that AgroPortal offers a robust and stable reference repository that will become highly valuable for the agronomic, agriculture, food, plant sciences and biodiversity domains.

AgroPortal's developments source code is open source and available on <https://github.com/agroportal>.

# Chapter IV.

## Challenges, propositions and results



*Devils Tower National Monument*

In the following sections, we describe some challenges we have identified by working on ontology repositories and exchanging with different user communities. In each case, we describe results obtained on the topic and point to the relevant publications for more details.

### IV.1 Challenge 1: Metadata, evaluation and selection

The first questions we ask ourselves when entering a bookstore are often: “Where is the book I am looking for?” or “Which book will I discover and pick up today?” The same questions are true for ontology libraries. To address them:

**We need better description of the ontologies, with precise and harmonized metadata and we need means (including automatic ones) to facilitate evaluation, identification and selection of the ontologies of interest.**

Concerning metadata: As any resources, ontologies, thesaurus, vocabularies and terminologies need to be described with relevant metadata to facilitate their identification, selection and reuse. Metadata is now identified as a requirement to make the data FAIR [15]. But **as any other data, ontologies have themselves to be Findable, Accessible, Interoperable, and Re-usable**.<sup>9</sup> Although there are multiple dimensions to make ontologies FAIR, one will agree developing open ontology repositories and libraries is one of them. For ontologies to be FAIR, there is a need for metadata authoring guidelines and for harmonization of existing metadata vocabularies –taken independently none of them can completely describe an ontology. Ontology libraries and repositories also have to play an important role. Indeed, some metadata properties are intrinsic to the ontology (name, license, description); other information, such as community feedbacks, or relations to other ontologies are typically information that an ontology library shall capture, populate and consolidate to facilitate the processes of identifying and selecting the right ontology(ies) to use.

When someone is interested in an ontology, he/she may like to know: Who edited or contributed? When? What methodology or tool was used? Which natural language is used? Which formats are available? What are the metrics? Is it free to use or licensed? Who is using it? In addition, when someone is interested about ontologies of a domain, he/she may like to know: How ontologies can be grouped together? Which are most used? What are the relations between them? What are the common practices? Who are the key contributors of the domain? Or the most important organizations? All this information can be represented by metadata properties. Capturing

---

<sup>9</sup> This is also identified by the FAIR principle I2: “(meta)data use vocabularies that follow FAIR principles.”

that information is **both a technical challenge** –we need models, tools and automated population– **and a data curation challenge**. Indeed, the information or metadata about an ontology is often dispatched within websites, scientific articles, documentation or sometimes not existing at all, except in the brain of the original ontology developers. Clear guidelines are necessary on what to describe and how. For instance, the recent *Minimum Information for Reporting of an Ontology* initiative (<https://github.com/owlcs/miro>) [66] proposes the MIRO guidelines to ontology developers when reporting an ontology, e.g., in a scientific article.

In reviewing the current practices related to describing ontologies and using ontology metadata vocabularies, we have observed some limitations, lack of harmonization and confusions in the practices. This is not surprising when considering the efforts needed to just identify the potentially relevant vocabularies that could be used to describe ontologies.<sup>10</sup> Indeed, a few of these vocabularies are **dedicated to ontologies** (e.g., the *Ontology Metadata Vocabulary* (OMV) [67], the *Descriptive Ontology of Ontology Relations* (DOOR) [68], the *Vocabulary of a Friend* (VOAF) [69]), or **datasets** (e.g., the *Vocabulary of Interlinked Datasets* (VOID) [70], the *Data Catalog Vocabulary* (DCAT), or Schema.org) and others capture **more general metadata** (e.g., *Dublin Core* (DC) and *DCMI Metadata Terms* (DCT) [71], the *Provenance Ontology* (PROV), *Description of a Project* (DOAP)). They are often not maintained anymore, sometimes very specific or too general and of course, they are rarely aligned one another despite their significant overlaps. Furthermore, there have been several ontology repository projects that did not also took the problem seriously enough to support the description of their ontologies with standard vocabularies [36, 72]. With the exception of the Linked Open Vocabularies registry [34], the MMI Ontology Registry and Repository [35], and to some extent, the NCBO BioPortal [CJ19], **the question of harmonization and standardization of ontology descriptions have not really been a central matter**, although this is changing now (e.g., the OBO Foundry community metadata effort). The Linked Open Vocabularies was a good counter example; it has developed and adopted VOAF as a unified model to describe metadata and relations between vocabularies. Now, even if the metadata vocabulary is limited (16 properties), the platform has more than 600 resources described with the same model.

Concerning evaluation and selection: When available and properly harmonized, metadata facilitate the ontology evaluation, identification and selection processes, which has been assessed as crucial to enable ontology reuse [17, 66, 73, 74]. Ontology *evaluation* has been defined as the problem of **assessing a given ontology from the point of view of a particular criterion**, typically in order to determine which of several ontologies would best suit a particular purpose [75, 76]. *Identification and selection* (or recommendation) of an ontology are the processes of **choosing the right ontology for a given task** when searching for ontologies in an ontology library.

Early contributions in the field of ontology evaluation date back to the beginning of 90s and were motivated by the necessity of having evaluation strategies to guide and improve the ontology engineering process [77–79]. Some years later, with the birth of the semantic web, the need for reusing ontologies across the web motivated the development of the first ontology search engines [9, 25–27], which made it possible to retrieve all ontologies satisfying some basic requirements.

Ontology recommendation is fundamentally an ontology evaluation task because it addresses the problem of evaluating and consequently selecting the most appropriate ontologies for a specific context or goal [17, 80]. The process of recommending ontologies is a complex process that comprises not only enumerating a list of ontologies with class names matching a specific term, but also evaluating all candidate ontologies according to a variety of criteria, such as coverage, richness of the ontology structure [81–83], correctness, frequency of use [84], connectivity [81], formality, user ratings [85], and their suitability for the task at hand. In the following, we will present our work in building the NCBO Recommender [CJ12][CJ13].

#### IV.1.1 Harnessing the power of unified metadata in an ontology repository

In [CJ5], we adopt the perspective of designers of an ontology repository and report on our effort to develop a unified ontology metadata model for this repository. We measure the model impact on facilitating ontology descriptions, identification and selection. To do so, **we reviewed the current practices related to describing ontologies and using ontology metadata vocabularies**. This review was made to build a list of metadata properties that can be used to describe ontologies inside our own ontology repository. The objective of this work is not to propose another “vocabulary” for ontology metadata but to address the need of a common metadata

---

<sup>10</sup> Here, we consider the terms ontologies, terminologies, thesaurus and vocabularies as the type of knowledge organization systems [226] or knowledge artifacts [227]. Those are the subjects we are interested in describing. However, to facilitate the reading, we use the word *ontology* to identify the subject that is described by metadata (e.g., Movie Ontology, Human Disease Ontology, MeSH thesaurus, etc.) and the word *vocabulary* to identify the semantic resources used to described ontologies (e.g., OMV, DC, DCAT, etc.).

model inside an ontology repository i.e., implementing a way to compare ontologies side by side and describe the global landscape of all the ontologies in a library or repository. We have:

- Reviewed the most standard and relevant vocabularies (23 totals e.g., DC, VOID, OMV, DCAT, etc.) to describe metadata for ontologies. For each of these vocabularies, we have selected the significant properties to describe objects that an ontology could be considered a certain type of e.g., dataset, an asset, a project or a document. For instance, an ontology may be seen as a `prov:Entity` object and then the property `prov:wasGeneratedBy` may be used to describe its provenance.
- Reviewed the current use of metadata vocabularies by sampling 805 ontologies and measuring which vocabularies (and which properties in those vocabularies) are actually used by ontology developers.
- Studied some of the most common ontology repositories available in the semantic web community, and especially the NCBO BioPortal to capture in our list, the properties that were actually implemented by the repositories but that would represent an information not specific to the portal.

As the result, we obtained a list of 346 relevant properties to describe different aspects of ontologies that we have categorized for better understanding. Someone developing an ontology will of course not have to fill them all but can consider them as a list of candidate properties to use. We then grouped those properties into a **unified and simplified model of 127 properties** that includes the 46 properties originally offered by the NCBO BioPortal and reuses properties of the reviewed metadata vocabularies for the rest. We have implemented this new ontology metadata model within AgroPortal.

With a new edition interface and a common model available for all the ontologies in the repository, we have then spent a significant amount of time to edit and curate ourselves ontology descriptions, and we have asked the ontology developers to validate our edits and complete them. Now **all the ontologies within AgroPortal are described with the same unified metadata model**. This has resulted in three important new features summarized Table 4, including our capability to automatically aggregate information about ontologies to facilitate the comprehension of the whole agronomical ontology landscape by displaying diagrams, charts and networks about all the ontologies in the repository (grouping, types of ontologies, average metrics, most frequent licenses, languages or formats, leading contributors & organizations, most active ontologies, etc.) as illustrated in Figure 6 and Figure 7.

*Table 4. Summary of metadata use within AgroPortal ontology repository [CJ5].*

	Ontology Summary page	Browse Ontologies page	Landscape page
<b>Description</b>	Gives all the metadata information about a specific ontology.	Allows to search, order and select ontologies using a faceted search approach, based on the metadata.	Allows to explore the agronomical ontology landscape by automatically aggregating the metadata fields of each ontologies in explicit visualizations (charts, term clouds and graphs).
<b>New compared to BioPortal</b>	The whole “Additional Metadata” block which corresponds to properties from our new model. Plus the “Get my metadata back” buttons.	Four additional ways to filter ontologies in the list (content, natural language, formality level, type) as well as two new options to sort this list (name, released date).	This page did not exist in the original NCBO BioPortal.
<b>Example (user interface)</b>	<a href="http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE">http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE</a>	<a href="http://agroportal.lirmm.fr/ontologies">http://agroportal.lirmm.fr/ontologies</a>	<a href="http://agroportal.lirmm.fr/landscape">http://agroportal.lirmm.fr/landscape</a>
<b>Example (API call)</b>	<a href="http://data.agroportal.lirmm.fr/ontologies/ANAEETHES/submissions/2?display=all">http://data.agroportal.lirmm.fr/ontologies/ANAEETHES/submissions/2?display=all</a>	<a href="http://data.agroportal.lirmm.fr/ontologies">http://data.agroportal.lirmm.fr/ontologies</a>	E.g., to get omv:hasLicense property <a href="http://data.agroportal.lirmm.fr/submissions?display=hasLicense">http://data.agroportal.lirmm.fr/submissions?display=hasLicense</a>

An evaluation survey conducted with AgroPortal’s users shows evidence of the influence of ontology metadata on ontology identification and selection and reports on the very positive evaluation of the new functionalities by AgroPortal’s users. Thanks to this new unified model served by a stable application programming interface, metadata descriptions of AgroPortal ontologies have already been automatically harvested by two external ontology libraries: the Agrisemantics Map of Data Standards (<http://vest.agrisemantics.org>) [CJ111] and FAIRsharing (<http://fairsharing.org>).

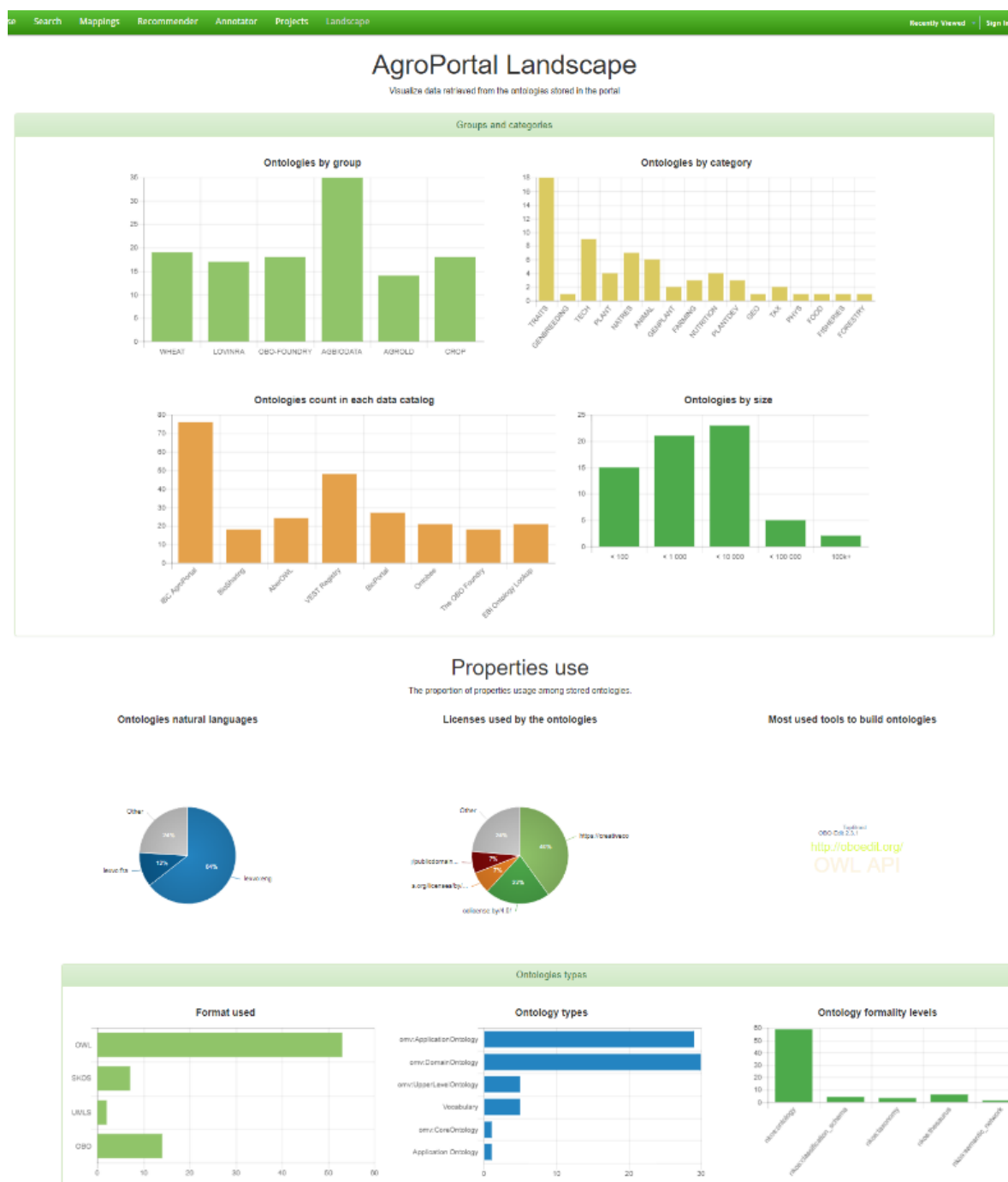


Figure 6. AgroPortal's new Landscape page: <http://agroportal.lirmm.fr/landscape> (part 1).

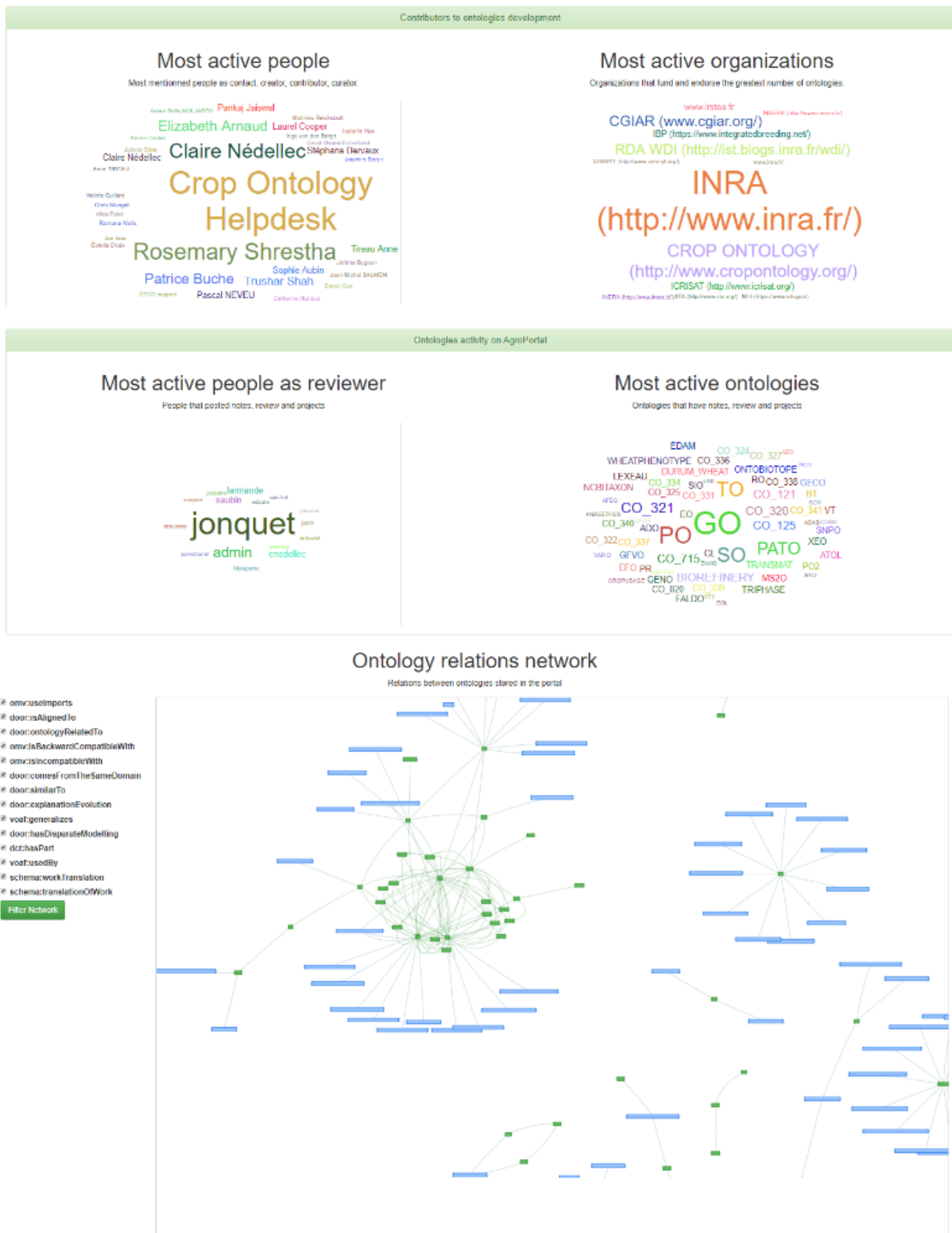


Figure 7. AgriPortal's new Landscape page: <http://agroportal.lirmm.fr/landscape> (part 2).

In [CJ5], we present how to harness the potential of a complete and unified metadata model with dedicated features in an ontology repository, however we did not pursue the goal of mixing all the reviewed vocabularies into a new “integrated vocabulary” that could become a standard for describing ontologies (e.g., a new OMV); although the clear need for metadata authoring guidelines and for harmonization of existing metadata vocabularies has been identified. A generalization of this work is studied in a community driven standardization effort presented hereafter.

### IV.1.2 Metadata vocabulary for Ontology Description and Publication (MOD)

The Research Data Alliance's Vocabulary and Semantic Services Interest Group (VSSIG)<sup>11</sup> seeks to develop community-based approaches and recommendations to make knowledge organization systems (i.e., controlled vocabularies, ontologies, and their associated services) FAIR. The VSSIG develops recommendations to address the needs of research communities and software developers for discovering and using multi-disciplinary controlled vocabularies and ontologies published on the web. In this context, several task groups were created, including the "Ontology metadata" one in which we (C. Jonquet, A. Toulet and B. Dutta) are involved as leaders. The work of this task group consists in **discussing and prototyping a new integrated ontology metadata standard** that can be used to describe any semantic resources and will be based mostly on previous metadata vocabularies.

In [CJ24], we have generalized our work done within AgroPortal and in collaboration with B. Dutta from the Indian Statistical Institute, Bangalore, we propose a new version of the *Metadata vocabulary for Ontology Description and publication*, called MOD 1.2 which succeeds previous work published in 2015 [72] and shall be a successor of OMV [67] which did not reuse any standard vocabularies at that time. The criteria for inclusion within MOD 1.2 of a property to describe its primary class `mod:Ontology`, were the following, considered by order of importance:

1. Relevance for describing an ontology –the property may have a sense if used to describe an ontology.
2. Semantic consistency –there must not be any conflict (e.g., disjoint classes) if someone would describe an ontology with all the listed properties. For instance, an ontology may be of type `omv:Ontology`, `foaf:Document`, `owl:Ontology`, `prov:Entity`.
3. Being included in a W3C or Dublin Core recommendation.
4. The frequency of use as found in studying ontology metadata vocabularies.
5. Priority to vocabularies specific for ontologies rather than to the ones specialized for the more general objects (`cc:Work`, `dcat:DataSet`, `sd:Service`, etc.).

MOD 1.2 is defined in OWL and consists of 19 classes and 88 properties most of them to describe the `mod:Ontology` classe.<sup>12</sup> Figure 8 provides a representation of the model in terms of its main classes, object & data properties. MOD 1.2 may serve as (i) a vocabulary –such as an application profile– to be used by **ontology developers to annotate and describe their ontologies**, or (ii) an explicit OWL ontology to be used by **ontology libraries to offer semantic descriptions of ontologies as linked data**.

Using the MOD 1.2 OWL model, we manually created a small knowledge base consisting of metadata about eight agronomical ontologies selected from AgroPortal<sup>13</sup> (AGROVOC, Gene Ontology, NAL Thesaurus, NCBI Organismal Classification, Protein ontology, AnaEE Thesaurus, IBP Crop Research Ontology, and Sequence Types and Features Ontology) which are very precisely described thanks to the work presented in the previous section. The knowledge base supports a variety of new queries, for instances: which is the most popular ontology editing tool? Who are the key contributors in a domain? How many ontologies are produced by OBO Foundry group? What are the projects using the Protein Ontology? *What are the ontologies endorsed by the RDA Wheat Data Interoperability WG and the National Science Foundation?* These queries were expressed in SPARQL and successfully run over the knowledge base. The above *italicized* query is shown below. It returns the title and the creator of the ontologies endorsed by RDA WDI and NSF. A couple of such sample SPARQL queries are also available on GitHub.

```
SELECT DISTINCT ?Ontology ?Author
WHERE {
  {?x a mod:Ontology; omv:endorsedBy <https://www.rd-alliance.org/groups/wheat-data-
  interoperability-wg.html> ; dct:title ?Ontology .}
  UNION
  {?x a mod:Ontology; omv:endorsedBy
  <http://dbpedia.org/resource/Category:National_Science_Foundation> ; dct:title ?Ontology .}
  OPTIONAL {?x dct:creator ?Author .} }
```

<sup>11</sup> <https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html>

<sup>12</sup> The OWL file and versions are publicly available (<https://github.com/sifrproject/MOD-Ontology>).

<sup>13</sup> The process was since then semi-automatized and we can now build MOD 1.2 knowledge bases with a significant number of instances.

Our future goals are: (i) to automatize the process of creating `mod:Ontology` instances using the application programming interfaces of the main ontology libraries (e.g., BioPortal, AgroPortal, OBO Foundry). This will enable to export the content of these libraries without doing any change to their internal data models; (ii) to release knowledge base as linked open data consisting of metadata for ontologies covering a significant amount of ontologies; and (iii) to offer a SPARQL endpoint to provide local and remote advanced queries on the knowledge base.

**Figure 8. A snapshot of MOD 1.2 [CJ24].** A complete diagram is available at <https://github.com/sifrproject/MOD-Ontology>.

#### IV.1.3 NCBO Recommender: a biomedical ontology recommender web service

- Re-use existing ontologies when constructing new ones;

- Identify the most appropriate ontology for a given domain;
- Support an annotation workflow.

Therefore, ontology recommendation has emerged as a key issue in biomedicine [84, 87–90]. Recommending biomedical ontologies is a challenging task. The **great number, size, and complexity of biomedical ontologies, as well as the diversity of user requirements and expectations, make it difficult to identify the most appropriate ontologies** to annotate biomedical data. The manner in which recommendation occurs depends on user settings. In some cases, the recommendation process can be long and non-automatic; the user can participate in the process (e.g., answer questions to refine the query) to enhance the accuracy of results. In other cases, a quick and fully automated approach is required, such as when ontology selection occurs at runtime in an application.

In [CJ13], we conceived and developed within BioPortal the *Biomedical Ontology Recommender web service*, later called simply the *NCBO Recommender*. The system provided a quick automated recommendation with minimal user burden. We considered two main recommendation scenarios differentiated by the type of input provided by the user:

- *Corpus-based recommendation*: Given a corpus of textual metadata describing some elements of a biomedical dataset, our system recommends appropriate ontologies to annotate the dataset with ontology concepts.
- *Keyword-based recommendation*: Given a set of keywords/terms representative of a domain of interest, our system recommends appropriate ontologies to consider for re-use or extension for researchers building new ontologies or semantic applications.

The service recommended based on three criteria. The first one was *coverage*, or the ontologies that provide most terms covering the input text. The second was *connectivity*, or the ontologies that are most often mapped to by other ontologies. The final criterion was *size*, or the smallest ontologies in number of concepts. The service scored the ontologies as a function of scores of the annotations created using the NCBO Annotator web service [CJ41] and relied on all the ontologies in BioPortal. The approach used both a syntactic concept recognition step (string matching with concept names & synonyms) and a mapping expansion step to enforce reference ontologies (expand annotations with mappings). In [CJ13], we evaluated and discussed recommendation results generated by different heuristics in the context of three real world use cases. Overall, evaluators agreed on the utility of the recommendations provided both for their keyword and corpus datasets.

To the best of our knowledge, the NCBO Recommender was the first biomedical ontology recommendation service, and it became widely known and adopted by the community.<sup>14</sup> However, the service had some limitations, and a significant amount of work has been done in the field of ontology recommendation since its release. This motivated us to analyze its weaknesses and to design a new recommendation approach.

In [CJ12], we have applied our previous experience in the development of the original NCBO Recommender and the BiOSS system [84] to conceive a new version of the NCBO Recommender (2.0). The new recommendation approach evaluates the relevance of an ontology to biomedical text data according to four different criteria: (1) *coverage*, or the extent to which the ontology covers the input data; (2) the *acceptance* of the ontology in the community; (3) the level of *detail* of the ontology classes that cover the input data; and (4) the *specialization* of the ontology to the domain of the input data. This new version of the service combines the strengths of its predecessor with a range of adjustments and new features that improve its reliability and usefulness. The user interface is illustrated in Figure 9.

In [CJ12], to evaluate our approach, we compared the performance of the NCBO Recommender 2.0 to the previous version of 2010 using data from a variety of well-known public biomedical databases (PubMed, the Gene Expression Omnibus (GEO) and ClinicalTrials.gov). We used the API provided by the NCBO Resource Index [CJ15] to programmatically extract data from those databases. Our evaluation shows NCBO Recommender 2.0 returns higher quality suggestions than the original approach, providing better coverage of the input data, more detailed information about their concepts, increased specialization for the domain of the input data, and greater acceptance and use in the community. In addition, it provides users with more explanatory information, along with suggestions of not only individual ontologies but also groups of ontologies to use together. It also can be customized to fit the needs of different ontology recommendation scenarios.

---

<sup>14</sup> In 2015, before the release of the new Recommender described hereafter, there were 57 citations to the NCBO Recommender paper [CJ17] and the service received 95 calls per month on average in 2013 (which is quite significant for a web service that is intrinsically made to be used only once per recommendation scenario need).

Input  
☐ Text ☒ Keywords (separated by commas)

Output  
☒ Ontologies ☐ Ontology sets

Insert sample input

headaches, anemia, abnormal behavior, irritability, floppy head, cellulitis, lameness, stiff neck, facial tremor, backache, abdominal pain, weight gain, congestion, sneezing, respiratory failure, vascular alteration, atrial fibrillation, sleepy, sweaty, tired, weak

advanced options

Edit Input

Recommended ontologies

POS.	ONTOLOGY	FINAL SCORE	COVERAGE SCORE	ACCEPTANCE SCORE	DETAIL SCORE	SPECIALIZATION SCORE	ANNOTATIONS	HIGHLIGHT ANNOTATIONS
1	SYMP	74.3	90.2	29.1	36.3	99.2	17	<input checked="" type="checkbox"/>
2	SNOMEDCT	68.9	67.9	95.3	59.2	55.9	16	<input type="checkbox"/>
3	NCIT	64.0	55.1	87.6	74.9	62.3	13	<input type="checkbox"/>
4	MEDDRA	60.3	64.2	96.5	28.5	41.3	14	<input type="checkbox"/>
5	MESH	52.5	38.7	88.2	92.6	27.4	9	<input type="checkbox"/>
6	RCD	51.5	50.2	86.7	34.5	38.0	10	<input type="checkbox"/>
7	CSSO	47.6	43.6	22.6	77.9	56.8	9	<input type="checkbox"/>

**Figure 9. Interface of the Recommender 2.0 in the NCBO BioPortal [CJ12].** For the text entered by the user (here a list of keywords), ontologies are ranked following a final score computed from different scores obtained for each recommendation criteria.

Our approach for ontology recommendation was designed for the biomedical field, but it can be adapted to work with ontologies from other domains so long as they have a resource equivalent to the NCBO Annotator, an API to obtain basic information about all the candidate ontologies, and their classes, and alternative resources for extracting information about the acceptance of each ontology. Because it is **integrated in the NCBO technology, the Recommender is already available within the SIFR BioPortal and AgroPortal**. We shall note that these services do not yet rely on the new metadata model presented Section IV.1.1 as discussed in Section V.2.

Because the subject is still very much of interest in the semantic web community, we are currently working with the NCBO group at Stanford and Loughborough University on a review of automatic ontology selection and recommendation systems. This work is currently in progress.

## IV.2 Challenge 2: Multilingualism

Scientific discoveries that could be made with help of ontologies to annotate, integrate, mine and search data, are often limited by the availability of ontology-based tools and services only for one natural language, usually English, for which there exist the most ontologies. Recently, ontology localization, i.e., “the process of adapting an ontology to a concrete language and culture community” [19], has become very important in the ontology development lifecycle, but when efforts are made to properly represent lexical (e.g., using Lemon [91]) or multilingual information (e.g., using LexOMV [92] or Lemon translation module [93]) are made, it is rarely leveraged by ontology libraries and repositories. In the future:

**We need ontology repositories to entirely support interface and content internationalization and be multilingual by enabling a complete use of their functionalities and services for multilingual ontologies or monolingual ontologies linked one another.**

We distinguish *interface internationalization* –which consists of displaying static elements of the user interface (e.g., menu names, help, etc.) in different languages and enabling to switch from one language to another– from *content internationalization* –which consists in displaying an ontology repository content (e.g., ontology labels, mappings, etc.) in another language. However, the need goes beyond internationalization (which is mainly related to display) to to enable a complete use of the functionalities and services of the repository for any

multilingual or monolingual ontologies and data. We call a *multilingual ontology*, an ontology that provides labels or lexicalizations in different *natural languages* and uses the standard ways to differentiate them (e.g., `rdfs:label` et `xmllang` property with values in ISO-639-3) or a rich lexical representation (e.g., Lemon). For instance, Orphanet ontology [94] was constructed with labels in 5 languages. We call a *language specific ontology*, or a *monolingual ontology*, an ontology with labels in a unique natural language that usually serves as the basis for conceptualization. These ontologies are either being originally developed in a given language or are the result of a translation of an ontology in another language. For instance, MeSH-fr is the specific French version of MeSH translated by the French INSERM organization (<http://mesh.inserm.fr>).

Multilingualism must be handled in a proper semantically rich and consistent manner (i.e., using the appropriate semantic web mechanisms and vocabularies) **enabling use of ontologies independently of the language** and therefore enabling cross lingual search, annotation and mining of data indexed with ontologies. Multilingualism became an important issue with the explosion of data being released and linked over the web today. Within the semantic web community research about multilingualism has gained a lot of interest in the last years [18]. Several approaches have been proposed to add lexical information to ontologies such as SKOS-XL, Lexvo [95], Lingvoj, resulting on the proposition of the Ontolex-Lemon standard [91]. For instance, instead of using `rdfs:label` or `skos:*Label`, one can use the SKOS-XL extension to define labels as classes with property `skosxl:literalForm` for the label itself. This reification of the label property allows defining further properties for labels e.g., acronym, short forms, translations. This solution offers a richer description of what a label is and support entailment to SKOS. The state-of-the-art for **adding complex lexical information to an ontology is the Ontolex-Lemon** (LExical Model for ONtologies) model done within the Monnet EU project, which is designed to represent lexical information about words and terms relative to an ontology. Lemon allows for instance, to add part-of-speech information to terms thanks to a clear separation of the lexicon and ontology layers in the model. Lemon perfectly defines how to represent translations within a multilingual ontology<sup>15</sup> and being multilingual certainly means to parse Lemon translation descriptions in an ontology repository. A recent extension offers mechanisms to represent even more precisely multilingual content in ontologies [93] by reifying the translation relation into a class with specific attributes.

In the biomedical domain, the UMLS Metathesaurus, a set of terminologies which are manually integrated and distributed by the US National Library of Medicine [39], does contain terminologies in other languages than English. In addition, the HeTOP repository [33] also offers translated terms in multiple languages, especially French, and enables cross lingual search but most of its content is not publicly or easily accessible (e.g., no web service API or ontology download functionality). In both cases, the underlying approach is one of a common meta-model for all the integrated ontologies which means that there exists a unique class for concepts (e.g., the UMLS Concept Unique Identifiers (CUI)) and additional label properties offer translations to multiple languages. This is different from the **NCBO BioPortal approach which does not build a global thesaurus but keep each ontology separated and use alignments to interconnect them**.

In [CJ51], we presented a roadmap for addressing multilingualism in the NCBO BioPortal, which takes English as primary language. We proposed a set of representations to support multilingualism in the repository and to enable a complete use of the functionalities and services for any kind of ontologies and data. In the following sections, we will first review this roadmap and then explain how we have addressed some of the requirements for multilingual ontology repositories.

#### IV.2.1 A roadmap for making BioPortal multilingual

In [CJ51], after explaining why NCBO BioPortal is not multilingual, we established some elements required to implement representation of multilingual content (cf. Figure 10):

1. *Representation of natural language property for an ontology.* We call *natural language*, the language (French, English, Spanish, etc.) used when defining the class labels in an ontology. This language property has not to be confused with the *format language* used to describe the ontology (OWL, RDFS, RRF, etc.). We proposed to use the property `omv:naturalLanguage` from OMV [96].
2. *Representation of translation relations between ontologies.* We call a *translation*, the relation between two monolingual or multilingual ontologies, in different languages, that represent mainly the same knowledge resource (domain, topics, classes, relations). For instance, MeSH-fr is a translation of MeSH. Other relations between ontologies can also be used to be more specific. We proposed to use and extend the DOOR ontology [68] which is the state-of-the-art about ontology relationships.

<sup>15</sup> "A Translation is a special case of SenseVariation involving two lexical senses in different languages that stand in a translation relation in the sense that they can be exchanged for each other without any meaning implications."

3. Representation of the distinction between ontologies with multilingual content. We proposed to extend OMV within BioPortal's metadata model to include and formalize the distinction between multilingual ontologies and language specific ontologies.
4. *Representation of multilingual mappings.* A multilingual mapping states that the terms in the mapped ontologies are a translation of one another (between the natural languages of the ontologies). For instance, Mesh-fr/mélanome has a multilingual translation mapping to Mesh/melanoma. We proposed to represent multilingual mappings (i.e., one-to-one mappings) between concepts from ontologies in different languages as any other BioPortal mapping, but with a specific relation taken from the GOLD ontology (<http://linguistics-ontology.org/>) [97] i.e., the gold:translation property (or sub-properties).

These elements are mainly at the level on metadata description of the ontology. We have addressed the three firsts when working on AgroPortal's new metadata model and the fourth one when working on reconciling alignments (cf. next section). We are currently consolidating the recommendations within our work on MOD, presented in Section IV.1.

Then, in a roadmap for a multilingual BioPortal, we proposed two steps:

5. *Reconciliation of multilingual mappings.* Language specific ontologies that have been produced by translating another ontology will not always precisely describe a way to resolve translations between concepts. If the two ontologies do not use the same URIs, then a one-to-one multilingual mapping need to be reconciled between the ontologies to connect each concept to its translation. Therefore, we need to implement several methods to extract multilingual translation mappings between translated ontologies and then reconcile them into the BioPortal mapping repository.
6. *Internationalization of the portal.* Once multilingual mappings are reconciled within BioPortal, and multilingual ontologies are properly handled, content internationalization of the portal becomes possible. One can switch from a user interface display to another using a contextual link (e.g., clicking on a language flag): in the case of a multilingual ontology a simple change of the label displayed is necessary whereas in the case of a monolingual ontology the concept being displayed has to change using the multilingual translation mapping if exists. Services, such as the Annotator can be used with a language parameter for the language of the given text data. In addition, we will have to translate the user interface (menu, help) and make sure the portal can switch from one language to the other (as any other web application).

We have addressed #5, described hereafter. But not yet worked on #6.

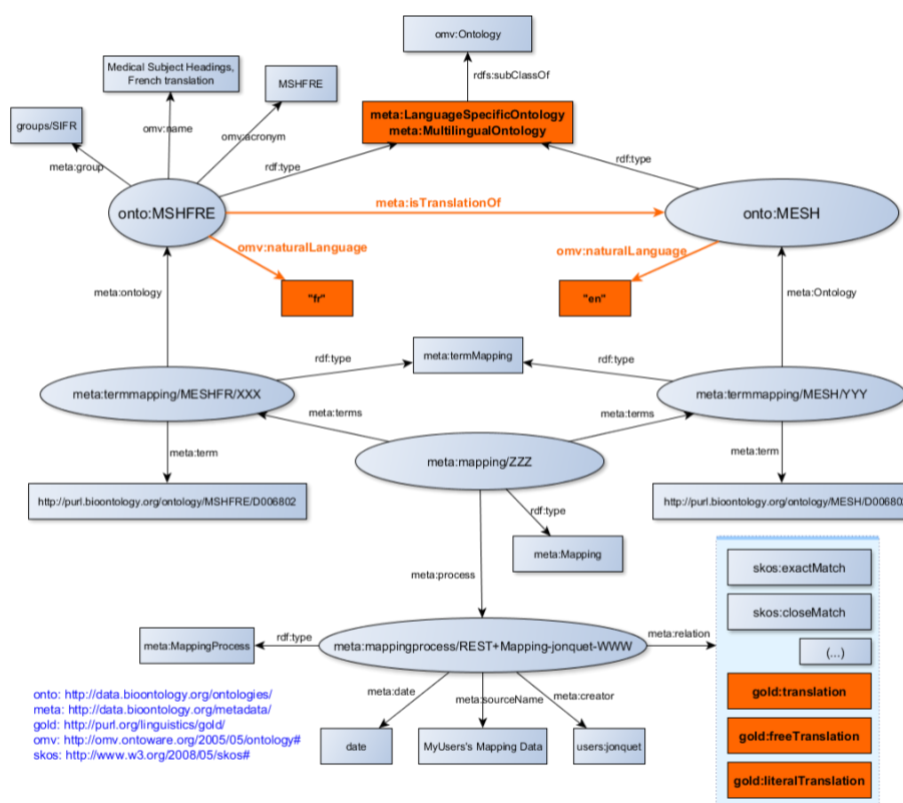


Figure 10. Representations of multilingual content in BioPortal [CJ51]. New elements proposed are in orange.

## IV.2.2 Multilingual mapping reconciliation between English-French biomedical terminologies

Even if multilingual ontologies are now more common, for historical reasons, in the biomedical domain, many ontologies or terminologies have been translated from one natural language to another resulting in two potentially aligned monolingual ontologies but with their own specificity (e.g., format, developers, and versions). Most often, there is **no formal representation of the translation links between translated ontologies** and original ones and those mappings are not formally available as linked data. However, these mappings are very important for the interoperability and the integration of multilingual biomedical data.

To ensure semantic interoperability, it is not enough to just translate ontologies, we must also formally keep the link between objects of the translated ontologies and the original ones [98, 99]. We call reconciliation the process of re-establishing this link formally.<sup>16</sup> These multilingual mappings, once established and represented in a formal way, can have multiple applications [100]. For example, they allow performing a multilingual indexing of biomedical resources, which allows multilingual semantic search. A user types in a query using French terms and retrieves results within English data resources (and vice-versa). In the context of the SIFR project, we wanted to be able to retrieve from a French concept in the SIFR BioPortal, its corresponding English concept in the NCBO BioPortal and vice versa. Our goal was also to improve the French Annotator workflow and enable the annotation of French text with English ontologies.

In [CJ29], we conducted a study on ten French terminologies hosted on the SIFR BioPortal that we wished to formally align with their original English versions hosted on the NCBO BioPortal. All English terminologies came from the UMLS Metathesaurus (version 2015AA) and were imported by the NCBO team in the NCBO BioPortal using the umls2rdf tool (<https://github.com/ncbo/umls2rdf>). The French terminologies came from the UMLS or were provided by the CISMeF group as an OWL file. In this second case, the translations were generally produced or synthesized by CISMeF.

In order **to store our multilingual mappings, we had to change their representation in the SIFR BioPortal's architecture**, especially: (1) To allow tagging the same mapping with several semantic web properties to avoid duplicating the mappings (semantic mapping and translation mapping); (2) To allow the SIFR BioPortal to store mappings that target ontologies in another instance of BioPortal (inter-portal), or in any external resource (external mappings).

Our methodology consisted of (cf. Figure 11): (1) To download ontology files in .ttl or .owl formats from the NCBO and SIFR BioPortal. (2) To parse them with the Jena API to extract the necessary data for multilingual alignment. (3) To store the data in a SQL table (one table per ontology). (4) To make the relevant “join” queries between the two tables on the field/property used to reconcile the mappings. (5) Finally, to post the mappings produced to SIFR BioPortal after choosing the relevant GOLD and SKOS properties.

The complexity of the task mainly came from: (i) Identifying the right property to use to do the join: Often, the original concept code field could be used, but in some cases, it was not available, and another identifier system had to be used. In one case, we also used automatic translation of the labels. (ii) Automatically choosing the right SKOS mapping property to use in the case of broad and narrow match: We had to treat each pair of ontologies apart with its specificities especially in the choice of alignment property and how to recover it. Refinements were needed when translated ontologies did not follow exactly the content of the original ontology (in English).

As a result, **we have reconciled more than 228K mappings** between ten English ontologies hosted on NCBO BioPortal and their French translations hosted on the SIFR BioPortal. We stored these mappings into the modified SIFR BioPortal's mapping repository. Then, we adapted the user interface so that when browsing a concept, we can see in the “Class Mappings” tab the multilingual alignments classified as “Interportal mappings” with a flag to indicate that it is a linguistic mapping to English, we can also observe the properties used. The aligned concept link allows the user to switch from the SIFR BioPortal to the target concept in the NCBO BioPortal. Like all the content of the SIFR BioPortal, in addition to the graphical interface, these multilingual mappings are also available in JSON directly via the REST web service API and a SPARQL endpoint which makes them part of the web of data; easily readable and reusable by any semantic web applications.

---

<sup>16</sup> We use the term reconciliation to avoid the confusion with ontology alignment extraction or creation approaches, which challenge is aligning automatically different ontologies (possibly multilingual). This aspect is addressed in Section IV.3.

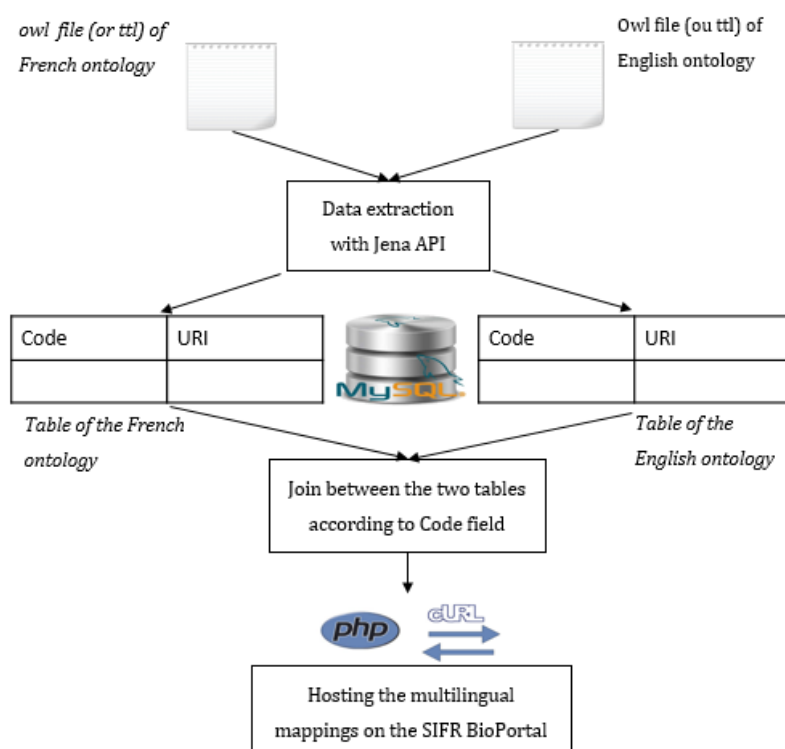


Figure 11. Multilingual mapping reconciliation methodology [CJ29].

### IV.3 Challenge 3: Ontology alignment

Ontologies, or other semantic resources, inevitably overlap in coverage and they are heterogeneous because designed independently, by different developers, and following diverse modeling principles and patterns. To achieve interoperability and integration, one solution is to identify/generate mappings (or correspondences) between different ontologies of the same domain. This process is known as *ontology matching* or *ontology alignment*. Ontology heterogeneity makes the matching process complex [20, 101] but generating the mappings is not the only challenge. Indeed:

**Ontology repositories shall include mapping repositories and support the representation, extraction, harvesting, generation, validation, merging, evaluation, visualization, storage and retrieval of mappings between the ontologies they host and other ones.**

This need has been explicitly expressed by almost all our partner organizations in biomedicine, agronomy or ecology. Surprisingly, it seems there is a gap between the state-of-the-art results obtained in automatically *generating* mappings at each edition of the Ontology Alignment Evaluation Initiative (OAEI – <http://oaei.ontologymatching.org>) and the day-to-day reality of ontology developers. Tools are often hardly reusable, and results cannot be easily reproduced outside of the benchmarking effort; already existing mappings are not uniformly described or not shared/available; mappings quality and provenance is always in doubt; multiple mappings cause conflicts. Plus, there is no recognized standard way to represent mappings (with provenance information) and no shared repository to merge, store and retrieve them. We have identified several **important aspects related to mappings when building a mapping repository** aside of an ontology repository:

- By mapping *representation*, we mean a standard, shared and adopted way to represent a mapping with its metadata/provenance information. Today, there exist no standard, shared and recognized way of representing mappings.
- By mapping *extraction*, we mean being able to extract and load in the repository or exploit in any other way, mappings –with relevant metadata– explicitly declared inside the ontology source file for classes or properties (typically using `owl:sameAs` or SKOS mapping properties).
- By mapping *harvesting*, we mean localize and import previously generated mapping datasets. An equivalent effort as the one made to harvest ontologies, must be made to harvest the mappings between

these ontologies and describe them with metadata and provenance information in the repository. This is a very tedious task.

- By mapping *generation*, we mean the process of automatically identifying correspondences between two given ontologies. An ontology repository shall offer such a capability and state-of-the-art approaches and pluggable tools are desired.
- By mapping *validation*, we mean to offer ontology developers a mechanism to validate an automatically generated alignment. This concerns only the developer or expert loading the ontology in the repository or explicitly interested in the alignments.
- By mapping *merging*, we mean the process of integrating together alignments extracted or generated from different sources (automatic or not) and previously validated. For example, this consists in assigning a score based on the frequency and multiplicity of sources of a mapping in the repository.
- By mapping *evaluation*, we mean to offer a community (i.e., several ontology developers and users) a mechanism to evaluate all the mappings of an ontology repository once they have been merged. Indeed, conflicts or redundancies shall occur when merging validated mappings into a unique repository; therefore, a community-based evaluation of the mappings is also needed.
- By mapping *visualization*, we mean user interfaces allowing to easily navigate and explore the mapping repository content. Such visualizations shall allow to discover links between concepts and ontologies.
- By mapping *storage and retrieval*, we mean the mechanisms to consider mappings first class objects in a repository and especially being able to load, store and retrieve them with appropriate interfaces and APIs (such as JSON or SPARQL).

To the best of our knowledge, of all the ontology repositories discussed in Section III.1 only the NCBO BioPortal offers a mapping repository and some mappings capabilities; consequently, also the SIFR BioPortal and AgroPortal do. Table 5 shows how the NCBO technology address each aspect previously presented.

**Table 5. How mappings are handled in the NCBO technology.**

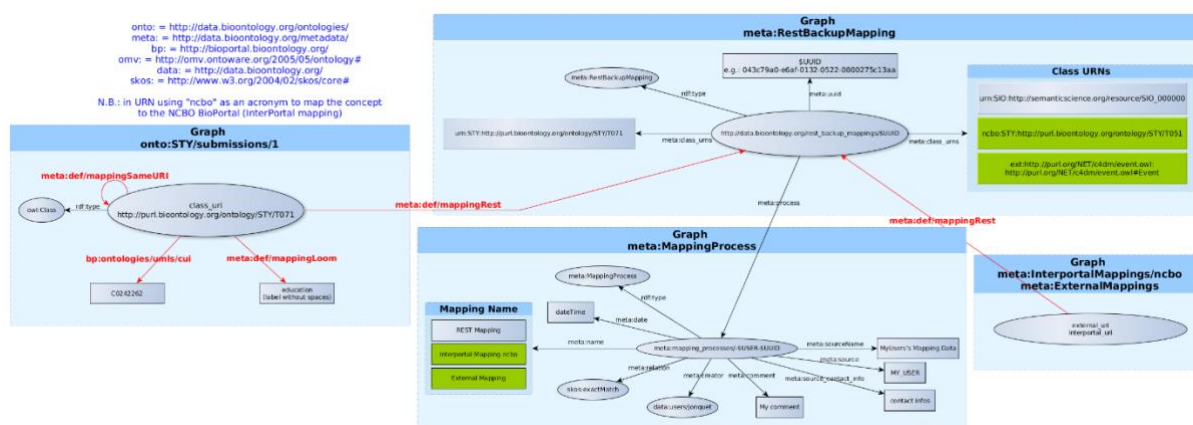
Aspect	How it is addressed in BioPortal	Example
<b>Representation</b>	Mappings are reified into an RDF resource with multiple properties including mapped classes and metadata about the mapping as illustrated in Figure 12. Mappings are exported in JSON by the web service API.	
<b>Extraction</b>	Not addressed. Mappings defined inside the ontologies are not extracted to be included in the repository.	
<b>Harvesting</b>	Not addressed. The NCBO does not harvest the mappings previously generated by external parties (e.g., UMLS, OBO Foundry, etc.).	
<b>Generation</b>	Automatically generates concept-to-concept mappings between two ontologies hosted within the portal when two classes share the same identifiers properties (URI or UMLS CUI), or when they share a common normalized preferred label or synonym.	
<b>Validation</b>	Not addressed.	
<b>Merging</b>	Not addressed.	
<b>Evaluation</b>	Not addressed. Although users (especially the ontology owner) can be notified when a mapping is manually created by another user on their ontology of interest.	
<b>Visualization</b>	Offer two basic interfaces: (i) when browsing a concept, one can see the links to the mapped classes in other ontologies; (ii) a global mapping tab allows to see the numbers of mappings between one ontology and all the other ones in the portal and download them.	
<b>Storage and retrieval</b>	The REST web service API allows to create (PUT), modify (POST) and retrieve (GET) mappings inside the repository. They are also available via the SPARQL endpoint.	

Each of these aspects related to mappings have been addressed partially by the semantic web community. However, there is a challenge for ontology repositories to address all of them in the same environment. For instances:

- To be able to participate to the competition, participants of the OAEI campaigns must use a common XML-based representation format called EDOAL (Expressive and Declarative Ontology Alignment Language) defined by the Alignment API [102].

- The European Bioinformatics Institute is currently working on OXO [103] in which they extract mappings declared inside the source file using OBO-XREF (formally `oboInOwl:hasDbXref`) property.<sup>17</sup> They also extract mappings from the UMLS.
- The US National Library of Medicine harvests in the UMLS some mappings produced by the biomedical community (MRMAP table), but this is a very tedious task and limited to the sources the UMLS Metathesaurus actually imports.
- The question of automatically generating mapping is certainly the most studied by the community [20, 101] and the multiple OAEI campaigns have allowed to reach very good theoretical performance.
- With YAM++ online [104] (<http://yamplusplus.lirmm.fr>) LIRMM has designed a prototype interface which allows users of the YAM++ ontology matcher to validate the mapping results one-by-one.
- Multiple works have addressed the question of fancy and relevant visualizations for ontology mappings [105–107].

In the following we will present three different works in the area of ontology alignment. First, we will motivate the importance of ontology mappings inside ontology repositories by showing what mappings tell us about the ontologies themselves, the structure of the ontology repository, and the ways in which the mappings can help in the process of ontology design and evaluation. Second, we will present our work on mapping generation algorithms using existing ontology alignments as background knowledge. Third, we will present our current work in AgroPortal to address the mapping extraction, harvesting, validation, merging, evaluation, and visualization aspects previously presented.



**Figure 12. Graphical representation of a mapping in BioPortal.** In green are the new elements that we have added to the model to represent interportal and external mappings.

### IV.3.1 What four million mappings can tell you about two hundred ontologies

Since the very beginning, we viewed mappings between concepts in different ontologies as an essential part of the NCBO BioPortal ontology repository: **mappings between ontology concepts are first class objects**. Users can browse the mappings, create new mappings, upload mappings created with other tools, download the mappings stored in BioPortal, or comment on the mappings and discuss them [108]. A previous study has shown that in the case of biomedical **ontologies simple lexical techniques, such as comparing preferred names of concepts and their synonyms, are extremely effective** in generating mappings. In fact, these techniques often perform better than advanced techniques [109] even if resulting lexical mappings can be sometime inaccurate and should be used with caution [110, 111].

In 2009, the NCBO BioPortal had around 140 ontologies and was about to include around 70 terminologies of the UMLS. In [CJ40], we constructed a resource of approximately 4 million mappings automatically generated between concepts in these semantic resources based on the lexical similarity of concept names and synonyms [109].<sup>18</sup> In doing this, we had two goals: (i) we wanted to create a mapping repository in BioPortal that

<sup>17</sup> Database cross references are used by the OBO community to interconnect an ontology term to another related entity generally in a database. However, they have a very poor semantics and are used in a idiosyncratic way to capture any kind of link (similar to the `rdf:seeAlso` property) including mappings between ontologies.

<sup>18</sup> The algorithm in BioPortal was later called LOOM.

other applications can access and use; and (ii) we wanted to learn more about the characteristics of the ontologies and the relationships between them. Using our set of more than 4 million mappings generated over our ontology set, we wanted to answer several practical questions with implications for ontology reuse and development, such as:

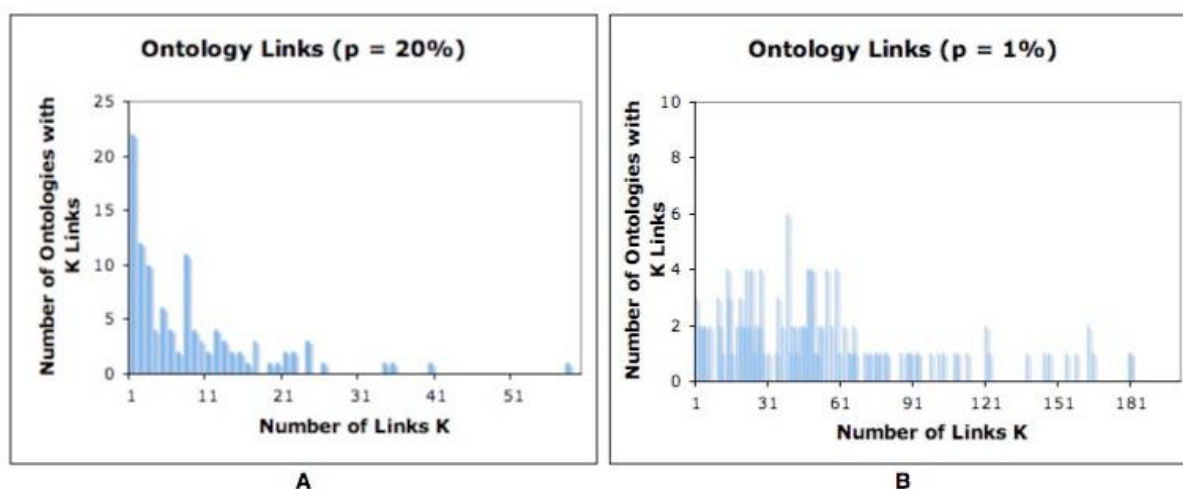
- To what degree are the domains covered by different ontologies connected?
- If you are new to a domain, what are the important or representative ontologies with good coverage?
- Can we identify domains that are not well covered by existing ontologies?
- If you want to build domain-specific tools for creating ontology mappings, what are good ontologies to use for background knowledge?
- What can we learn about the characteristics of the ontologies themselves and the ontology repository from the mappings between them?

We use network analysis methods to answer these practical questions and to reason about the distribution of mappings among the ontologies. In the following, we illustrate some of the results.

We defined a *percent-normalized link* between ontologies as:

Given two ontologies, the source ontology  $S$  and the target ontology  $T$ , and a set of mappings  $M(S, T)$  between them, we say that there is a percent-normalized link between  $S$  and  $T$ ,  $L_p(S, T)$  where  $p \geq 0$  and  $p \leq 100$ , if and only if at least  $p\%$  of the concepts in the ontology  $S$  are sources for the mappings in  $M(S, T)$ . For instance, if an ontology  $S$  has 1000 concepts, and 500 of these concepts are mapped to concepts in an ontology  $T$ , then  $L_p(S, T)$  is true for all values of  $p$  from 0% to 50%.

For instance, Figure 13 shows a distribution of the number of links that ontologies have for two values of  $p$ . The graph demonstrates the power-law distribution for  $p = 20\%$ : there is a small number of ontologies that have a large number of links (hubs) and a large number of ontologies with just a few links. If we use  $p = 1\%$  (there is a link from one ontology to another if at least 1% of its concepts are mapped), the distribution becomes essentially random.



**Figure 13. Number of links between ontologies for(a)  $p = 20\%$  and (b)  $p = 1\%$  [CJ40].** The x-axis represents number of links to other ontologies that each ontology has. The y-axis represents number of ontologies with that number of links.

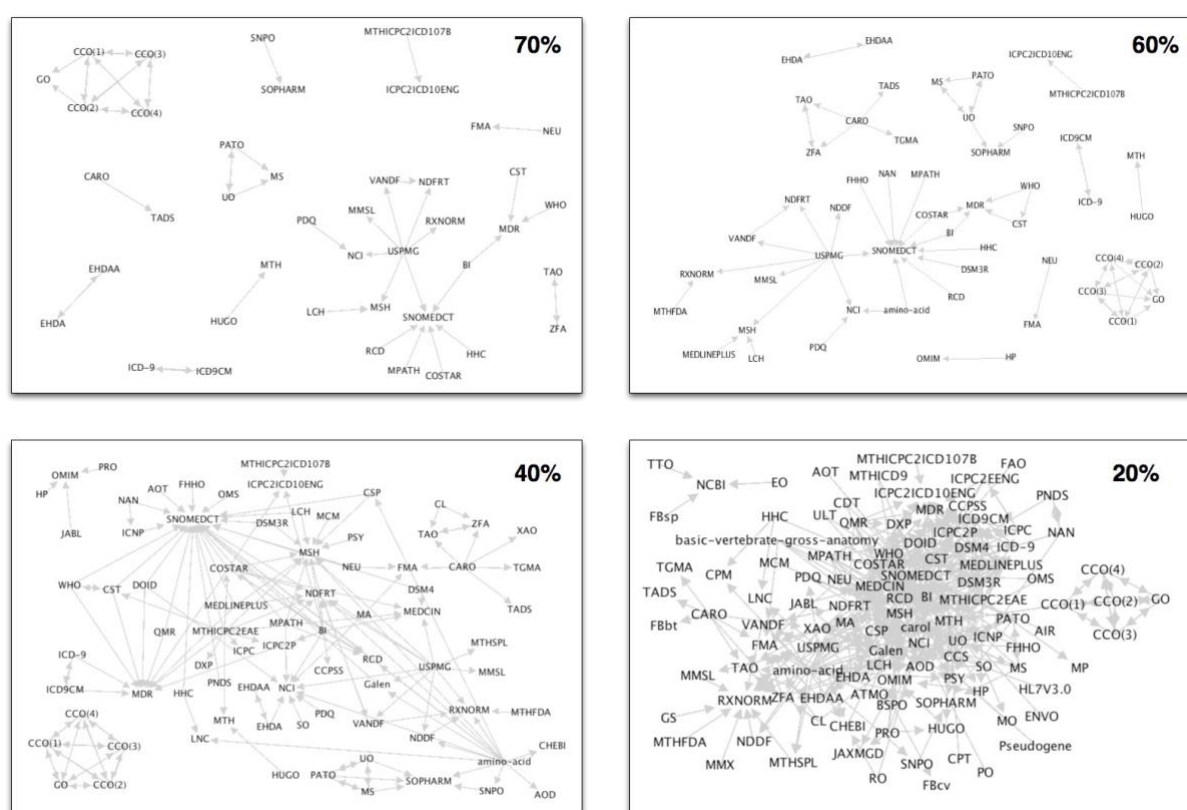
In Figure 14, we constructed directed graphs of ontologies at several different thresholds of  $p$ . We used these graphs to identify connections between different ontologies, based on varying levels of overlap. The graphs identified clear hubs—ontologies to which many other ontologies link—such as SNOMED-CT.<sup>19</sup> Hubs with many outgoing links show shared domains, particularly at high threshold values for  $p$ . For these hub ontologies, a large portion of their concepts is mapped to several different ontologies. Thus, ontologies that are linked through such a hub likely share the content that is represented in the hub ontology. At  $p=40\%$  we can distinguish a cluster of ontologies about anatomy (around the CARO node) or at  $p=70\%$ , one about drugs (around node USPMG).

Overall our analysis showed:

<sup>19</sup> Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT) is one of the most comprehensive and precise clinical health terminology.

- The biomedical ontologies in our set are very closely connected, with 33% of them having at least half of their concepts mapped to concepts in other ontologies.
- With such a large overlap among the ontologies, one can say a large number of concepts in biomedicine are already represented in many different ways.
- The domain of biomedicine is such that there is a little bit of overlap in everything, resulting in the extremely connected model we see at  $p=1\%$ . At  $p=20\%$ , however, we see a meaningful power-law distribution. At even higher thresholds, we can see ontologies that are very closely related.
- We can identify cluster of ontologies for a subdomain and identify prominent ontologies (i.e., an ontology with lots of mappings to other ontologies is an “important” one).

Although the work done in [CJ40], had several limitations and would need to be reproduced on a more recent ontology set in the NCO BioPortal, our study showed the importance of ontology alignment in an ontology repository. It showed what the mappings tell us about the ontologies themselves, the structure of the ontology repository, and the ways in which the mappings can help in the process of ontology design and evaluation. A similar, more recent study about ontology terms reuse have been done by Kamdar et al. [112].



**Figure 14. Network analysis of the links in our ontology sets [CJ40].** The graphs show percent-normalized links between ontologies that are true for  $p = 20\%$ ,  $40\%$ ,  $60\%$ , and  $70\%$ .

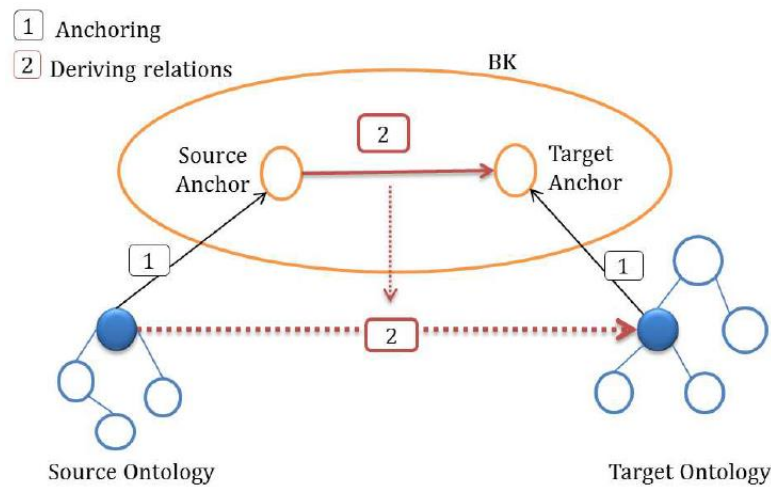
#### IV.3.2 Enhancing ontology matching with background knowledge (A. Annane’s PhD project)

During Amina Annane’s PhD project (2015-2018), directed by Pr. Zohra Bellashene and Pr. Faical Azouaou (ESI Alger), we investigated the question of mapping generation [41]. We experimented and evaluated our approaches and algorithms in the biomedical domain, thanks to the profusion of knowledge resources in biomedicine (ontologies, terminologies and existing alignments). The following is a summary of her contributions and accepted publications.

Original ontology matching methods usually exploit the lexical and structural content of the ontologies to align; this is known as *direct matching* or *content-based matching*. To that end, many syntactic and structural similarity measures have been developed [101, 113, 114]. However, **direct matching is less effective when equivalent concepts have dissimilar labels and are structured with different modeling views**. To overcome this semantic heterogeneity, the ontology matching community has turned to the use of external *background knowledge (BK)*

resources as a semantic bridge between the ontologies to align. This approach is known as *indirect matching*, *BK-based matching* or *context-based matching* [115], as it **exploits external resources to identify mappings between the ontologies to align**. The BK-based matching approach raises two main issues: (i) how to select (or build) background knowledge resource(s) for a given ontology matching task? and (ii) how to concretely use the selected background knowledge resource(s) to enhance the quality of the matching result? In the literature, several works have dealt with these issues jointly or separately [115–120].

Exploiting background knowledge resources in the matching process includes three steps (cf. Figure 15). The first one, called *anchoring*, aims at linking the entities of the ontologies to align to the entities of the selected resources. This is usually performed by a direct matching between the ontologies to align and the selected resources. The second one, called *deriving*, deduces semantic relationships between the anchored entities according to the relationships linking the anchors in the background knowledge resource. Finally, the third step (not on figure) aggregates the derived mappings and selects the most relevant ones to produce the final alignment.



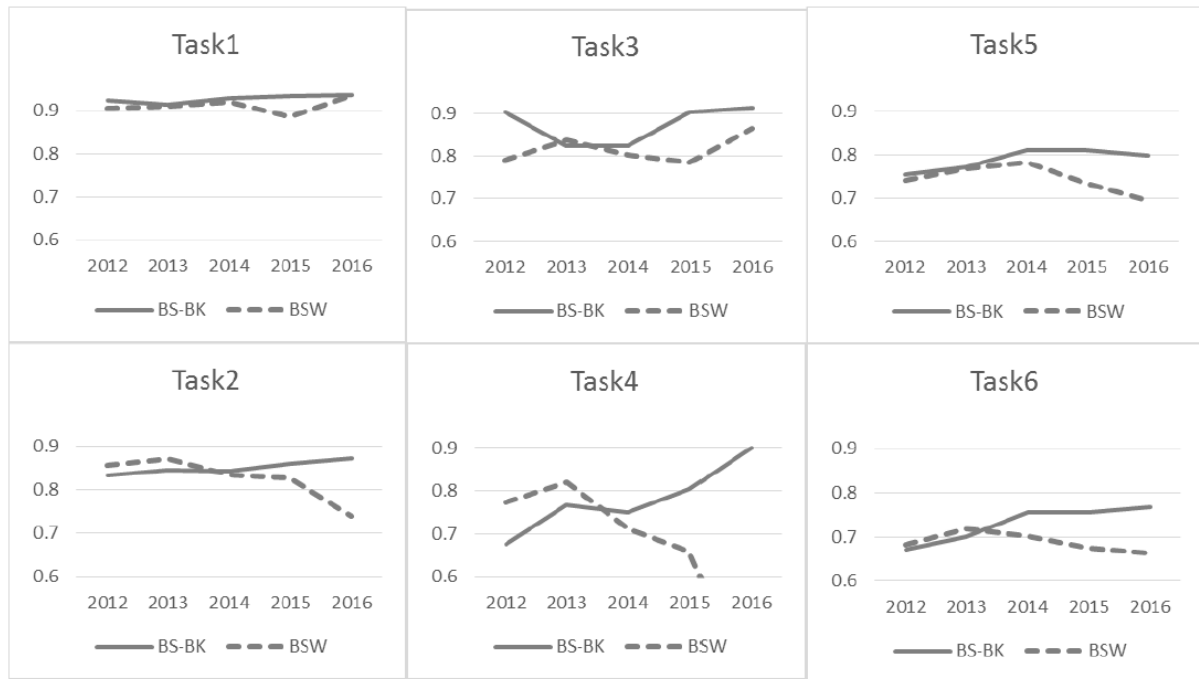
**Figure 15. Exploiting a background knowledge resource to generate mappings [CJ6].**

The use of BK resources is one of the main challenges of ontology matching [121]. In [CJ-UR3], we made a systematic review and historical evaluation comparison of state-of-the-art ontology alignment approaches using background knowledge resources. We provide a **synthetic classification and present a comparative evaluation by analyzing system performance** results obtained during Ontology Alignment Evaluation Initiative (OAEI) 2012-2016 campaigns. We thus evaluate the benefit of using BK resources and the improvement achieved by this approach regarding the systems that do not use background knowledge. Our survey shows:

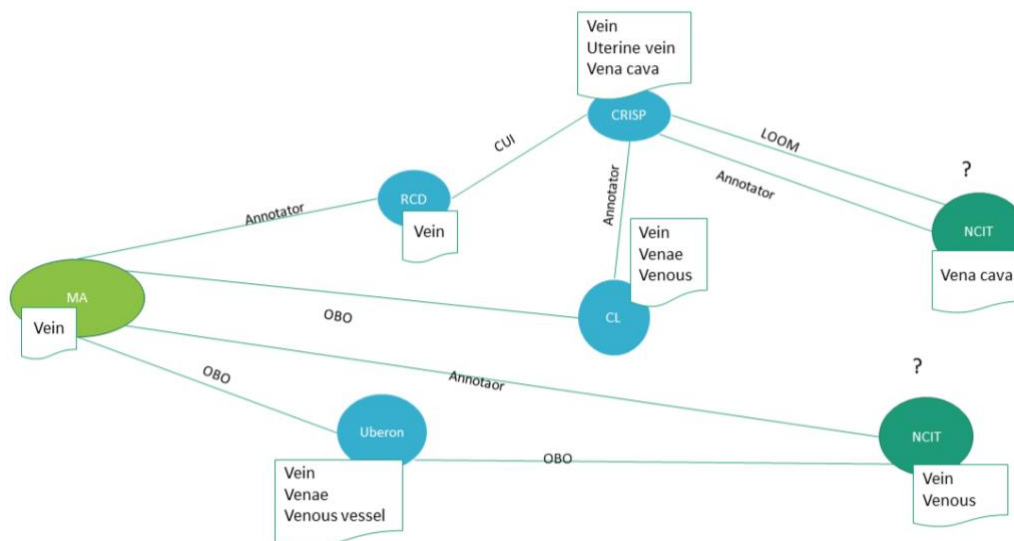
- BK-based matching systems outperformed all the systems that do not use BK resources in OAEI campaigns during the last four years as illustrated by Figure 16 (except for the large biomedical track in 2012-2013). Moreover, methods using BK resources allow to discover new mappings which have not been found by lexical and structural measures. The background knowledge plays the role of the semantic bridge between the initial ontologies to align. The use of BK is thus necessary in the presence of important semantic heterogeneity.
- BK-based matching methods are domain independent. The use of generic lexical resource such as WordNet for lexical enrichment allowed to obtain high scores in different tracks of the OAEI competition. However, experiments show such generic resource are prone to produce erroneous mappings in domains with specialized vocabularies, such as the biomedical domain. In this case, specialized BK resources seem to be better than generic ones. The best systems in the biomedical tracks of the OAEI competition in biomedical tracks (AML [122] and LogMapBio [123]) use biomedical ontologies as background knowledge.
- We studied if the use of the appropriate BK resources may replace the combination of different direct matching systems (based on several similarity measures). Currently the use of background knowledge is considered as an extension of the systems and not the main component.

**Ontologies, others than the ones to align, are the most frequently used type of background knowledge resources.** Several methods have been proposed to select ontologies, other than the ones to align, as background knowledge, however, these methods return a set of complete ontologies, while, in most cases, only fragments of the returned ontologies are effective for discovering new mappings. Related works often select a set of

complete ontologies. In [CJ27] and then in [CJ6], we proposed a novel BK-based ontology matching approach to select and **build a background knowledge resource with just the right concepts chosen from a set of ontologies**. We picked up only relevant concepts and relevant existing mappings linking these concepts all together in a specific and customized background knowledge graph. Then we used **paths within this graph to discover new mappings** between the ontologies to align as illustrated in Figure 17. We have implemented and evaluated our approach using the content of the NCBO BioPortal repository and the Anatomy benchmark from the OAEI. We used the mapping gain measure [116] to assess how much our final background knowledge graph improves results of state-of-the-art alignment systems. Our experiments showed that our BK selection approach improves efficiency without loss of effectiveness. Furthermore, the evaluation shows that our approach discovers mappings that have not been found by state-of-the-art systems.

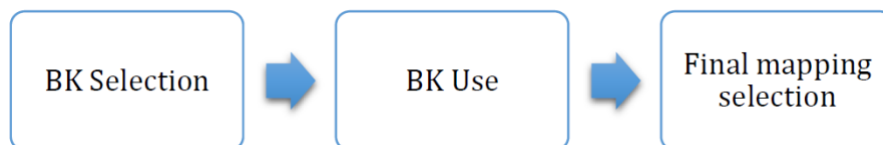


**Figure 16.** OAEI's LargeBio tracks results evolution (2012-2016) [CJ6]. The dashed (resp. continuous) line represents the best F-measure obtained by systems that do not use BK resources (resp. use BK resources)).



**Figure 17.** Using the graph of mappings as background knowledge to discover new mappings between source and target ontologies (here Mouse Anatomy (MA) and NCI Thesaurus (NCIT) [CJ27]. Node in the graph are ontology concepts. Edges are different types of already existing mappings used in the BK (LOOM, OBO XREFS or generated with the NCBO Annotator).

Exploiting background knowledge resources in ontology matching is a double-edged sword: while it may increase recall (i.e., retrieve more correct mappings), it may lower precision (i.e., produce more incorrect mappings) [115]. Consequently, **selecting correct mappings from the candidate ones is particularly challenging in the context of BK-based matching**. In [CJ6], we extended our previous work on selecting/building and using the background knowledge resource and proposed two new selection methods. The first one is based on a set of rules, while the second one is based on supervised machine learning, as described in the following paragraphs.



**Figure 18.** Main steps of our BK-based ontology matching approach [CJ6]. Only the 3<sup>rd</sup> one is described here.

Because of the structure of our BK resource, candidate mappings consist in a set of paths linking the source concept to the target one. Several paths may represent the same candidate mapping. Thus, to compute the final score  $k$  for a given candidate mapping, we must address two issues:

- How to *compose* the different mapping scores of the same path?
- How to *aggregate* the scores of different paths representing the same candidate mapping?

Related work suggested to use algebraic functions, such as multiplication, average, maximum, etc. to compose and aggregate different mapping scores [124]. In our approach, we use the term *configuration* for a given pair of composition and aggregation functions.

**Rule-based selection:** We defined a set of rules to decide whether or not to keep a given candidate mapping in the final alignment:

1. Mappings returned by direct and indirect matching are selected.
2. Mappings resulting from the composition of only manual mappings are selected.
3. For each source concept, the target candidate with the highest mapping score is retained.
4. For each target concept, the source candidate with the highest mapping score is retained.

For rules 3 and 4, the score may be controlled by a given threshold. The score of the candidate mappings is computed with the multiplication-maximum configuration.

**Machine learning-based selection:** As testing the performance of multiple rules and configurations can be long and fastidious, Machine Learning (ML) is an interesting alternative for the selection of candidate mappings assuming training data are available:

- Test data are candidate mappings between the source and target ontologies to be classified true or false.
- Training data are a set of candidate mappings already classified as true or false. These candidate mappings are completely distinct from test data.
- Attributes (or features) which describe each candidate mappings are the different configurations and any variable that can help to classify a given candidate mapping.

In our case, we used data from the OAEI tracks as training data (with cross validation and separate learning) and proposed 27 selection attributes such as: direct score (if the candidate mapping belongs to the alignment returned by the direct matching), number of paths in the BK graph, path length (min, max, avg), scores (21 different). We used the *RandomForest* algorithm (a non-linear method based on decision trees) [125] available in the Weka platform [126] and Neo4j (<https://neo4j.com>) to store and compute the BK graph.

We evaluated our approach with extensive experiments on two Ontology Alignment Evaluation Initiative (OAEI) benchmarks: Anatomy and LargeBio tracks. According to the OAEI 2016 campaign, AML [122] and LogMapBio [123] are the best BK-based ontology matching systems. To establish a fair comparison with these systems, our evaluation employs the same set of preselected ontologies to build the BK resource. We also added to the BK some OBO XREFS mappings between our BK ontologies. We used YAM++ to generate all the required direct alignments for our experiments (e.g., anchoring to BK). YAM++ is a state-of-the-art direct ontology matcher previously developed at LIRMM [114]. It combines several syntactic, lexical and structural similarity measures; it was top ranked in OAEI 2013 [127].

Our results confirmed the effectiveness and efficiency of our approach; we showed [CJ6]:

- Our BK selection method builds a smaller-size BK than the preselected ontologies and the small size of the built BK does not affect its effectiveness;

- Deriving mappings across several intermediate concepts generates more correct mappings than deriving across one intermediate concept;
- Our selection methods are effective however, ML-based selection promotes precision, while rule-based selection promotes recall;
- Our built BK reduces the computation time of using ontologies as background knowledge;
- Our method slightly overcomes or competes with state-of-the-art matchers exploiting background knowledge resources on experimented OAEI tracks (AML and LogMapBio).

During Amina Annane's PhD project, **we have successfully participated twice to the OAEI competition** (in 2017 and 2017.5). In 2017 [CJ112], we used a system called YAM-BIO which was an extension of YAM++ but dedicated to aligning biomedical ontologies with a new component that uses existing mappings between multiple biomedical ontologies as background knowledge.<sup>20</sup> We applied the indirect matching technique only for the source concepts that have not been previously matched directly by YAM++. We participated in two tracks: Anatomy and LargeBio and obtained results very close to top ranked state-of-the-art systems. We ranked 2<sup>nd</sup> in the Anatomy track (with results very close results to winner) and 1<sup>st</sup> in Task 1 and 4 of the LargeBio track.<sup>21</sup> We do not report results from 2017.5 edition here.

### IV.3.3 Building a “Lingua Franca” in agri-food and biodiversity

In the previous section, we have demonstrated that existing mappings between ontologies can be used to improve ontology alignment; in other words, a centralized mapping repository is an excellent resource to curate and generate new mappings. However, **results obtained in the biomedical domain, thanks to the profusion of knowledge resources (ontologies and existing alignments), will not easily be reproducible to other domains.** Within AgroPortal project, we are interested in building such a mapping repository for agronomy, plant sciences, food and biodiversity.

In 2018, we started the “Lingua project” to investigate other aspects relative to ontology alignment such as mapping extraction, harvesting, validation, merging, evaluation, and visualization. Our goal is to make AgroPortal the reference platform for ontology alignment in agri-food and biodiversity by adopting a complete semantic web and linked open data approach and by engaging the community. We will **build a complete ontology alignment framework, based on BioPortal/AgroPortal and YAM++, that covers the whole ontology alignment life cycle** from hosting/accessing ontologies to semi-manual and community-based evaluation of the merged mappings (illustrated in Figure 19). From AgroPortal or YAM++ online, users will select the ontologies to align then select the specific matcher components (algorithms) to use and the system will support manual validation of mappings after the execution of the matchers. We envision a two-phased evaluation of mappings: (i) mappings automatically generated with YAM++ will be manually validated inside the YAM++ web application; (ii) then, once the mappings are uploaded to AgroPortal (with provenance information) and the mappings have been shared with the community, they will be incorporated into a global evaluation mapping page that will merge and display mappings from all sources into a unique view (with scores) that will facilitate evaluation of the mappings. Such two-phase mapping validation + evaluation will reinforce the trust of the community in the mappings being generated and hosted by the platform. Finally, existing mappings in the repository will themselves be used in following executions of other ontology alignment algorithms using background knowledge (as presented in Section IV.3.2). This work is in progress.

<sup>20</sup> At that time, our complete BK-based methodology was not fully designed and implemented yet, we therefore decided to participate with a hybrid tool (YAM-BIO) inspired from our research.

<sup>21</sup> Detailed results are available on the OAEI web page: <http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2017>

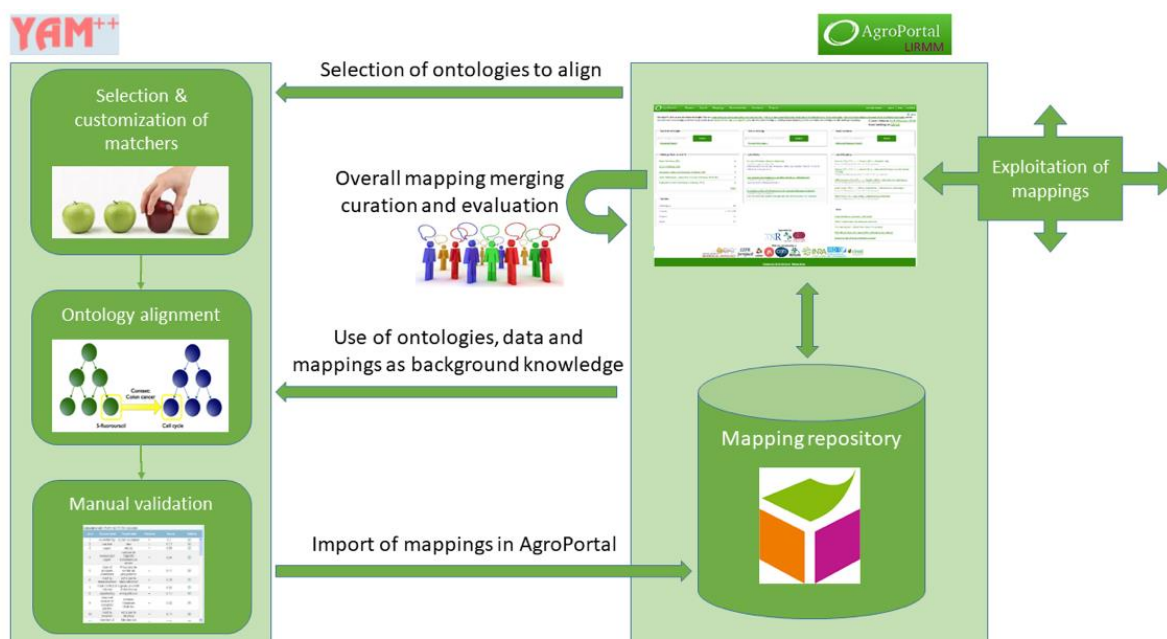


Figure 19. Ontology alignment framework between AgroPortal and YAM++. Work in progress.

#### IV.4 Challenge 4: Generic ontology-based services (especially for free text data)

Ontology repositories offer a large span of services: file hosting, versioning, search/browse content, visualization, metrics, notes, mapping, etc. These services are ‘generic’ if they are domain independent i.e., not specific to a domain, group of ontologies, specific format or design principles. Therefore:

**Ontology repositories must continue to enhance ontology-based services and integrate new generic ones to enlarge the spectrum of possible use of ontologies, especially related to data annotation.**

Using standard formats such as OWL or SKOS has facilitated the development of a wide range of tools and applications for semantic resources. The challenge is now to package them inside ontology repositories and keep vertical quality (i.e., one ontology) while enabling quantitative horizontal use (i.e., multiple ontologies). By integrating an application, tool or service within an ontology repository, researchers and developers face different type of issues that the ones tackled when designing the original tool: it will have to work for a wide range of heterogeneous semantic resources including some designed with different styles and formats, it will have to scale up to extremely large semantic resources, it will have to adopt standards input/output format and technologies (e.g., the semantic web languages such as RDF, OWL, SKOS, SPARQL, etc.), it will have to be technically robust, stable and long term maintained, it will have to be smartly integrated with other services of the repository while staying decoupled as much as possible for future maintenance and evolution, and finally, it will have to keep and guaranty its original quality/performance when used within the repository.

Those are the scientific and technological challenges we have faced up when working on ontology-based services, especially for text data annotation.

**One important use of ontologies is for annotating and indexing text data.** Indeed, ontologies allow representing data with clear semantics that can be leveraged by computing algorithms to search, query or reason on the data. One way of using ontologies is by means of creating semantic annotations. An annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the term. When doing ontology-based indexing, we use these annotations to “bring together” the data elements from these resources. However, explicitly annotating data is still not a common practice for several reasons [CJ41]:

- Annotation often needs to be done manually either by expert curators or directly by the authors of the data;
- The number and format of ontologies available for use is large and ontologies change often and frequently overlap;
- Users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves;
- Annotation is often a boring additional task without immediate reward for the author.

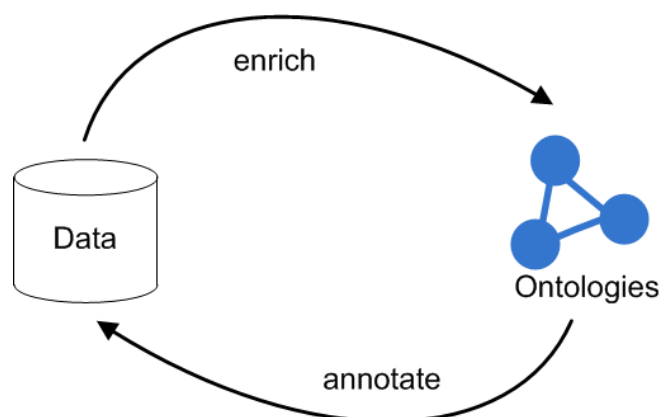
Semantic annotation is an important research topic in the semantic web community [7, 128]. Tools vary along with the types of documents that they annotate (e.g., image annotation [129]). For an overview and comparison of semantic annotation tools the reader may refer to the study by Uren et al. [6]. In the following we restrict our review and contributions to biomedicine.

Previous work has encouraged and exalted the use of biomedical ontologies for annotation at various levels [130–134]. For a while, the prevalent paradigm in the use of ontologies was that of manual annotation and curation. However, several researchers have shown that such **manual annotation, though highly desirable, will not scale to the large amounts of data being generated in the life sciences** [135]. If one examines the reasons for the low adoption of ontology-based annotation methods among database providers [45, 136], the high cost of manual data curation remains the main obstacle. In light of this situation, researchers have called for the need of automated annotation methods [137, 138] and for leveraging natural language processing tools in the curation process [128, 139]. Related efforts in the community that aim to facilitate the use of ontologies in automated annotation and curation pipelines include: Terminizer (Manchester University) [140], OnTheFly and Reflect (EMBL) [141, 142], Whatizit (EBI) [143], MetaMap (NLM) [138] and older projects such as: IndexFinder [134], SAPHIRE [144], CONANN [145]. Recent reviews on semantic annotation in biomedicine include [146–149].

Logically, ontology-based annotation services often accompany ontology repositories. For instances, BioPortal has the NCBO Annotator [CJ41] (described in the following sections), OLS had Whatizit [143] and now moved to ZOOMA, CISMef HeTOP had FMTI [150] and now ECMT [151] and UMLS has MetaMap [138].

Within the NCBO project, between 2007 and 2010, we have **developed methods to annotate large numbers of data resources automatically, and prototyped several systems for ontology-based annotation and indexing of biomedical data**. As presented in Section IV.4.1 and then IV.5.1, we have integrated these applications within the NCBO BioPortal platform and they became some of the most frequently used services in the platform. After 2010, after working on semantic annotations for English data –and with English ontologies– our motivation was to offer the same kind of resources for the French context. Indeed, French is not in the same situation: there is little readily available technology (i.e., “off-the-shelf” technology) that allows the use of ontologies uniformly in various annotation and curation pipelines with minimal effort. This was the inception of the SIFR project.

**Working on semantic annotation usually goes pairwise with working on knowledge extraction.** Indeed, ontologies and data are two elements of a repeated life cycle, as illustrated in Figure 20: on one side the community use ontologies to annotate data (sometime develop them with mainly this goal) [7] and then exploit the semantics of the ontologies to search or mine the data annotated –or semantically indexed; on the other side, the data can themselves be used to enrich the ontologies with manual, automatic or hybrid methods to extract new concepts and terms, relationships or rules to include in the ontologies (e.g., [152]).



**Figure 20. Data-ontologies life-cycle.** Ontologies are often developed to annotate data. Data are often used to enrich ontologies.

In the following, we will first focus on the bottom part of the lifecycle Figure 20, and present our work on semantic annotation of free text, first building in 2009 the NCBO Annotator, one of the most used ontology-based annotation web service in biomedicine and second by investigating similar questions but for French biomedical data within the SIFR project and exploring the challenges of dealing with clinical text. Second, we will look at the top part of the lifecycle Figure 20, and present our work on automatic terminology extraction and ontology enrichment.

#### IV.4.1 Semantic annotation of biomedical text with the NCBO Annotator

In [CJ41] and then in [CJ18], we presented the NCBO Annotator (<http://bioportal.bioontology.org/annotator>), a web service which provides a mechanism to employ ontology-based annotation in curation, data integration, and indexing workflows, using any of the several hundred public ontologies in the NCBO BioPortal repository. The NCBO Annotator **tags raw text descriptions with relevant biomedical ontology concepts** and returns the annotations to end users.

The NCBO Annotator workflow is composed of two main steps illustrated in Figure 21. First, the user submitted text is given as input to a *concept recognition tool* along with a dictionary. The dictionary (or lexicon) consists of a list of strings that identify ontology concepts. The dictionary is constructed by pooling all concept names and other lexical identifiers, such as synonyms or alternative labels that identify concepts. The Annotator uses Mgrep [153], a concept recognizer developed by the University of Michigan that **enables fast and efficient matching of text against a set of dictionary terms to recognize concepts** and generate *direct annotations*. Second, **semantic expansion components use the ontology structure to create additional annotations**. For example, the *is\_a transitive closure* component traverses an ontology parent-child hierarchy to create additional annotations with parent concepts. The *ontology-mapping* component creates additional annotations based on existing mappings between ontology terms. The direct annotations and the set of *semantically expanded annotations* are scored and returned to the user.

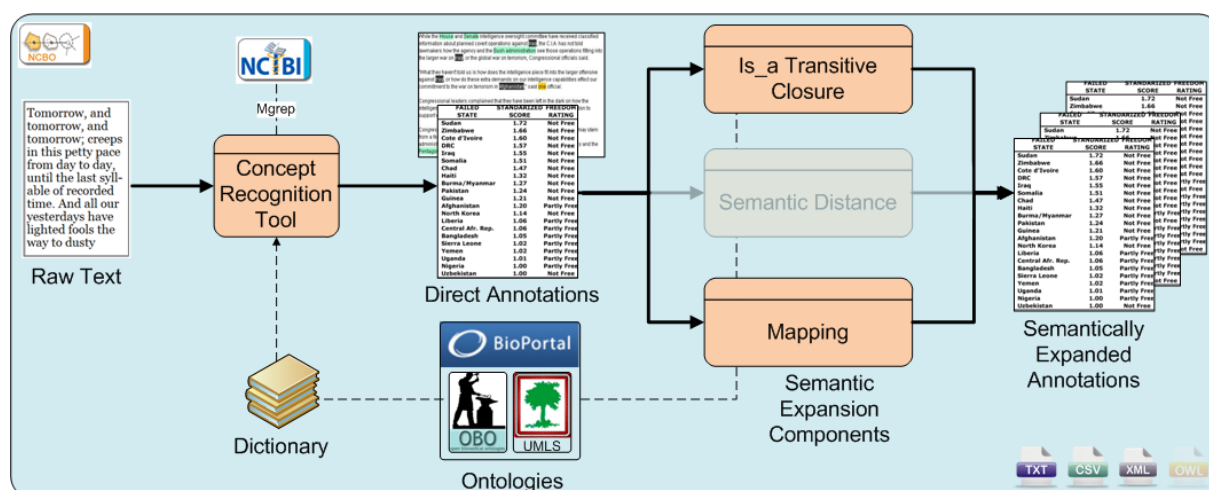


Figure 21. NCBO Annotator original workflow [CJ41].

Mgrep and/or the NCBO Annotator have been evaluated on different datasets [148, 154–156][CJ18] and usually perform very well in terms of precision e.g., 95% in recognizing disease names [157] but low on recall (due to the simple string matching strategy). A comparative evaluation of MetaMap [138] (the reference free text annotation tool for UMLS terminologies) and Mgrep within NCBO Annotator was made in 2009 [CJ14] when the NCBO Annotator was first released. Based on the results of our comparison, and because Mgrep had significantly faster execution time and can work with non-UMLS dictionary sources (which MetaMap cannot at this time), we decided to use Mgrep as the initial concept recognizer for building the NCBO Annotator. Although experiments have been carried out –both by NCBO and later by LIRMM– to swap the underlying concept recognizer with another (MetaMap, Alvis, Mallet, UniTex), Mgrep is still the default recognizer. It uses a simple label matching approach but offers a fast and reliable (precision) matching that enables its use in real-time high load web services.

When using the Annotator users can **adjust several parameters**: select the ontologies to use as well as the UMLS Semantic Types [158]; set up the service to recognize only the longest ontology term found in the text or recognize sub-terms as well (for example, recognize *breast cancer* vs. *breast* and *cancer*); enable or disable

matching based on synonyms terms and can also specify a minimum required term length; specify stop-words or use the default list of stop words; activate or disable the semantic expansion components as well as define the maximum parent level and the type of mappings used when expanding annotations.

In the original web service, **annotations returned had a score**, which was a number assigned to an annotation to indicate its importance. The scoring algorithm gave a weight to an annotation based on the kind of annotation (e.g., direct or expanded) as well as the type of the underlying matching term. Although this functionality was later removed from the native service on the NCBO BioPortal, we reoffered it through our NCBO Annotator+ presented Section IV.4.4 and inside the SIFR Annotator, thanks to a study made on scoring annotation results [CJ32].

In the three first years of its existence (2009-2011), the NCBO Annotator has been accessed approximately 90 million times and has processed approximately 700GB of data. It is the most used web service from the NCBO BioPortal (in number of API calls). At that time it was also embedded in commercial platforms such as Elsevier's SciVerse platform (<http://www.hub.sciverse.com>), Laboratree (<http://laboratree.org>) or Collabrx (<http://collabrx.com>). As **examples of published biomedical results, supported by the Annotator service** we can cite several use cases: (i) the use of a nanoparticle ontology to annotate a knowledge base for nanoparticles enables novel information retrieval queries [159]; (ii) rat gene expression data mining [154]; (iii) novel types of analyses that can associate classes of diseases with specific mutation types [160]; (iv) the use of annotation services enables morphology-based phylogenetic revisionary studies [155]; (v) mining of electronic health records [161, 162], mining for adverse drug events [163].

Several reviews or evaluations of biomedical named entity recognition or annotation tools have been made since 2009 [146–148, 164], and the NCBO Annotator is systematically included. We believe the NCBO Annotator in addition of its technical quality had two main advantages: (i) the web service allows end users to utilize ontologies for annotation of biomedical data with minimal effort; (ii) it is the most comprehensive annotation tool in terms of the diversity of ontologies available for use in the annotation task. **By making ontology-based annotation available for “plugging-in” to curation, data integration and annotation mining workflows, the NCBO Annotator has enabled a significant shift** in the way annotations are created and mined for biomedical research.

The three communications: poster/demo at ISWC 2009 [CJ102], article at AMIA Summit on translational bioinformatics [CJ41] and journal article at BMC Bioinformatics [CJ18] cumulate a total of 448 citations, respectively: 61, 254, 133.

#### IV.4.2 SIFR Annotator: a publicly accessible ontology-based annotation tool to process French biomedical text

Despite a wide adoption of English in science, a significant amount of biomedical data are produced in other languages, such as French. Yet, a majority of natural language processing or semantic tools as well as domain terminologies or ontologies are only available in English, and cannot be readily applied to other languages, due to fundamental linguistic differences. One of the main motivation to build the SIFR BioPortal was to design a semantic annotation workflow capable of processing French biomedical text.

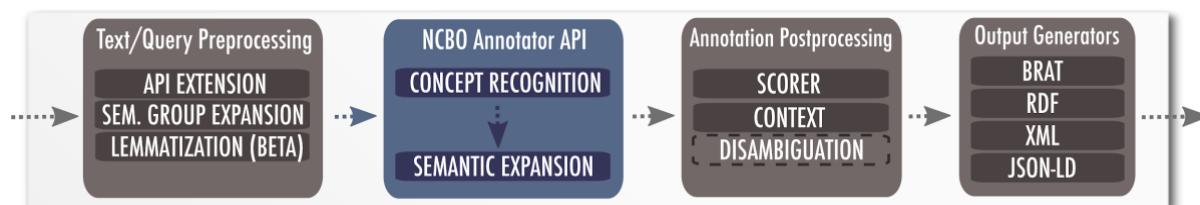
In [CJ65] then in [CJ2], we present the SIFR Annotator (<http://bioportal.lirmm.fr/annotator>), a publicly accessible and easily usable ontology-based annotation web service to process biomedical text and clinical notes in French. The annotator service processes raw textual descriptions, tags them with relevant biomedical ontology concepts, expands the annotations using the knowledge embedded in the ontologies and contextualizes the annotations before returning them to the users in several formats such as XML, JSON-LD, RDF or BRAT. **We have adapted the NCBO technology to French and significantly enhanced the original annotator [CJ2]**, including:

- *Cleaning dictionary heuristics*. To augment the SIFR Annotator's recall performance, we have implemented some heuristics to extend/clean the dictionary: Remove useless description at the end of concept labels e.g., "SAI", separate individual clauses from conjunctive sentences, normalize punctuation, remove parenthesized or bracketed precisions. Our experiments have shown that recall increases with such heuristics, while precision decreases, therefore the heuristics are currently deactivated by default.
- *UMLS Semantic Groups filtering*. The original NCBO Annotator offered the possibility to filter out the annotations by UMLS Semantic Types [158]. We have extended that functionality to UMLS Semantic Groups [165] which are a coarser-grained grouping of concepts. During query preprocessing, the Semantic

Group parameter is expanded into appropriate Semantic Types that are then handled by the original core Annotator components.<sup>22</sup>

- *Scoring.* When doing ontology-based indexing, the scoring and ranking of the results become crucial to distinguish the most relevant annotations within the input text. Higher scores reflect more important or relevant annotations. We have implemented and evaluated a new scoring method allowing to rank the annotations and enabling to use such scores for better indexing of the annotated data. By using a natural language processing-based term extraction measure, called C-Value [166], we were able to offer three relevant scoring algorithms which use frequencies of the matches and positively discriminate multi-words term annotations. This work is reported and evaluated in [CJ32].
- *Score filtering.* We have also implemented a thresholding feature that allows to prune annotations based on absolute or relative score values.
- *Lemmatization.* We have developed a beta lemmatization feature in the SIFR Annotator that is not yet properly evaluated but preliminary tests indicate that it would fix many morphosyntactic recognition errors. If lemmatization parameter is activated, then the text is being lemmatized by an external lemmatizer and send to a specific instance of Mgrep which runs with a lemmatized dictionary instead.
- *Clinical context detection.* This feature, which required significant research and evaluation is described in Section IV.4.3.
- *Additional output formats.* NCBO Annotator supports JSON-LD and XML outputs, but while JSON-LD is a recognized format, it is not sufficient for many annotation benchmarks and tasks, especially in the semantic web and natural language communities. SIFR Annotator adds support for standard linguistic annotation formats for annotation (BRAT and RDF) and task-specific output formats (e.g., CLEF eHealth/Quaero). The new output formats allow us to produce outputs compatible with evaluation campaigns and in turn to evaluate the SIFR Annotator. Moreover, they enable interoperability with various existing annotation standards/tools.

To generalize the features developed for French in the SIFR BioPortal to annotators in other BioPortal instances, we have **adopted a proxy architecture, that allows the implementation of features on top of the original REST API**, thereby extending the service by pre-processing inputs and post-processing outputs. Figure 22 describes the extended SIFR Annotator workflow, where the blue frame represents components from Figure 21.



**Figure 22. Proxy service architecture implementing the SIFR Annotator extended workflow [CJ2].** During preprocessing, parameters are handled, and text can be lemmatized, before both are sent to the core annotator components. During annotation postprocessing, scoring and context detection are performed. Subsequently, the output is serialized to the requested format.

Thanks to this proxy architecture, **some enhancements have not been implemented only for French but have been generalized for the original English NCBO Annotator (or any other Annotator based on NCBO technology) [CJ8]** as described Section IV.4.4. Especially, the new contextualization features, described Section IV.4.3, make SIFR Annotator the first general annotation workflow with a complete implementation of the ConText/NegEx algorithm [CJ-UR2] for French.

SIFR BioPortal, across all the ontologies indexed in the repository, currently represents **the largest open French-language biomedical dictionary/term repository**,<sup>23</sup> with over 330K concepts and around twice that number of terms. Enabling the SIFR Annotator service to use additional ontologies is as simple as uploading them to the portal (the indexing and dictionary generation are automatic) and take only a few minutes. The Annotator is

<sup>22</sup> For most of the six ontologies in the UMLS group, produced by CISMef in OWL format the relevant UMLS identifiers (CUI) and Semantic Type (TUI) were missing or improperly attached to the concepts. We therefore enriched them to reconcile their content with UMLS concepts and Semantic Type identifiers [CJ46]. For this, we used the set of previously reconciled multilingual mappings described in Section IV.2.2 and [CJ29].

<sup>23</sup> CISMef's HeTOP repository is larger, but the content is not accessible publicly.

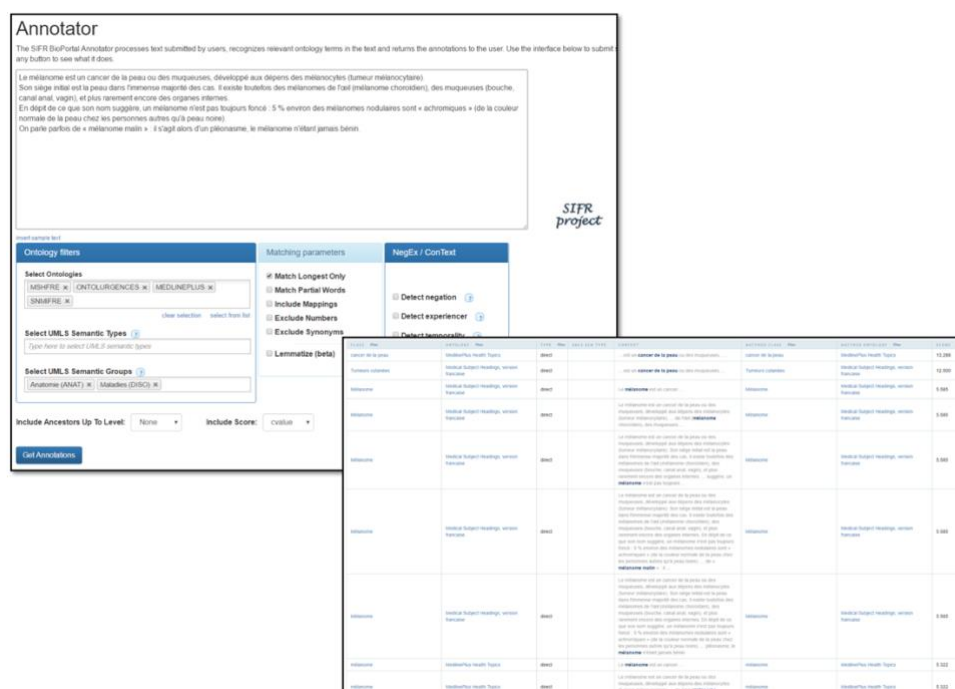
meant to be accessed through a REST API but there is also a user interface that serves as a demonstrator and that allows a full parametrization (cf. Figure 23).

A quantitative evaluation of annotation performance is of critical importance to enable comparison to other state-of-the-art annotation systems. A preliminary evaluation of the SIFR Annotator done in 2016 has shown that the web service matches the results of previously reported work in French, while being public, of easy access and use, and turned toward semantic web standards [CJ65]. However, the previous evaluation was limited in scope and new French benchmarks have since been published. Indeed, since 2015, the main venue for the evaluation of French biomedical annotation are the CLEF eHealth information extractions tasks [167–169]. In 2017, we participated to the CLEF eHealth campaign with both the SIFR Annotator for French and the NCBO Annotator for English. Our results are reported in [CJ48].

In [CJ2], we draw a more exhaustive evaluation of the SIFR Annotator and its new capabilities with the following corpora: (i) the Quaero corpus (from CLEF eHealth 2015 [170]) which includes French Medline citations (titles & abstracts) and drug labels from the European Medicines Agency, both annotated with UMLS Semantic Groups and Concept Unique Identifiers (CUIs); (ii) the CépiDC corpus (from CLEF eHealth 2017 [167]) which gathers French death certificates annotated with ICD-10 codes produced by the French epidemiological center for medical causes of death (CépiDC<sup>24</sup>). By evaluating and comparing the SIFR Annotator to state-of-the-art results, **we showed the web service performs comparably to other knowledge-based annotation approaches in recognizing entities in biomedical text and reach state-of-the-art levels in clinical context detection** (negation, experimenter, temporality).

Additionally, **the SIFR Annotator is the first openly accessible web tool to annotate and contextualize French biomedical text with ontology concepts leveraging a dictionary currently made of 28 terminologies and ontologies and 330K concepts**. The SIFR Annotator has significant other advantages that are not highlighted in the evaluation tasks. For instances, the ability to exploit the ontology hierarchy or mappings.

We believe that SIFR Annotator can help in a wide range of text mining or annotation problems, but of course not universally. In [CJ2], we have also highlighted the shortcomings of our SIFR Annotator and proposed some possible solutions for their mitigation in future technical evolutions of the service (e.g., disambiguation module). The code is openly available, and we also provide a Docker packaging for easy local deployment to process sensitive (e.g., clinical) data in-house (<https://github.com/sifrproject>).



**Figure 23. The SIFR Annotator user interface [CJ2].** The upper screen capture illustrates the main form of the annotator, where one inputs text and selects the annotation parameters. The lower screen capture shows the table with the resulting annotations.

<sup>24</sup> Centre d'épidémiologie sur les causes médicales de décès, Unité Inserm US10, <http://www.cepidc.inserm.fr>

### IV.4.3 Detecting negation, temporality and experienter in French clinical notes

In the context of the ANR PractiKPharma project, we **had to improve the SIFR Annotator to process clinical text**. Our use case, working with HEGP hospital (Paris) and LORIA (Nancy), is to extract pharmacogenomics knowledge from French electronic health records (EHRs) to compare them to state-of-the-art knowledge published in scientific articles and references databases (<http://practikpharma.loria.fr>).

EHRs often include unstructured elements (free text) that contain valuable information for medical research [171]. Researchers have developed systems to automatically detect clinical conditions and extract valuable knowledge in order to facilitate decision support [172], the identification of patients [173] and surveillance [174]. **When annotating clinical text, the context of the annotated clinical conditions is crucial:** distinguishing between affirmed and negated conditions (e.g., “no sign of cancer”); whether a condition pertains to the patient or to others (e.g., family members); or temporality (is a condition recent or historical). NegEx/ConText, is one of the best performing and fastest (open-source) algorithms for clinical context detection in English medical text [175, 176]. NegEx/ConText is based on lexical cues (trigger terms) that modify the default status of medical conditions appearing in their scope. For instance, by default the system considers a condition affirmed, and marks it as negated only if it appears under the scope of a negation trigger term. Each trigger term has a pre-defined scope either forward (e.g., “denies”) or backward (e.g., “is ruled out”), which ends by a colon or a termination term (e.g., “but”).

In [CJ63] then in [CJ-UR2], we present *French ConText*: an adaptation and enrichment of NegEx/ConText to the French language.<sup>25</sup> We compiled an extensive list of French lexical cues by automatic and manual translation and by enrichment. We integrated French ConText in SIFR Annotator, and thanks to the proxy architecture plugged the original ConText (for English) in the NCBO Annotator (cf. Section IV.4.4). We offer now, **both for English and French a unique open ontology-based annotation service that both recognize ontology concepts and contextualize them** allowing non-natural-language-processing experts to both annotate and contextualize medical conditions in clinical notes.

To evaluate French ConText, we manually annotated the context of medical conditions present in two types of clinical narratives: (i) death certificates and (ii) electronic health records.<sup>26</sup> We reported an evaluation of the SIFR Annotator with F1 scores between 83.7% & 86.3% for negated concepts (better by more than 5% of previously reported results adapting NegEx to French), F1 between 88.9% and 91.7% for the detection of historical entities and between 79.2% and 90.9% for concepts pertaining to an experienter other than the patient. The results are on-par with other state-of-the-art approaches (NegEx for negation, machine learning, etc.), independently from the concept recognition performance. Furthermore, **French ConText outperforms previously reported French systems for negation detection when compared on the same datasets and it is the first implementation of temporality and experienter identification reported for French**. This work is reported and evaluated in detail in [CJ63] but an extended English journal publication is currently under review [CJ-UR2].

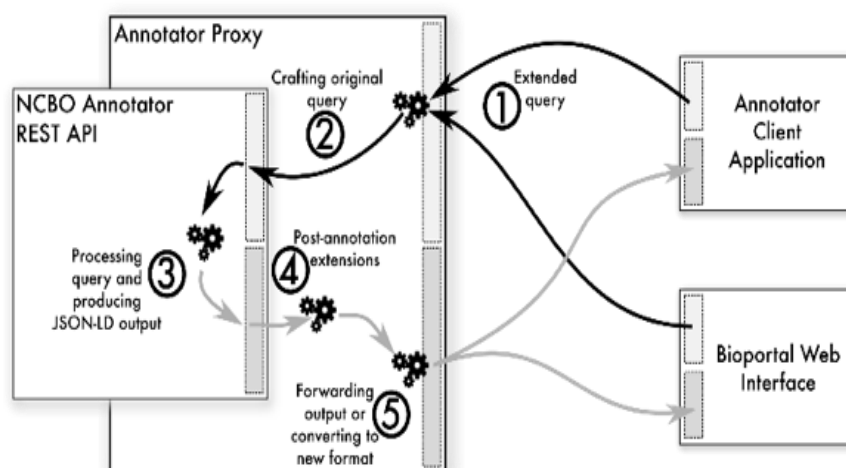
### IV.4.4 Annotating and indexing English clinical text with the NCBO Annotator+

In [CJ8], we present the proxy architecture –previously mentioned when presenting the SIFR Annotator– which allowed us to **add new functionalities to the NCBO Annotator without hosting or modifying the original web service**. Some of these new functionalities are particularly relevant to process electronic health records including annotation scoring, clinical context detection or coarse-grained entity type annotations. We present the NCBO Annotator+, a web service which incorporates these new functionalities, based on the proxy architecture illustrated Figure 24. The Annotator+ has been successfully integrated into the SIFR BioPortal platform to annotate English text. A web user interface is available for testing and ontology selection ([http://biportal.lirmm.fr/ncbo\\_annotatorplus](http://biportal.lirmm.fr/ncbo_annotatorplus)) however, the Annotator+ is meant to be used through the web service application programming interface.

---

<sup>25</sup> Although an implementation of NegEx was available for French [228], we extended it to the complete ConText algorithm.

<sup>26</sup> We evaluated French ConText on a sub-corpus of death certificates from the CLEF eHealth Task 1 corpus (6 sentences for experienter, 150 for temporality, 1030 for negation) and on a clinical corpus from the European Hospital Georges Pompidou (630 lines for experienter, 475 lines for temporality, and 400 lines for negation).



**Figure 24. NCBO Annotator+ proxy web service architecture [CJ8].** (1) requests are sent to the proxy with extended parameters that are parsed to select/apply the additional features; (2) a query is crafted for the original service without any extended parameters; (3) the original NCBO Annotator processes the query and returns the results; (4) the proxy retrieves annotations and applies post-processing/filtering (e.g., scoring); and finally, (5) the output is generated in the original format or in one of the new output formats from Annotator+.

We briefly report on the performance of the NCBO Annotator+ for: (1) annotating and contextualizing concepts in English clinical text on the CLEF eHealth 2017 task 1 corpus, created for the automatic annotation of death certificates with ICD-10 codes; (ii) the SemEval 2015 Task 14.2 development corpus, created for the identification of biomedical concepts (i.e., names and identifiers in UMLS) and of clinical context features (we covered negation and experienter).

When annotating the death certificates with the NCBO Annotator, we obtained median results compared to the rest of the competitors (cf. Table 6); **ahead of other knowledge-based systems but behind specifically tailored supervised learning systems**. The results are encouraging considering that we have not customized the service in any way for the task. We acknowledge the better performance of supervised learning approaches, but claim that in the health domain, they are often not applicable for lack of training data.

For the task of concept recognition in the SemEval corpus, the NCBO Annotator obtained average scores, given that we performed no adaptation to the task (and we did not use the training data at all), the concept recognition accuracy is fair (66.6%). We did not have access to the test gold standard and thus cannot compare to other participants (we ran on the development corpus). For negation, Annotator+ obtained state-of-the-art performance (balanced weighted average performance) and for experienter detection, we obtained results that are not substantially lower than existing evaluations of ConText [176]. These results confirm both the potential of the NCBO Annotator as a concept recognition service (never evaluated on standardized evaluation campaign tasks) and the nonreduced performance of NegEx/ConText when implemented in Annotator+.

**Table 6. Evaluation for concept recognition (NCBO Annotator) and clinical context detection (NCBO Annotator+) expressed by Precision, Recall, F-measure, Accuracy [CJ8].**

Task (Corpus)	P (%)	R (%)	F1 (%)	A (%)
Concept recognition (CLEF eHealth)	69.1	51.4	58.9	
Concept recognition (SemEval)	46.9	62.0	53.4	66.6
Negation detection (SemEval)	87.0	88.9	88.0	89.3
Experienter detection (SemEval)	52.9	70.4	60.4	52.7

#### IV.4.5 Terminology extraction and ontology enrichment (J-A. Lossio's PhD project)

During Juan-Antonio Lossio's PhD project (2012-2015), directed by Dr. Mathieu Roche (CIRAD) and Dr. Maguelonne Teisseire (IRSTEA), we investigated research issues related to automatic biomedical terminology extraction and sense induction for ontology enrichment [177]. The following is a summary of his contributions and accepted publications.

A few semi-automatic methodologies have been proposed for the construction/enrichment of ontologies from text. They are mostly achieved using natural language processing techniques to assess texts. Methods must take into account both the lexical and semantic complexity of biomedical data. Our first contribution in this area

concerns the automatic extraction of specialized biomedical terms (lexical complexity) from corpora. We focus here on terms that do not exist in an ontology/terminology, called new biomedical candidate terms. We proposed **a methodology based on linguistic, statistic, graph, and web features to improve the ranking of new biomedical candidate terms**. New ranking measures for single –and multi-word– term extraction methods are proposed and evaluated. In addition, we present **BioTex, an application that implements the proposed measures**. The second contribution concerns concept extraction and semantic linkage of the new extracted terms (semantic complexity). We **detect if a term is polysemic or not, then identify its possible senses** and induce the most relevant sense to attach the new candidate terms in an existing biomedical ontology or terminology. We experimented our approach with the MeSH terminology.

These two contributions, which represent useful feature for ontology developers, have not yet been incorporated within an ontology repository technology.

**Automatic biomedical term extraction.** The huge amount of biomedical data available today often consists of plain text. These texts are written using a specific language (expressions and terms) of the associated community. Therefore, there is a need for formalization and cataloging of these technical terms or concepts via the construction of terminologies and ontologies. These technical terms are also important for information retrieval, for instance when indexing documents or formulating queries. However, as the task of manually extracting terms of a domain is very long and cumbersome, researchers have striving to **design automatic methods to assist knowledge experts in the process of cataloging the terms and concepts** of a domain under the form of vocabularies, thesauri, terminologies or ontologies. **Automatic term extraction (ATE), aims to automatically extract technical terminology from a given text corpus** [178].

The main issues in ATE are: (i) extraction of non-valid terms (noise) or omission of terms with low frequency (silence), (ii) extraction of multi-word terms having various complex various structures, (iii) manual validation efforts of the candidate terms, and (iv) management of large-scale corpora. Recent studies have focused on multi-word (n-grams) and single-word (unigrams) term extraction. Techniques can be divided into four broad categories: linguistic, statistical, machine learning, and hybrid. Graph-based approaches have not yet been applied to ATE, although they have been successively adopted in other information retrieval fields and could be suitable for our purpose. Existing web techniques have not been applied to ATE but can be adapted for such purposes. We especially mention **C-value/NC-value**[166] **which combines statistical and linguistic information for the extraction of multi-word and nested terms**. This is the most well-known measure in the literature which obtained best results compared to other measures [179]. It has been used for recognizing terms in the biomedical literature [180] and applied to different languages other than English. We have been much inspired by C-value in our work and have chosen this measure as baseline. Tools and applications for biomedical term extraction include: TerMine (<http://www.nactem.ac.uk/software/termine>), Java ATE [179], FlexiTerm [181], BioYaTea [182], and our application, BioTex [CJ93] based on the methodology presented hereafter.

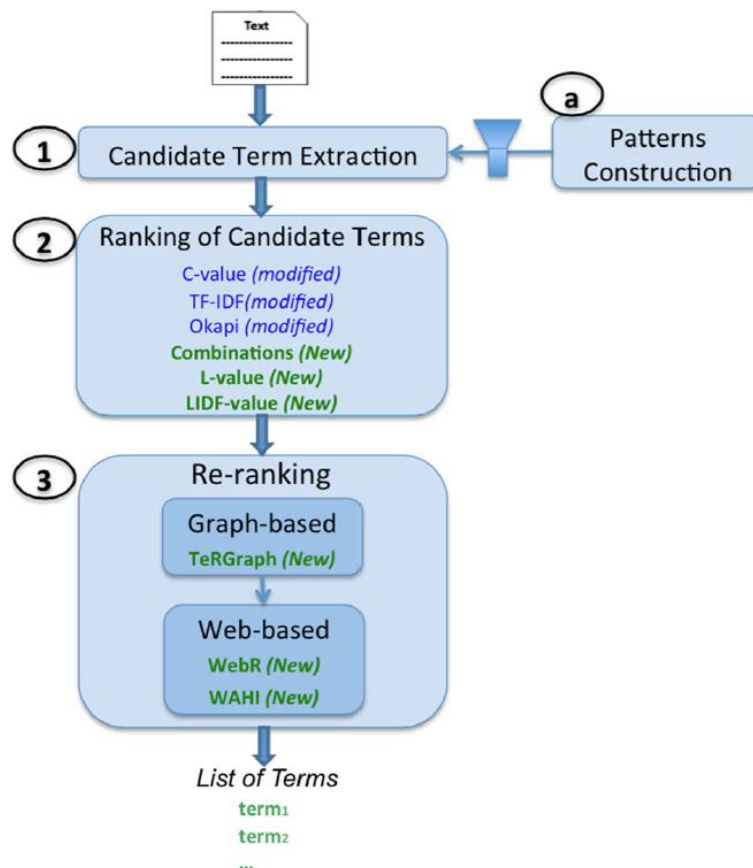
In [CJ13], we propose a **cutting-edge methodology to extract and to rank biomedical terms**, covering all the previously mentioned issues. This methodology offers several measures based on linguistic, statistical, graph and web aspects. These measures extract and rank candidate terms with excellent precision: we demonstrate that they outperform previously reported precision results for automatic term extraction, and work with different languages (English, French, and Spanish). We also demonstrate how the use of graphs and the web to assess the significance of a term candidate, enables us to outperform precision results.

Our methodology has three main steps (as illustrated Figure 25). One of our hypothesis (as of C-value's) is that biomedical terms have a similar syntactic structure (linguistic aspect). Therefore, in a preliminary step, using Part-of-Speech (POS) tagging,<sup>27</sup> we built a list of the most common linguistic patterns according to the syntactic structure of terms present in the UMLS. We will later use this list to filter out the content of our input corpus (also POS tagged) and retain only terms whose syntactic structure is in the patterns list.

We then propose new measures and some modifications of existing baseline measures to rank candidate terms. Our ranking measures are statistical- and linguistic-based. Consequently, to the application of these measures, we re-rank the results to increase the top k term precision using two new measures: **TeRGraph, a graph-based measure which exploits co-occurrence in sentences in the corpus; and WAHI, a web-based measure which uses web search to measure word associations**.

---

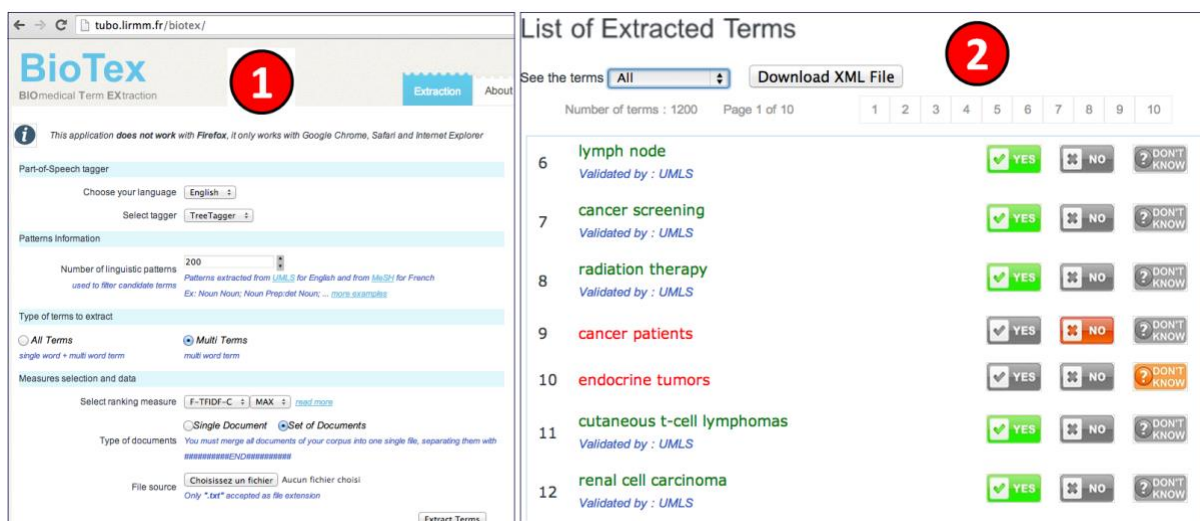
<sup>27</sup> Part-of-Speech (POS) tagging is the process of assigning each word in a text to its grammatical category (e.g. noun, adjective).



**Figure 25. Workflow methodology for biomedical term extraction [CJ13].** New proposed measures for ranking or re-ranking are in green, and already existing measure slightly modified are in blue. From a text corpus as input, the methodology returns a ranked list of new candidate terms.

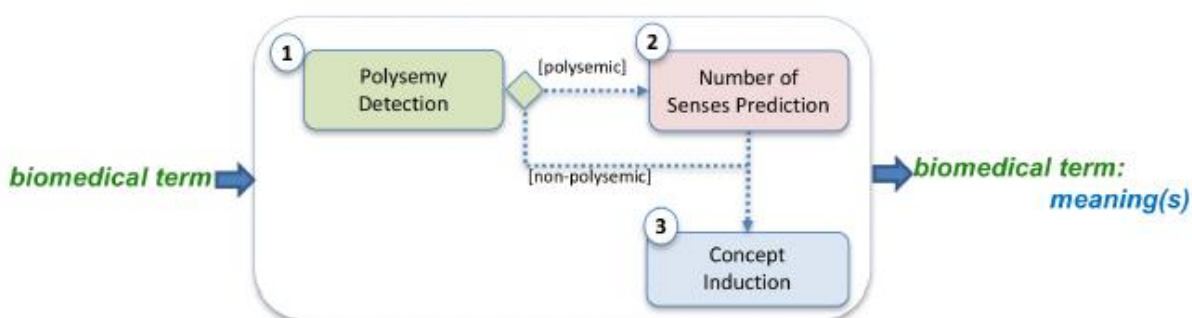
We evaluated our methodology on the biomedical GENIA ([www.geniaproject.org](http://www.geniaproject.org)) and LabTestsOnline (<https://labtestsonline.org>) corpora which are respectively made of titles/abstracts of journal articles and textual information for patients or family caregivers about clinical lab tests. We compared results of all the measures listed Figure 25 with previously reported measures. We experimentally showed that LIDF-value (based on the linguistic patterns, inverse document frequency and C-value) outperformed a state-of-the-art baseline for extracting terms while obtaining the best precision results in all intervals (i.e., P@k). With three languages the LIDF-value trends were similar. For all cases, our re-ranking measures improve the precision obtained with LIDF-value. WAHI (based on Yahoo Search) obtained better precision for the first P@100 extracted terms with 96 % precision.

In [CJ93], we presented **BioTex**, a web application (<http://tubo.lirmm.fr/biotex>) that implements state-of-the-art measures (including some of our new ones) for automatic extraction of biomedical terms from English and French free text. After the extraction process, BioTex automatically validates the extracted terms by using UMLS (Eng) & MeSH-fr (Fr). As illustrated in Figure 26 (2), these validated terms are displayed in green, specifying the used knowledge source used for validation and the others in red. Once validated the last ones may be considered candidates for ontology enrichment.



**Figure 26. BioTex user interface [CJ93].** During term extraction step (1), users upload their text corpus and select the measure to use; during term validation (2) users can manually validate the results not yet automatically validated (in red) and export the final candidate terms.

Sense induction and ontology enrichment. Once a term has been extracted from a corpus, we need to identify the concept (sense) behind to suggest a relevant position in the ontology to enrich. **Word-sense induction (WSI) is the task of automatically inducing the different senses of a word in a piece of text.** Most existing WSI approaches are based on unsupervised machine learning with senses represented as clusters of tokens (e.g., words or phrases). In general, existing WSI approaches only consider sense induction for individual words, such as verbs, nouns, and adjectives [183, 184]. However, biomedical terms are often composed of more than one word—80% of UMLS terms are composed of two or more words. Another issue with existing WSI methods is that they do not first check whether a target word is polysemic (i.e., ambiguous) or not. Thus, a significant amount of computing time is wasted on identifying the different senses for non-polysemic words. In addition, clustering algorithms used to predict the number of senses often suffer from poor performance [185]. To address these challenges associated with applying WSI in biomedicine, we proposed in [CJ89] then in [CJ7], a complete workflow for automatically enriching biomedical ontologies or terminologies starting with executing BioTex to extract candidate terms and continuing with three steps to induce the concept senses described in Figure 27.



**Figure 27. Methodology for biomedical entity sense induction [CJ7].**

In [CJ30], we focused on polysemy detection to predict if candidate terms are polysemic or not. **We introduced 23 new features for machine-learning-based polysemy detection** extracted directly from text (11) and from a cooccurrence graph itself produced from the text corpus (12). For examples, the number/min/max of UMLS terms contained in the set of abstracts obtained in PubMed with the candidate term or the number of neighbors in the co-occurrence graph. We also used two terminology resources: UMLS (i.e., biomedical) and AGROVOC (i.e., agronomy) to derive these features. These two thesauri have a certain degree of overlapping concepts, which can be considered as polysemic entities that belong to both biomedical and agricultural domains. For instance, the term “cold” can represent either a disease (i.e., the common cold) or the feeling of no warmth in UMLS, as well as the temperature of the weather in AGROVOC. Thus, we hypothesized that candidate terms (that did not appear in these two thesauri) that co-occurred with existing polysemic ones were more likely to be polysemic as well. We implemented the method for polysemy detection using multiple supervised machine learning

algorithms from Weka [126] and experimented with a standard corpus of polysemic terms –the MSH WSD<sup>28</sup> dataset, which consists of 203 ambiguous entities. Our method showed an F-measure of 98%.

In [CJ89], we also drafted **methods to predict the number of senses and induce the concept**. We suggested to use clustering algorithms for the first issue (e.g., CLUTO<sup>29</sup>) and our method achieved an F-measure of 93%. Then we experimented the second issue by automatically enriching the 2009 version of MeSH terminology with 60 terms that have been added by experts to MeSH between 2009 and 2015. We showed that 50% of our predictions for positioning the new term in the terminology were correct.

Finally, in [CJ7], we have consolidated this work by completing the methodology with concept induction, compared our methods with others and strengthen our evaluation.

#### IV.4.6 MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums

As a parallel work on terminology extraction, we have studied the extraction of lay user vocabulary out of web forums. With the explosion of Web 2.0 and social medias, doctors have definitively realized the enormous potential of data generated by patients [186]. According to a 2011 Health On the Net Foundation survey [187], the web has become the second source of information for patients after consultations with a doctor. **Semantically analyze patient-generated text from a biomedical perspective is challenging because of the vocabulary gap between patients and health professionals.** Indeed, health consumers, i.e., patients, are generally laypersons who do not have the technical or scientific expertise and hence expressions and vocabulary [188]. Laypersons use abbreviations, misspellings, neologisms or existing words that are diverted from their standard professional use. The medical expertise and vocabulary are well formalized in standards terminologies and ontologies, which enable semantic analysis of expert generated text; however, resources which formalize the vocabulary of health consumers (patients and their family, laypersons in general) remain scarce. The situation is even worse if one is interested in another language than English.

Many researchers have been working to **reduce this vocabulary gap between laypersons and health care professionals** by identifying CHV constituents and/or mapping them to their equivalents in the standard biomedical semantic resources [189, 190]. However, these efforts do not often result in reusable open access resources. Indeed, one of the only freely available CHV is the (English) Open-Access and Collaborative CHV (included in the UMLS Metathesaurus) that was developed by Univ. of Utah and recently updated by mining social network data [191].

Semantically representing CHVs' content and using them inside forum applications will enhance the patient's access to information by connecting the formal medical expertise to the actual content of the forums, inside the forums. It would also enable to process semantically patient-generated text. For instance, topics discussed will be more easily mined in order to identify what are the principal concerns of the patients[192]; forum providers would be able to connect their users to reference data resources that are indexed with standard medical terminologies but that are targeted for patients e.g., MedLinePlus.

In [193]s, the authors focused on a methodology to extract a preliminary CHV out of forum patient posts, & Facebook groups about breast cancer. In [CJ26], we built-up on this work and presented **a concrete machine-readable formalization of the extracted vocabulary, the provenance information and the alignment to standard terminologies, using the semantic Web languages (RDF, SKOS and PROV)** as illustrated Figure 28. We used a sample of 173 relations built around 64 expert concepts which have been automatically (89%) or manually (11%) aligned to standard biomedical terminologies, in our case: MeSH, MedDRA and SNOMEDint. The resulting vocabulary, called MuEVo (Multi-Expertise Vocabulary) and the mappings are publicly available in the SIFR BioPortal French biomedical ontology repository (<http://biportal.lirmm.fr/ontologies/MUEVO>).

---

<sup>28</sup> <https://wds.nlm.nih.gov/collaboration.shtml>

<sup>29</sup> <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>

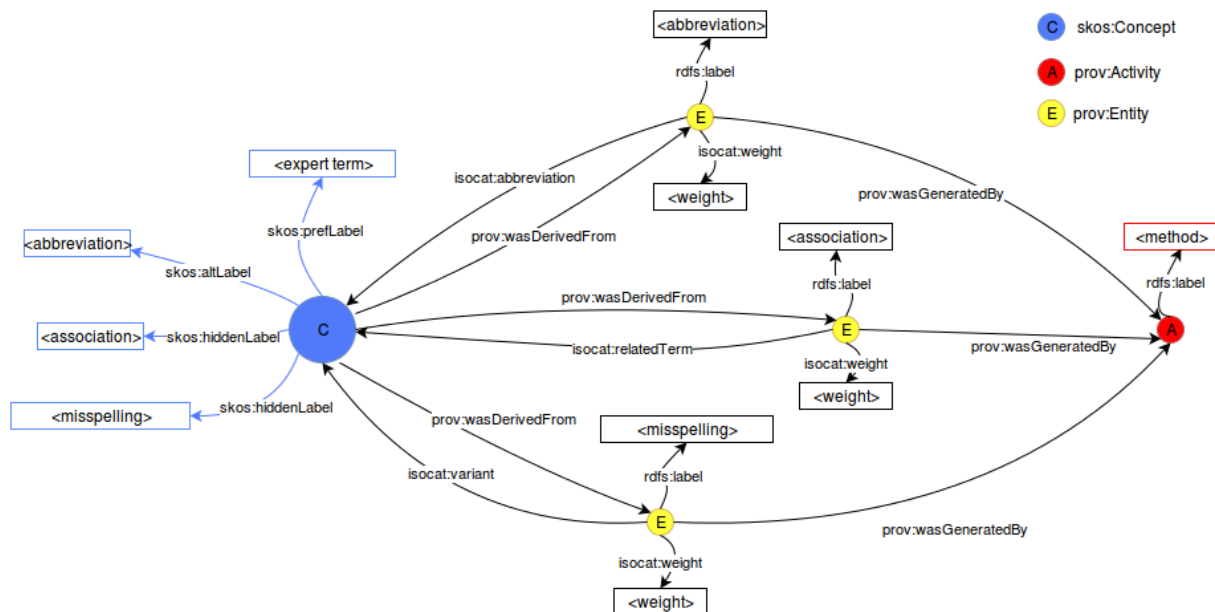


Figure 28. Model to formalize lay-expert relations in MuEVo using SKOS+PROV [CJ26].

In this preliminary study, we focused on breast cancer and French language, but our model is generalizable to other domain or language. Although the size of the current vocabulary is quite small, this is the result of an automatic process that could be reproduced on other datasets to augment it.

## IV.5 Challenge 5: Annotations and linked data

Datasets produced in science are highly heterogeneous: they are stored and accessible in many different databases, using idiosyncratic schemas and access mechanisms. For instance, a researcher studying allelic variations in a gene can find all the pathways that the gene affects, the drug effects that these variations modulate, any disease that could be caused by the gene, and the clinical trials that involve the drug or diseases related to that specific gene. The information that we need to answer such questions is available in public biomedical resources; the problem is finding that information.

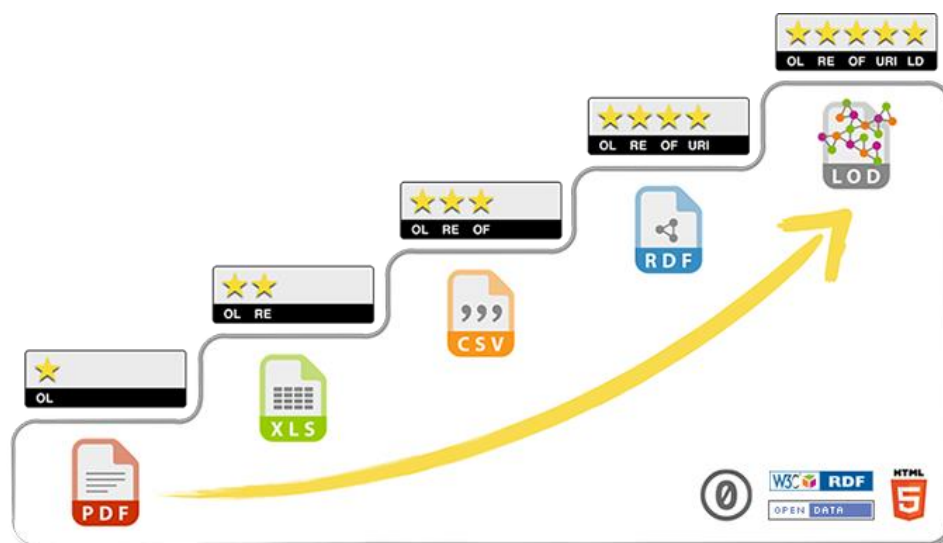
Data integration and semantic interoperability enable **new scientific discoveries that could be made by merging different currently available data**. This is one major reason for adopting ontologies. Ontologies can be used to search, mine and analyze uniformly the information stored in these diverse resources. They are used to design semantic indexes [194, 195] of data and linked open datasets [196–198] that could be used for various type of cross datasets studies.

**Ontology repositories must facilitate indexing/annotation, search and access to semantically described, interoperable, linked open data either directly from within the repositories or via uniform automatic access to ontologies.**

We have seen in presenting the preceding challenge possible indexing/annotation tool; this challenge is about exploiting those tools to process big data sources and turn them into knowledge. When building semantic indexes, big data represents a set of challenges for ontology repositories: **scalability, consistency, completeness in a context where both ontologies and data constantly evolve**. For example, indexed data consistency shall be checked by ontology repositories using OWL reasoning. Also, a semantic index –or the annotations– shall perpetually be updated by: (i) indexing the new data elements as they arrived (or re-index the ones that have changed) with all the ontologies necessary; (ii) indexing all the content of the data resource with every new ontology as they arrived in the repository; (iii) refresh the index to reflect changes in ontologies (removed/added concepts). See for example the work done in [199] on the evolution of ontologies and annotations.

For about 10 years, the *Linked Open Data* approach [10] has been adopted to represent data with semantic web technologies and indexing them with ontologies/vocabularies. The web of data –built out of linked open data– is the concrete and most salient outcome of 20 years of semantic web research. It is well represented by the

linked open data diagram (Figure 3, page 22). **Ontologies and vocabularies are the backbone of linked open data as they are used to semantically annotate and interlink datasets.** Methods and techniques have recently been developed, allowing the massive publication of structured data on the web. The general principles were established by Tim Berners-Lee –the inventor of both the web and the semantic web– as illustrated in Figure 29.



**Figure 29. Producing five-star linked data as suggested by T. Berners-Lee** (source: <https://5stardata.info>).  
 (\*) make your stuff available on the Web (whatever format) under an open license; (\*\*) make it available as structured data (e.g., Excel instead of image scan of a table); (\*\*\*) make it available in a non-proprietary open format (e.g., CSV instead of Excel); (\*\*\*\*) use URIs to denote things, so that people can point at your stuff; (\*\*\*\*\*) link your data to other data to provide context.

In the recent years, the biomedical community has strongly embraced the semantic web vision as demonstrated by a number of initiatives to use ontologies for producing semantically rich data such as in Bio2RDF [196, 200], OpenPHACTS [198], Linked Life Data [201], KUPKB [202], and the EBI RDF Platform [197]. For example, OpenPHACTS serves as a good example of what can be achieved by using semantic web knowledge bases: the explorer provides use case driven tools that aid in browsing and visualizing the underlying knowledge represented in RDF which is very convenient for biologists.

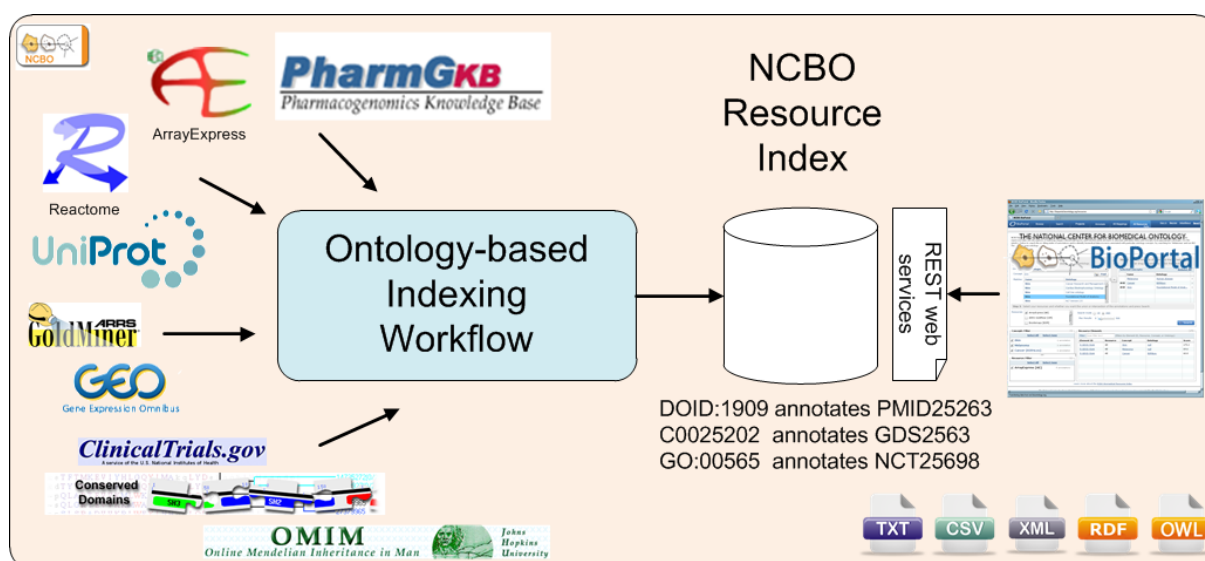
We believe that **ontology repositories have a role to play in building the web of data as they are the infrastructure hosting and serving the ontologies and vocabularies.** In this section, we will present some previous work done before the emergence of big data technologies and linked open data principles when building the NCBO Resource Index, a large-scale ontology-based index of more than 50 heterogeneous biomedical resources, integrated within the NCBO BioPortal. Then, we will quickly present other related work –on which we are involved but not as a primary actor – on using ontologies to build knowledge bases: (i) the AgroLD project which builds a database of agronomy resources described in RDF and annotated with ontologies; (ii) the PGxLOD knowledge base which offers linked open datasets for exploring and assessing pharmacogenomics knowledge. These two efforts have been respectively developed within the AgroLD and PractiKPharma projects. Finally, as an alternative to using ontologies, we will introduce our contribution in the ViewpointS project, an exploratory research designing a brain-inspired knowledge representation approach where semantic and social web contributions are merged into an adaptive knowledge graph.

#### IV.5.1 NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources

Researchers in biomedicine produce and publish enormous amounts of data describing everything from genomic information and pathways to drug descriptions, clinical trials, and diseases. The biomedical research community agrees that terminologies and ontologies are essential for data integration and translational discoveries to occur [43, 45, 131]. However, the metadata that describe the information in data resources are usually unstructured, often come in the form of free-text descriptions and are rarely labelled or tagged using terms from ontologies that are available for the domains. Users often prefer labels from ontologies because they provide a clear point of reference during their search and mining tasks [128, 203]. Semantic annotation of biomedical resources is still minimal and is often restricted to a few resources and a few ontologies as discussed in the preceding section. Usually, the textual content of these online resources is indexed (e.g., using Lucene) to enable

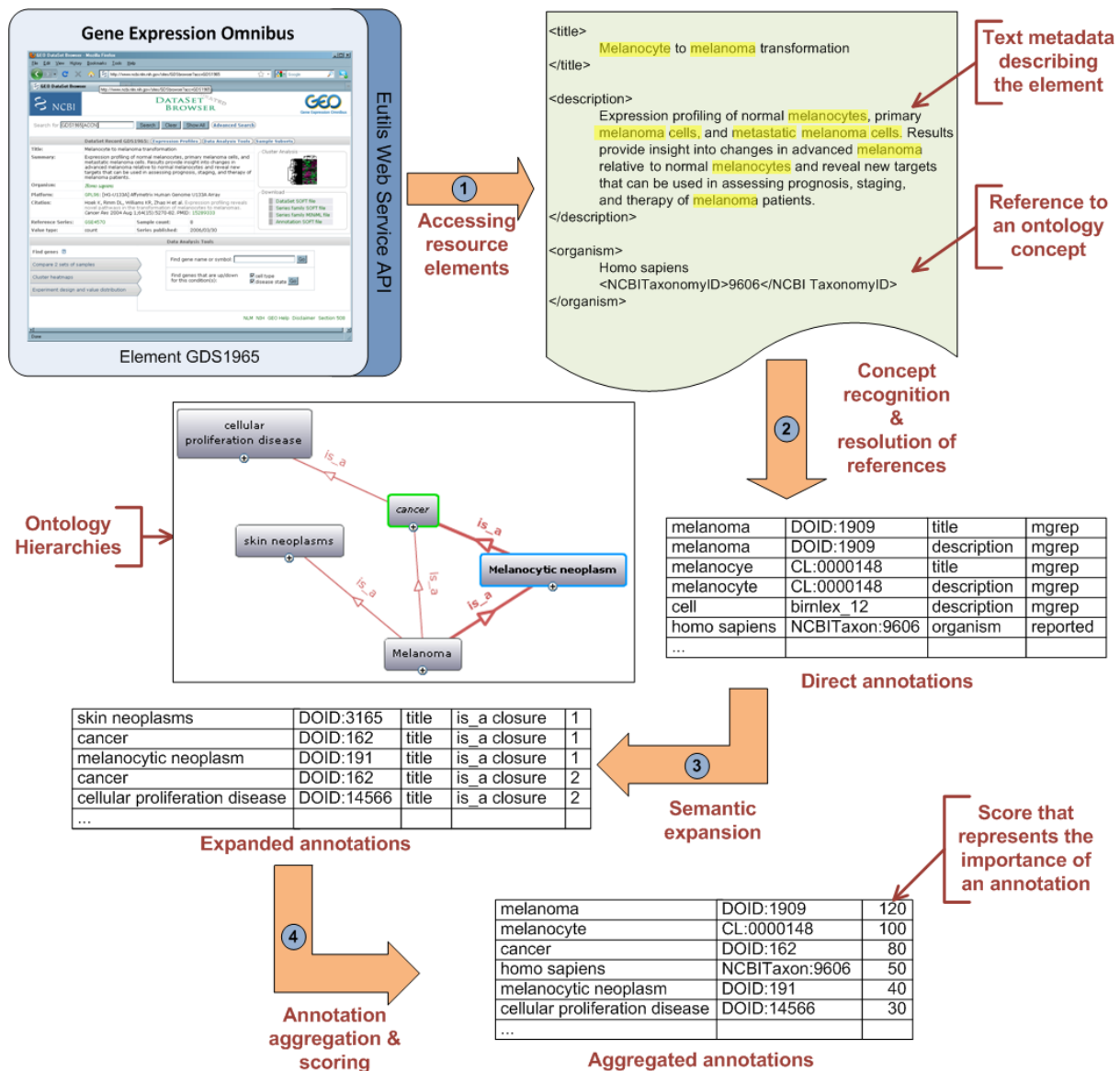
querying the resources with keywords. However, there are **obvious limits to keyword-based indexing, such as the use of synonyms, polysemy, lack of domain knowledge**. Furthermore, having to perform keyword searches at each web site individually makes the navigation and aggregation of the available information extremely cumbersome, if not impractical. Search engines, like Entrez ([www.ncbi.nlm.nih.gov/Entrez](http://www.ncbi.nlm.nih.gov/Entrez)), facilitate search across several resources, but they do not currently use as many of the available and relevant biomedical ontologies.

In [CJ15], based on our preliminary work described in [CJ42][CJ20], we have built the NCBO Resource Index, an ontology-based index of more than twenty heterogeneous biomedical resources (later extended to 50) included within BioPortal (<http://bioportal.bioontology.org/resources>). The resources came from a variety of data repositories maintained by organizations from around the world. They included gene expression datasets (Gene Expression Omnibus, ArrayExpress), clinical report descriptions (ClinicalTrials.gov), scientific literature (PubMed), proteins (UniProt KB), etc.



**Figure 30. NCBO Resource Index overview [CJ15].** We process each biomedical resource using the ontology-based indexing workflow (NCBO Annotator). We store the resulting annotations in a database and make them available in several formats via REST web services. BioPortal provides user friendly interfaces to search and navigate the Resource Index.

Semantic indexing in the Resource Index relied on the NCBO Annotator’s workflow presented Section IV.4.1. We use the terms from BioPortal ontologies to annotate, or “tag,” the textual descriptions of the data elements that reside in biomedical resources and we collect these annotations in a searchable and scalable index (cf. Figure 30). We used the **semantics that the ontologies encode**, such as different properties of classes, the class hierarchies, and the mappings between ontologies, in order **to improve the search experience** of the Resource Index user. The indexing workflow of a data element taken from the Gene Expression Omnibus database is illustrated Figure 31. When browsing the ontologies in the NCBO BioPortal, or using a dedicated search engine, users can discover datasets of interest. Our user interface within BioPortal enables scientists to search the multiple resources quickly and efficiently using domain terms, without even being aware that there is semantics “under the hood.”



**Figure 31. Example of annotations generated for a GEO element in the NCBO Resource Index [CJ15].** Direct annotations are generated from textual metadata and already existing ontology references of the data element. Then, expanded annotations are created using the ontology is-a hierarchy. Finally, all the annotations are aggregated and scored taking into consideration their frequency and context.

Ontology-based indexing was not new in biomedicine; however, it was usually restricted to indexing a specific resource with a specific ontology (vertical approach). We adopted a horizontal approach, enabling annotations of many important resources using a large number of ontologies. When the NCBO Resource Index was released in 2010, it included 22 resources, and more than 200 ontologies included in BioPortal. It was made of a 1.5Tb MySQL database, which stores the 11 Billion annotations of 3.2 Million data elements with 3.3 Million ontology concepts. It was later extended to more than 50 databases by the NCBO.

The Resource Index developed in 2008 did not rely on big data technologies and did not followed linked open data principles; both were in their infancies at that time. However, the underlying challenges of scalability and use of ontologies were already here. We would certainly implement it in a completely different way today. Later, in [CJ38], we have analyzed the metrics on ontologies in order to re-structure the database backend for the Resource Index. This restructuring has enabled us to reduce the indexing processing time for one of our larger datasets from one week to one hour. Although the NCBO Resource Index is still available within the NCBO BioPortal, it is not maintained and updated anymore since a few years. The impact of the application is rather difficult to evaluate considering the speed of change of science and technology in that area. Even if not explicitly used anymore, the NCBO Resource Index proposed an interesting effort illustrating the challenges of ontology-

based indexing at large scale as acknowledged by the 1<sup>st</sup> price at the *Semantic Web Challenge* organized at ISWC 2010. The four publications on the topic [CJ15, CJ20, CJ38, CJ42] gather today more than 310 citations total.

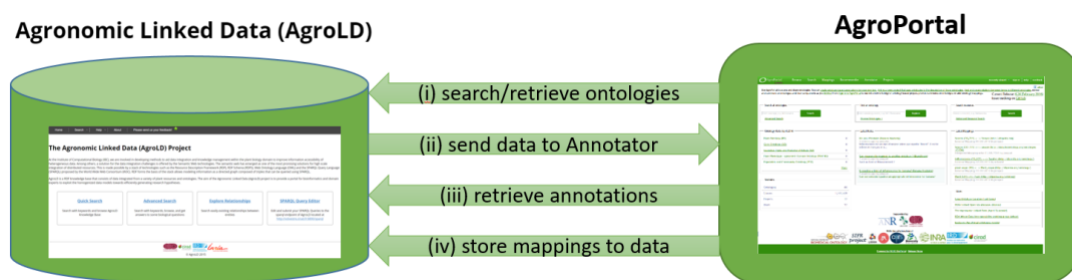
When building AgroPortal and the SIFR BioPortal, we have not included the “Resource Index components” in our ontology repositories. Our vision was to adopt another approach and set of technologies to play that role, as presented in the next two sections.

#### IV.5.2 AgroLD, an RDF knowledge base for agronomy

More recently, in agronomy, we have been involved in the AgroLD (Agronomic Linked Data) project which objective is to build a knowledge-based system relying on semantic web technologies and exploiting standard domain ontologies, to **integrate data about plant species of high interest for the plant science community** e.g., rice, wheat, arabidopsis. AgroLD’s goal is to offer a domain specific knowledge platform to solve complex biological and agronomical questions related to the implication of genes/proteins in, for instances, plant disease resistance or high yield traits.

In [CJ3], we present integration results of the first phase of the project, which focused on genomics, proteomics and phenomics. **AgroLD ([www.agrold.org](http://www.agrold.org)) is now an RDF knowledge base of 100M triples created by annotating and integrating more than 50 datasets coming from 10 data sources with 10 ontologies.** AgroLD offers information on genes, proteins, gene ontology associations, homology predictions, metabolic pathways, plant traits, and germplasm, on the following species: rice, wheat, arabidopsis, sorghum and maize. It provides integrated agronomic data, as well as the infrastructure to aid domain experts answering relevant biological questions. Original database contents were parsed and converted into RDF using a semi-automated pipeline implemented in Python.<sup>30</sup>

The conceptual framework for knowledge in AgroLD is based on well-established ontologies in plant sciences such as Gene Ontology, Sequence Ontology, Plant Ontology, Crop Ontology and Plant Environment Ontology. AgroLD needed a dedicated application programming interface to these ontologies, as well as a means to annotate database fields (header and values) with ontology concepts. In addition, it requires a system to store mappings annotations between key entities in the AgroLD knowledge base and reference ontologies. When building AgroLD, AgroPortal was used to retrieve ontologies (it was convenient to find them all in one place, and to use a unique and consistent API). Plus, we also **used the AgroPortal Annotator web service to annotate more than 50 datasets and produced 22% additional triples**, which were validated manually. Building such an annotation service for all these ontologies was one of the driving needs for AgroPortal since the very beginning. Finally, AgroPortal is also used to store annotations/mappings between high level concepts created in AgroLD and references ontologies. The interaction between AgroPortal and AgroLD are summarized in Figure 32. In the long-term vision for AgroPortal and AgroLD, the former might be an entry point to the knowledge stored in the latter, enabling users to easily query and locate data annotated with ontologies.



**Figure 32. Interaction between AgroPortal and AgroLD [CJ10].** (i) AgroPortal provides a unique endpoint to retrieve heterogenous ontologies; (ii) AgroLD’s annotation pipeline send data to the AgroPortal Annotator and (iii) retrieves annotations with ontology terms used to build AgroLD; finally (iv) AgroPortal offers a link from the ontologies to data stored in AgroLD with the ‘inter portal’ mapping mechanism.

#### IV.5.3 Pharmacogenomics Linked Open Data (PGxLOD)

Pharmacogenomics (PGx) studies how individual gene variations cause variability in drug responses. Knowledge in PGx is typically composed of units that have the form of ternary relationships *gene variant–drug–adverse event* –stating that an adverse event may occur for patients having the gene variant when being exposed to the drug– and can be formalized to different extents using biomedical ontologies. Most of state-of-the-art knowledge in

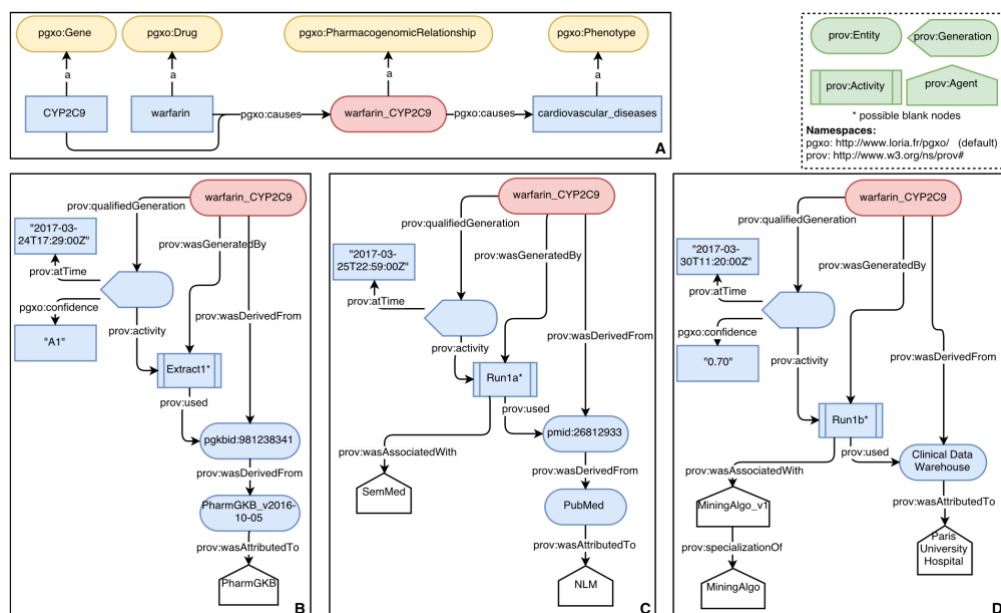
<sup>30</sup> <https://github.com/SouthGreenPlatform/AgroLD>.

PGx is not yet validated, consequently not yet applicable to medicine. During the PractiKPharma project (*Practice-based evidences for actioning Knowledge in Pharmacogenomics*) our objective is to **validate or moderate pharmacogenomics state-of-the-art knowledge on the basis of practice-based evidences**, i.e., knowledge extracted from EHRs. To achieve our goal, we extract state-of-the-art knowledge from PGx databases (i.e., PharmGKB) and literature (i.e., PubMed), and we extract observational knowledge from clinical data (in partnership with HEGP hospital); then we **compare knowledge units extracted from these two origins, to confirm or moderate state-of-the-art knowledge**, with the goal of enabling personalized medicine –a medicine tailored to each patient by considering in particular her/his genomic context.

PGx knowledge units available in reference databases, reported in the scientific biomedical literature and discovered by mining clinical data are heterogeneously described (i.e., with various quality, granularity, vocabulary, etc.). These knowledge units are also increasing: 40,000 PGx relationships were extracted from the 17,000,000 abstracts available on PubMed in 2008 [204] (there are now 27,000,000 abstracts available). It is consequently worth to extract, then compare, assertions from distinct resources. In [CJ49], we present a **lightweight and simple ontology named PGxO, that we developed to reconcile and trace knowledge in pharmacogenomics**. We captured the essential elements that constitute PGx knowledge and mapped them to existing standard ontologies. We also encode the provenance of the units using the PROV Ontology (PROV-O) [205]. An example of use of PGxO is illustrated in Figure 33.

Because PGxO's aim is to potentially represent multiple provenances for a unique PGx relationship, we defined a set of rules that, when satisfied, enable to decide when two PGx relationships with distinct provenances are in fact referring to the same knowledge unit.

By adopting PGxO and defining strict rules for its instantiation, we set up a first step toward a complete framework for PGx knowledge comparison. In [CJ1], we refined this work (especially the rules) and experimented our ontology and our proposed encoding for provenance information by **populating PGxO with data extracted automatically from PharmGKB (the reference PGx database) and the literature (PubMed)**, and manually from discoveries made from patient data studies. We called PGxLOD (Pharmacogenomic Linked Open Data) the resulting knowledge base that represents and reconciles knowledge units of those various origins. Our set of PGx linked data is available at <https://pgxlod.loria.fr>.<sup>31</sup> We believe PGxLOD will constitute a valuable community resource for PGx research. This work is still in progress.



**Figure 33. Example of instantiation of PGxO with a relationship (warfarin\_CYP2C9) and three distinct provenances. [CJ49].** Frame A represents the pharmacogenomic relationship. Frames B, C and D represent the three distinct provenances, respectively from PharmGKB, literature and EHRs. In these frames, the shape of the nodes refers to the type of PROV-O concepts they are instance of. Numeric IDs in B and C correspond to the PharmGKB annotation identifier and to the PubMed identifier used to extract the PGx relationship.

<sup>31</sup> The access is restricted as the data set contains some licensed PharmGKB data. An account will be provided upon request to users who have been granted a PharmGKB license.

#### IV.5.4 ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph (G. Surroca's PhD project)

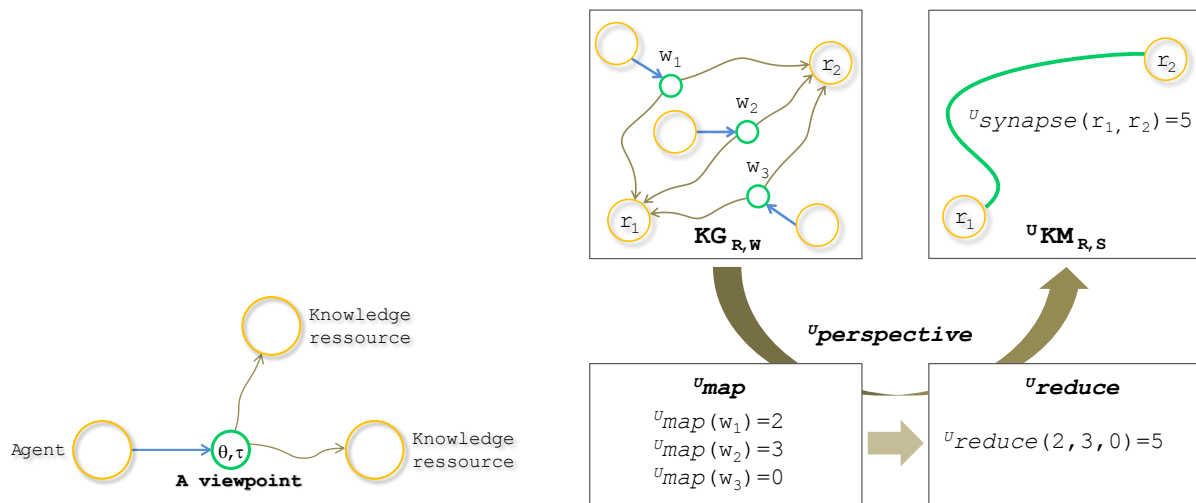
During Guillaume Surroca's PhD project (2013-2017), directed by Pr. Stefano Cerri (and co-supervised by Philippe Lemoisson (CIRAD)), we investigated an alternative way to represent knowledge called ViewpointS [206]. The following is a summary of the contributions and accepted publications on this project.

Formal data is supported by means of specific languages from which the syntax and semantics have to be mastered, which represents an obstacle for collective intelligence. In contrast, informal knowledge relies on weak/ambiguous contributions e.g., *I like*. Reconciling the two forms of knowledge is a big challenge. **The web explicitly exposes these two kinds of content:** with the Web 2.0, the *social web*, has democratized the sharing, recommendation and creation of content via social networks, blogs and fora; the *semantic web* offers to structure the knowledge deposited, generated and stored on the web. These types of content differ in the ways they are produced and structured. On one hand, contribution-based social web platforms allow the production of a wealth of data with little or no structure; these data evolve rapidly (e.g., folksonomies [207]). On the other hand, highly structured knowledge is constituted consensually by circles of experts (e.g., ontologies [208] or linked data [10]). Within this work our objective was to **create a knowledge representation formalism that retains the best qualities of each type of content and gives value to both** (i) the structure which characterizes semantic web datasets and (ii) the evolution and maintenance rates of shared knowledge on the social web as proposed in Gruber [209], [210] or surveyed in [211, 212].

We proposed a brain-inspired knowledge representation approach called *ViewpointS* where **formal data and informal contributions are merged into an adaptive knowledge graph** which is then topologically, rather than logically, explored and assessed. ViewpointS relies on three assumptions:

1. A *viewpoint* is a subjective connection between two objective knowledge resources; the aggregation of these connections between two given resources can be viewed as a synapse between two neurons;
2. The *knowledge graph* or "associative memory" formed by all the viewpoints (formal versus informal, proactive versus reactive) is a selectionist system evolving continuously according to user's interactions in the metaphor of a collective brain: each interaction is equivalent to the tuning of a synapse;
3. The topological structures which appear when adopting a user's perspective actualize assessable knowledge.

Rather than undertaking logical assessment (incompatible with informal content), we define *perspectives* (sets of quantification rules tuned to the interpretation of the user's context, which apply on the viewpoints) and topologically explore the *knowledge map* resulting from such an interpreted knowledge graph (Figure 34).



**Figure 34. The ViewpointS approach [CJ11].** A subjective connection called "viewpoint" is represented on the left; the blue arrow gives the provenance, ' $\theta$ ' gives the semantics, ' $\tau$ ' gives the time stamp. The right part illustrates the building of a knowledge map from the knowledge graph made of viewpoints.

The ViewpointS approach has evolved several times during the PhD project and has enabled us to demonstrate several interesting aspects summarized hereafter:

- In [CJ74], we showed how ViewpointS allows the **search and discovery of knowledge** through a search engine prototype for scientific publications (developed with HAL-LIRMM data).
- In [CJ69] and [CJ31], we wished to capture the phenomenon of Serendipity (i.e., incidental learning) using ViewpointS. To that effect, we built a simulation to study the **dissemination of knowledge** (with linked data and user contributions), similar to how the way the web is formed. Using a behavioural model configured to represent various web navigation strategies, we sought to optimize the distribution of preference systems. Our results outlined the most appropriate **strategies for incidental learning**, bringing us closer to understanding and modelling the processes involved in Serendipity.
- In [CJ28], we benchmarked the ViewpointS approach against other classic semantic distances (graph based or information content based) on a WordNet experiment. Our goal was to demonstrate the value of keeping the subjectivity of the represented knowledge, while having a generic approach that can handle any kind of knowledge and **compute similarity between any kinds of objects**. The perspective mechanism allows us to have generic methods achieving relatively close results to those of “classic” similarity/distance measures in the literature [213].
- Finally, in [CJ11], we firstly illustrate the model within a mock-up simulation, where the hypothesis of knowledge emerging from preference dissemination is positively tested. Then we use a real-life web dataset (MovieLens) that mixes formal data about movies with user ratings. We have defined two distances based on the viewpoints’ evaluation and aggregation –one using the shortest path, one using all the paths– and proposed a topology-based **measure for assessing emergent knowledge**. With simulation, we have proved the potential of our measurements to capture both the explicitly reified knowledge and the implicit knowledge hidden behind subjective contributions; we have also proved informal learning by watching the evolutions of a knowledge map. Then by experimenting with the MovieLens dataset, we have shown the ability of our model to capture the semantics of the data, and compared the impact of the subjective contributions (the ratings) on the formal knowledge under different perspectives. Our results show that ViewpointS is a relevant, generic and powerful innovative approach to capture and reconcile formal and informal knowledge and enable collective intelligence.

Several prototypes have been developed during G. Surroca’s project. The latest is available at <http://viewpoints.cirad.fr>.

## IV.6 Challenge 6: Scalability and interoperability

The NCBO BioPortal, which is generally considered has the biggest ontology repository contains 770 ontologies as of end 2018. More and more vocabularies are being developed and hosted by the Linked Open Vocabularies platform. AgroPortal recently passed 100 semantic resources –with more than 2/3 of them not present in any similar ontology repository. Multiple domain specific ontology repository efforts have started often inspired by results in the biomedical domain and sometime by reusing NCBO technology (e.g., MMI OOR, AgroPortal, ESIPPortal, SIFR BioPortal). The semantic web and linked open data are being adopted widely, therefore:

**The more ontologies and ontology repositories are being developed, the more scalability and interoperability issues become important.**

By *scalability*, we mean the ability for ontology repositories to host more and more ontologies while:

- Keeping their central role of facilitating concept search, ontology identification and selection, mappings and annotation of data resource. Especially, with more ontologies the question of ontology alignment and reuse (by importing or reusing entities) is crucial as ontologies will necessarily overlap more and more.
- Continuing to offer robust, fast and reliable services. With more ontologies, infrastructure issues become important and repositories must ensure a certain level of technical quality of service.

By *interoperability*, we mean the ability for ontology repositories and libraries to interoperate one another. Some ontologies are necessarily useful to different communities and shall then be hosted in multiple repositories e.g., domain ontologies such as the Gene Ontology [214], or the Environment Ontology [50]. Because no repository will host them all, ontology repositories have to offer a certain level of interoperability to ensure their users that they will not have to work with multiple web applications and programming interfaces if their ontologies of interest are not all hosted by the same repositories.

Within the SIFR and AgroPortal projects (both presented Section III.3), we have been **particularly careful in not redeveloping features and functionalities that to our knowledge were already available**. We have designed and implemented two advanced prototype ontology repositories for the French speaking biomedical community and for the agronomy domain. Our choice to reuse the NCBO technology was of course justified by the large spectrum of features and services it offers, but in addition our motivation was: (i) to avoid re-developing tools that have already been designed and extensively used; (ii) to contribute on the long term to support a commonly used technology; and (iii) to offer the same tools, services and formats to different but still interconnected communities, to facilitate the interface and interaction between their domains (agro, bio, health). Now, relying on the same original technology enhance both technical reuse (for example, enabling queries to either systems with the same code), and semantic reuse (consuming resources from different repositories).

When we have developed new functionalities –as described along this manuscript– we have maintained our systems backward compatibles with the original NCBO technology to facilitate a convergence of the efforts. We strongly believe that sharing the technology is the best way to guaranty long term support and development by engaging different ontology practitioners and communities all around the world with their respective funding and supporting schemes.

Also, sharing the technology is the best way to make ontology repositories interoperable. In terms of interoperability, we have only contributed on three aspects:

- As presented Section IV.1.2, and explained in [CJ5], we have developed a new ontology metadata model for our repositories and now lead a standardization initiative on the MOD specification [CJ24]. Indeed, standard ontology metadata is a crucial aspect to achieve interoperability of ontology repositories.
- As presented in Section IV.4.4 and explained in [CJ8], we have adopted a proxy architecture to implement some of our new functionalities for the SIFR Annotator. This has enabled to quickly develop the NCBO Annotator+ and benefit of the new features also in AgroPortal. More recently, we have extended the proxy architecture to all services (not only the Annotator). Within the VisaTM project, this has enabled us to develop a **wrapper for any NCBO-like ontology repositories** to consume semantic resources within the OpenMinTed text and data mining platform [CJ62].
- As mentioned in Section IV.2.2 and explained in [CJ29], to store our multilingual mappings, we had to change their representation in BioPortal's architecture, especially allow a BioPortal virtual appliance to store mappings that target ontologies (i) in another instance of BioPortal (*inter-portal*), or (ii) not in any BioPortal instance (external). SIFR BioPortal now hosts mappings that interconnect its ontologies with the ones in the NCBO BioPortal.

# Chapter V.

## Conclusion and Perspectives



*Badlands National Park*

### V.1 Conclusion

In this manuscript, we have presented some scientific and technical challenges in building ontology repositories and ontology services. We have shown that **studying ontology repositories raises multiple informatics research questions** in varied areas such as knowledge representation, semantic web, data integration, natural language processing, and more. We have illustrated our thoughts with results obtained over the last 12 years within our projects in biomedicine and agronomy. We have not covered all related work on the cited challenges and we have certainly skipped other important challenges such as semantic consistency, ontology evaluation, visualization, community engagement. But we offered a short summary of multiple various contributions on ontology repository and ontology-based service research.

Building atop of our experience working on the NCBO BioPortal and by pursuing the collaboration on this topic with Stanford BMIR, we have started to develop an expertise on this area of research in Montpellier. We can now contribute to this field of research with concrete use cases, communities and outcomes.

Within the SIFR project: We have built an open and generic platform for hosting biomedical ontologies and terminologies in French language (or which contain French labels) and we offer a unique openly available resource for annotating French biomedical text data. These products are still very new and will need to be improved to fully find their place in the French speaking ecosystem. We are optimistic SIFR BioPortal will provide the French speaking biomedical community (e.g., clinicians, health professionals, researchers) with high quality ontology-based services, allowing them to improve their data production and consumption processes.

One of the main objectives of the SIFR project was to build the SIFR/French Annotator. We have shown in our publications the SIFR Annotator web service is comparable, in terms of quality and annotation performance to other knowledge-based annotation approaches, while being a generic easily accessible web service. We believe that **SIFR Annotator can help in a wide range of text mining or annotation problems**, but of course not universally. To drive these future evolutions, we are currently developing several partnerships in France to use SIFR Annotator within hospitals (CHRU Nancy, George Pompidou European Hospital in Paris) or for large-scale annotation efforts (e.g., to annotate the corpus of course of the French national medicine curriculum in the SIDES 3.0 project).

In addition to the SIFR BioPortal and Annotator, SIFR enabled us to investigate multiple related areas and obtain significant results in terminology extraction, ontology enrichment, ontology alignment, and more as presented in the manuscript.

Within the AgroPortal project: We have built an advanced prototype for ontologies and vocabularies in agronomy, food, plant sciences and biodiversity. We have reused the NCBO technology, customized it and completed it with new features to address the need of our community as described in [CJ10]. By specifically addressing the requirements of the agronomy community, **AgroPortal has kindled an important interest both at the national and international levels. It is now being adopted.**<sup>32</sup> The endorsement of associated partners

---

<sup>32</sup> The main journal paper [CJ10] and the four poster-demo articles related to AgroPortal cumulates today 38 citations.

(IRD, CIRAD, INRA, IRSTEA) illustrates the impact and interest not just in France, but also internationally (e.g., FAO, Bioversity International, IC-FOODS consortium, NCBO, AgBioData consortium, RDA working groups). Especially, the RDA Agrisemantics working groups and H2020 eROSA project consortium<sup>33</sup> have expressed interest in using AgroPortal as a key element of an open data infrastructure for agri-food. However, the current AgroPortal prototype only partially addresses the needs of the community: it is not multilingual, it is limited in terms of ontology alignment capabilities and does not provide semantic-search and retrieval of data. We have identified multiple perspectives for AgroPortal, some described in next subsections, that will be addressed during the new ANR project D2KAB starting in June 2019.

## V.2 Perspectives and research project

In the future, we will continue our efforts to address the identified challenges (and others), while continue to **offer to various scientific communities the means to share and leverage their ontologies or semantic resources and enable new science in their fields**. Each time possible, every theoretical or methodological result, we will do the effort of implementing a concrete service or tool in the context of ontology repositories to reproduce the results. There are three general objectives for continuation for our work:

### Objective 1: To encourage the adoption of semantics and develop state-of-the-art ontology repositories.

We would like to continue to implement a knowledge engineering vision in which ontologies capture domain knowledge and serve as a common denominator for data interoperability and integration. We will help different scientific communities develop new semantic resources and encourage them to embrace the semantic web standards when structuring their knowledge: SKOS to formalize interoperable vocabularies and thesauri and OWL to develop formal ontologies. We will design, develop and maintain tools to support ontology developers in producing, releasing, sharing, serving and interlinking their semantic resources. For this, we will transform the SIFR BioPortal and AgroPortal prototypes into widely adopted, long-term supported, robust and curated ontology repositories.<sup>34</sup> Always driven by our collaborators, we will develop new state-of-the-art methods and functionalities addressing the challenges presented in this manuscript.

### Objective 2: To contribute to producing FAIR data by building and exploiting Linked Open Data.

The ultimate reason for developing, sharing and aligning ontologies is to use them to semantically describe the data and make them FAIR. One perspective is to develop the methods and technologies to transform data into formalized and actionable knowledge that can be used by machines for search, reasoning and mining. For this, we will capitalize on experience acquired from previous projects in LIRMM's FADO team, on existing methods to lift legacy data into linked data (e.g., ANR Datalift – <https://datalift.org>) [215], ANR DOREMUS ([www.doremus.org](http://www.doremus.org)) [216] or to annotate/extract text data and map them with ontology concepts (e.g., ANR SIFR – <http://sifr.lirmm.fr/>). Additionally, we will develop new ontology-based methods tailored for the specificity and diversity of our domain of applications (currently health, biomedicine, agronomy, agriculture and related domains) e.g., measured by sensors, observed in the fields, extracted from literature, spatially described. A perspective is also to work on knowledge exploitation by tackling the challenges of visualizing and reasoning over linked data.

### Objective 3: To enable new semantically rich data driven science.

Our research project is not only to contribute by new methods and tools to the domain of knowledge engineering. In the context of future research projects, we will also demonstrate the validity of our vision by providing value-added knowledge-based services enabling new scientific discoveries in multiple application domains. For instance, in ANR D2KAB, each scenario will target its own scientific outcomes enabled by ontology-based exploitation of linked open data. We chose the scenarios for their methodological issues and potential impacts. In agronomy, they will enable the convergence of agronomy research (INRA) and agriculture (IRSTEA/ACTA) on two major issues: wheat breeding and food packaging. In biodiversity, they will enable a better assessment of impacts of a major global change factor on species and ecosystems. We hypothesize that semantically rich data will enable new scientific discoveries and will allow the translation of scientific research data (agronomy) to real world applications (agriculture) with potential economic impact and will lead to better agriculture, better food and more respectful ecological practices.

---

<sup>33</sup> The Agrisemantics initiative (<http://agrisemantics.org>) and H2020 eROSA ([www.erosa.aginfra.eu](http://www.erosa.aginfra.eu)) have for goal, among other, to set up a EU proposition for an e-infrastructure for open science in agri-food.

<sup>34</sup> The future of the SIFR BioPortal is naturally linked to the NCBO BioPortal as they are addressing the same scientific domains. We cannot say today if a language specific portal would still be required with a fully multilingual NCBO BioPortal.

These general objectives will be realized by following specific ideas in relation to the topics discussed within this manuscript.

#### About semantic resources interoperation:

Semantic resources may be used in different types of information systems. Vocabularies or thesaurus, like AGROVOC [217], are typically developed in SKOS and are used for document indexing and retrieval purpose. Formal ontologies, like the Plant Ontology [48] are developed in OWL, and used for data integration, knowledge modelling and reasoning. These two types of semantic resources –often simply called the same word– are different in term of content and usage. **One challenge is to facilitate the cohabitation, interoperation and appropriate use of each types of semantic resources** in a common shared environment. One technical challenge will be to make the NCBO technology fully SKOS compliant and enable the interoperation of very formal OWL ontologies with less formal vocabularies/thesauri. The original NCBO technology, which was mainly developed for ontologies, does not fully address less formal vocabularies.

#### About ontology metadata:

In the work presented Section IV.1.1, we did not pursue the goal of integrating all the reviewed vocabularies into a new “integrated vocabulary” that could become a standard for describing ontologies (e.g., a new OMV). However, the analysis of the existing metadata vocabularies and practices showed there is a clear need for better metadata authoring guidelines for the community of ontology developers and a need of harmonization of existing metadata vocabularies. With MOD1.2, we proposed the first elements of specification that would merge and harmonize existing metadata vocabularies, but it is still a temporary proposition. It is understandable that **to achieve community adoption, this work needs to engage more people, with the ultimate goal of producing a community standard endorsed by a standardization body** such as the World Wide Web Consortium (W3C). A similar work was done in the W3C HCLS working group to produce an application profile for datasets [218]. This is a perspective for the “ontology metadata” task group of the RDA VSSIG.

As concrete application of this work, we will continue to **implement new ontology metadata features** within AgroPortal and the SIFR BioPortal and continue the tedious effort of editing and curating the metadata with the ontology developers. In [CJ5], we have conducted a user survey which confirmed this was a relevant track for AgroPortal as it eases the processes of identification and selection of ontologies. We are also discussing with Stanford: (i) how to merge back our extended metadata model into the NCBO BioPortal; (ii) how to include in the Protégé ontology development tool (<http://protege.stanford.edu>) a mechanism to facilitate the creation of metadata from scratch, when the ontology is being developed; (iii) how to reuse the results of the CEDAR project (<http://metadatacenter.org>) in terms of metadata prediction and edition [219]; (iv) how to leverage a unified metadata model within the ontology Recommender service which currently relies mostly on the content of ontologies. As another possible perspective, we will propose a **set of indicators specific to semantic resources to assess their FAIRness level**. The indicators will be based on the recently proposed FAIR metrics [220] and standardized metadata. We could then develop a FAIRness scorer for automated computation of this score within our ontology repositories.

#### About multilingualism:

The roadmap described in [CJ51] was not realized completely. Ultimately, within the SIFR project, we changed our route and implemented a French version of BioPortal (and the Annotator) rather than a multilingual version of the whole NCBO BioPortal. This choice was made considering the enormous technical challenge of drastically changing the NCBO technology. In some sense, BioPortal is victim of its own success: the number of ontologies in the NCBO BioPortal grows faster than our abilities to develop new services, features and scalability mechanisms in the platform. By adopting an approach where we build alternative, sub domain ontology repositories, will facilitate dealing with such issues. Therefore, because it is now an important requirement for AgroPortal,<sup>35</sup> we expect to progress on the question of multilingualism within this project in the future. This is one of the tasks of the D2KAB project.

#### About ontology alignment:

Independently of the application domain of application, one challenge is the overlap between ontologies. As mentioned Section IV.3.3, one objective within ANR D2KAB is to set up the bricks of a *lingua franca* for agronomy, agriculture and biodiversity. We are currently harvesting (searching for already existing mappings) and extracting (from each source file uploaded in AgroPortal) the mappings available in agri-food and biodiversity semantic resources. Ultimately, we will make AgroPortal the reference platform for mapping representation, extraction, harvesting, generation, validation, merging, evaluation, visualization, storage and retrieval by adopting a

---

<sup>35</sup> Resources such as AGROVOC or AnaEE Thesaurus developed in European contexts are natively multilingual.

complete semantic web and linked open data approach and by engaging the community. We will investigate two complementary strategies: one-to-one ontology alignments, and alignments towards a common hub of concepts for agriculture and food with the GACS project:

(i) In terms of one-to-one mapping generation, we will investigate ontology alignment research using background knowledge (BK) approaches and experimenting in agronomy and biodiversity. We will use AgroPortal's mapping repository as a BK resource to improve state-of-the-art ontology alignment algorithms. In latest OAEI campaigns, machine-learning based BK-based approaches are the ones obtaining now the best results; but they are only applicable when relevant and clean knowledge sources are available. We know it will be a challenge to reproduce the results obtain in the biomedical domain (Section IV.3.2) to others by lack of training and reference alignments and BK resources. **One exploratory approach will be to adopt a graph-based mapping repository (using NoSQL property graphs) to facilitate the exploitation of concept-to-concept mapping paths to identify and select new ontology alignments.** We make the hypothesis that graph databases being particularly relevant for paths related queries, will help us to push state-of-the-art performance.

(ii) We are working with the RDA Agrisemantics working group in the development of Global Agricultural Concept Scheme (GACS) [221], which, with the support of major organizations of the domain among which the FAO, CABI, USDA- NAL and INRA, will become the **future pivot vocabulary for agriculture and food**. In the future, GACS will be extended to map vocabularies and ontologies in the agriculture and food domains and will provide stable URIs for common concepts and their types. GACS aims at reducing the proliferation of one-to-one ontology alignments by offering **a knowledge hub to which every semantic resource in the domain can be attached**. AgroPortal is expected to play a key role in achieving GACS project's goals and driving its evolution. We will then align all ontologies and vocabularies in AgroPortal to GACS.

#### About semantic annotation:

In [CJ2], we extensively discussed the limitations and perspectives for the SIFR Annotator and proposed some possible solutions for their mitigation in future technical evolutions of the service. In the future, we will continue to **introduce state-of-the-art natural language processing techniques in the semantic annotation workflow**. For instance, the SIFR Annotator obviously suffers from ambiguity between the general usage of a word and its medical usage (e.g., cold); a key element needed is a disambiguation module.

Processing clinical text is an enormous application for medical informatics; we believe a huge number of medical findings are hidden in the clinical data warehouse and electronic health records. Therefore, we will also have to **improve the clinical text features** presented in [CJ-UR2] and [CJ8]. Working with clinical data raises other challenges related to data privacy and ethics (anonymization, data access, etc.) that we will also have to address.

Our work on SIFR Annotator, is not limited to French, however, the technical efforts have mainly been focused on decoupling the architecture from English and for allowing an easy adaptation to other languages. Although our target language is French, we have made some of our new features also available for English and we believe our efforts and experience would **facilitate deployment of new instance of BioPortal and its Annotator in other language** (especially roman language or linguistically close to French/English) after minor configuration and adjustments. In the future, we will investigate the application of the technology to multiple language also.

Finally, it is inevitable to reconsider today our methods at the light of recent advances in machine learning approaches. In [CJ2], we compared our approach to machine learning based approaches which indeed in some cases obtain better results (especially on benchmark tasks for which training data are available). Even if machine learning is not always applicable –especially in the health domain where obtaining annotated training data is very difficult or not always applicable– we believe, in the future, we will have to **investigate more machine learning techniques for semantic annotation**.

#### About semantic search and linked open data:

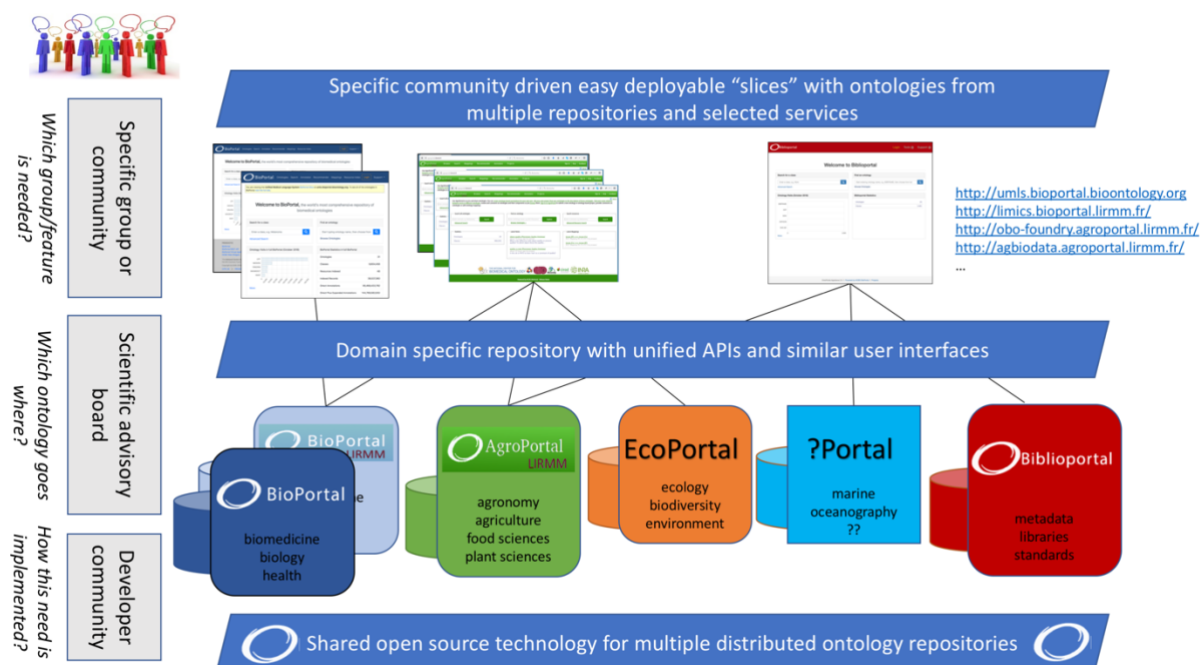
As explained in Section IV.5, data indexing and semantic search is one of the major use cases for ontologies. In the future we will therefore continue to investigate these issues using ontologies to annotate, index and represent miscellaneous data. While the original BioPortal had the NCBO Resource Index, presented in Section IV.5.1, we decided to take another approach in our ontology repository projects to provide data access through ontologies. The challenge is therefore to **enable semantic search, by implementing state-of-the-art algorithms leveraging the semantics of the ontologies in the repositories** (i.e., is-a relation, synonyms, mappings) to retrieve data. Among the technical challenge is also to address **multilingualism to enable multilingual search of data and browsing of ontologies**. In AgroPortal, we plan to rely on external annotated linked data resources to query directly with ontology terms to enabled ontology-based search (aka. semantic

search [222]), so that a user browsing ontologies can get direct access to data. We plan to rely on external resources such as AgroLD, presented Section IV.5.2 or other annotated datasets such as the CIARD RING directory (<http://ring.ciard.net>) [223], or Planteome [224]. However, in agronomy, we have not seen integrated semantic resources that have had a major impact such as the ones that have been developed in biomedical and health sciences (e.g., [Bio2RDF.org](http://Bio2RDF.org) [200], EBI RDF [197]). Indeed, we cannot yet measure the impact of the previously mentioned resources in terms of linked open data produced and made available to the rest of the world. Someone may ask: **where are agronomy and biodiversity in the famous LOD cloud diagram?** (<http://lod-cloud.net> – Figure 3, page 22). One perspective is then, with our partners, to build the agronomy, agriculture and biodiversity Linked Open Data cloud.

#### About ontology repositories interoperability:

The **role of ontologies and vocabularies for producing FAIR data has been clearly established** [15]. Besides biomedicine and agronomy, we have identified other application domains that are interested by building ontology repositories (geographic and environment systems, ecology & ecosystems, humanities) and new projects shall emerge with new partnerships in the future. One long term perspective would be to build a generic, still easily customizable solution, for making vocabulary and ontology repositories more interoperable, and possibly, sharing more technological components.

In the future, we will encourage the ontology repository providers to implement a common architecture as illustrated Figure 35. The bottom part illustrates how several ontology repositories, mostly based on the same technology (rounded rectangle), would exist side by side and serve their content via a common API and similar user interfaces. At a second layer, specific community-driven slices would consume the content from the source ontology repositories and offer their community a customized and simplified end-point. The NCBO technology already supports the deployment of specific “slices” i.e., a mechanism to allow users to interact (both via API or UI) only with a subset of ontologies in the repository. If browsing a slice, all the repository features will be restricted to the chosen subset, enabling users to focus on their specific set of interest. In the future, one challenge would be to enable the slice mechanism to consume ontologies from different source repositories so that a community (a specific project, data center, group of users, organization, etc.) could access and use ontologies while not being aware of the technical details of the platforms serving these ontologies. Every ontology repository involved would have to implement a common API enabling transparent consumption of the repository’s content; this could be done with the recently proposed SmartAPI approach [225].



**Figure 35. Ontology repositories working together.** Shared technology in the bottom part and community specific ‘slices’ in the top part.

### Project D2KAB, a roadmap for 2019-2023:

Multiple of these perspectives will be investigated in the context of D2KAB (Data to Knowledge in Agronomy and Biodiversity), an ANR project led by LIRMM, starting in June 2019 and briefly described hereafter.

D2KAB's objective is to **create a framework to turn agronomy and biodiversity data into knowledge – semantically described, interoperable, actionable, open–** and investigate scientific methods and tools to exploit this knowledge for applications in science & agriculture. Agronomy/agriculture and biodiversity (ag & biodiv) face several major societal, economical, and environmental challenges, a semantic data science approach will help to address. We shall provide the means **–ontologies and linked open data– for agronomy & biodiversity** to embrace the semantic web to produce and exploit FAIR data. To do so, we will develop new original methods and algorithms in the following areas: data integration, text mining, semantic annotation, ontology alignment and linked data exploitation. D2KAB project brings together a unique multidisciplinary consortium of 11 partners to achieve this objective: 2 informatics research units (LIRMM, I3S); 5 INRA/IRSTEA applied informatics research units (URGI, MaIAGE, IATE, DIST, TSCF) specialized in agronomy or agriculture; 2 labs in biodiversity and ecosystem research (CEFE, URFM); 1 association of agriculture stakeholders (ACTA); and 1 partnership with Stanford BMIR department. Each of the project **driving scenarios (food packaging, agro-agri linked data, wheat phenotype, ecosystems & plant biogeography)** will have a significant impact and produce concrete outcomes for ag & biodiv scientific communities and socio-economic actors in agriculture.

D2KAB's detailed objectives are an instantiated version, of the three general objectives described earlier in this Section. Moreover, the project gathers a consortium and a rich ecosystem (illustrated in Figure 36) to be productive beyond the span of the ANR support period.



**Figure 36. D2KAB project's ecosystem.**

D2KAB will rely on semantic web technologies to build a unified knowledge graph (illustrated in Figure 37) based on several data sources (already in RDF or not) and support new visualization and exploitation of this graph in each of the five scenarios. Ultimately, it will automatize the procedure to produce Linked Open Data for agronomy and biodiversity datasets used in the projects.

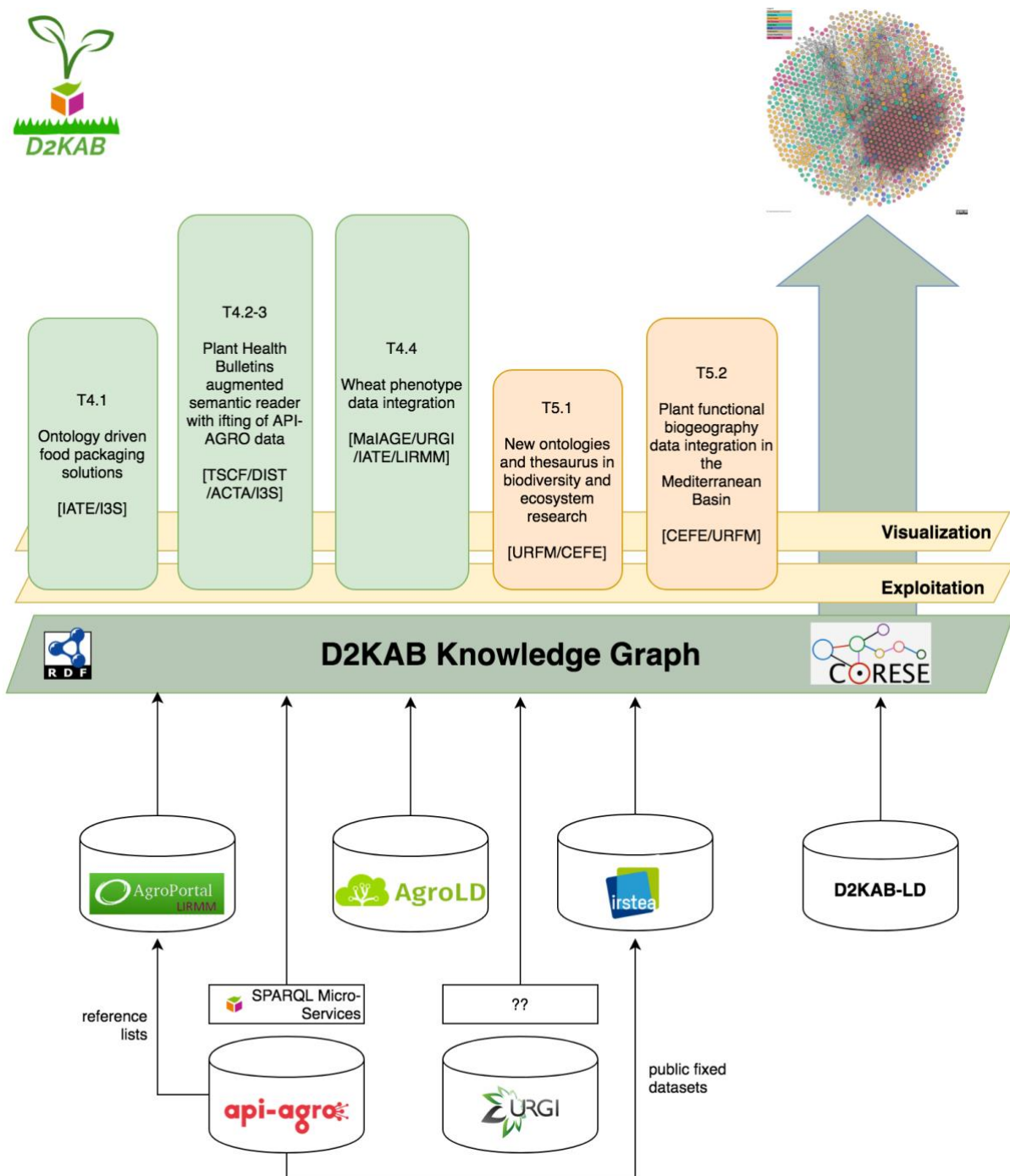


Figure 37. D2KAB's knowledge graph.



# Chapter VI.

## Curriculum Vitae



Great Sand Dunes National Park

PHD IN INFORMATICS  
ASSISTANT PROFESSOR, UNIVERSITY OF MONTPELLIER

### CONTACT & PROFESSIONAL SITUATION

---

Born May 26, 1980, Nîmes (Gard), France, married, 2 children, French nationality.

Skype: clementpro

Email: [jonquet@lirmm.fr](mailto:jonquet@lirmm.fr)

Twitter: [@jonquet\\_lirmm](https://twitter.com/jonquet_lirmm)

Web : [www.lirmm.fr/~jonquet](http://www.lirmm.fr/~jonquet)

ORCID: <http://orcid.org/0000-0002-2404-1582>

Public profiles: [Google Scholar](#), [Microsoft Academic](#), [ResearchGate](#), [DBLP](#), [HAL](#), [PubMed](#), [CiteSeer](#).



### CV STRENGTHS

---

- **Multidisciplinary research activities** (Ontologies, Semantic Web, Biomedical Informatics, Semantic annotation, Text mining, Service-oriented computing, Web Science, Agents).
- Experience in applied research (biomedicine, agronomy), software engineering & transfer skills.
- Collaborative work experience, **project funded research** (EU, ANR, NIH), management skills (project leading, outsourcing, supervision).
- Principal investigator of ANR JCJC (Young researcher program) SIFR project (2013-2017), co-PI of ANR PractiKPharma (2016-2019), recipient of **H2020 Marie Curie grant** (2016-2019), PI of ANR D2KAB (2019-2023).
- **9 years of lecturing** Informatics/Computer Science to different student grades. (Co)supervision of 12+ MSc. Students & 3 PhD candidates.
- Mobility: **3-year postdoc and later 3-year visiting scholar at Stanford University.**

### WORK EXPERIENCE

---



- **Since Sept. 2010: Assistant Professor, University of Montpellier, France.**  
Researcher in the Laboratory of Informatics, Robotics, and Microelectronics of Montpellier ([LIRMM](#)) and teacher at [Polytech Montpellier](#) Engineering School.
- **2015-2018: Visiting scholar, Stanford University, USA**  
Center for Biomedical Informatics Research ([BMIR](#)). Working with Pr. M. A. Musen & the [NCBO](#).
- **2007-2010: Postdoctoral scholar, Stanford University, USA**  
Center for Biomedical Informatics Research ([BMIR](#)). Working within Pr. M. A. Musen's group.
- **2006-2007: Lecturer, University Montpellier 3, France (humanities and social sciences) (~ French ATER).**  
Researcher at LIRMM.
- **2003-2006: French government PhD grant & young lecturer at University Montpellier 2 (sciences and techniques) (~ French 'allocataire MENRT' & 'moniteur CIES').**  
PhD achieved at LIRMM and supervised by Pr. Stefano A. Cerri.



## EDUCATION



- 2006: **PhD in Informatics/Computer Science** (First class with distinction) – UM2
- 2003: MSc in Computer Science (2.1 honours) – UM2 (~ French DEA & Maîtrise)
- 2001: BSc in Computer Science (2.2 honours) – UM2 (~ French Licence & DEUG)
- 1998: High School Diploma specialized in Maths (2.2 honours) – Uzès (Gard) (~ French Bac. S)

## RESEARCH ACTIVITY

### RESEARCH PROJECTS

- 2019-2023: **Principal investigator** of the Data to Knowledge in Agronomy and Biodiversity ([D2KAB project](#)) – ANR, 12-partner, 30-person project on semantics and linked data in agronomy & biodiversity.
- 2017-2019: [Visa<sup>TM</sup>](#) project. Text & Data mining infrastructure for French scientists. BSN-10 head by C. Nédellec, INRA, Jouy-en-Josas.
- 2015-2023: **Coordinator** of the [AgroPortal project](#), a vocabulary and ontology repository for agronomy, food, plant sciences and biodiversity, partially supported by ANR (SIFR, IBC, Labex NUMEV, Labex AGRO, EU-MSCA SIFRm, D2KAB).
- 2016-2019: **Co-principal investigator** of the [PractikPharma project](#) (Practice-based evidences for actioning Knowledge in Pharmacogenomics) – ANR headed by A. Coulet, INRIA, Nancy.
- 2013-2019: **Principal investigator** of the Semantic Indexing of French Biomedical Data Resources ([SIFR project](#)) – ANR Young Researcher & EU-MSCA. Building ontology-based services to leverage biomedical ontologies and terminologies in indexing, mining and retrieval of French biomedical data. Also supported by Univ. of Montpellier & CNRS, France-Stanford and Eiffel programs.
- 2012-2018: Institut de Biologie Computationnelle ([IBC](#)), axe 5 (workflow & data integration) – ANR Inv. d'Avenir BioInfo. Development and use cases for AgroPortal & AgroLD.
- 2011-2013: CR2i DiagnosTIC-Santé project (Centre de Recherche et d'Innovation Industrielle) – Inv. d'Avenir PFMI. Member of the metadata repository group (multi-omics platform development).
- 2007-2010: **National Center for Biomedical Ontology** ([NCBO](#)) – Part of the National Centers for Biomedical Computing supported by the NIH Roadmap; provider of the NCBO BioPortal.
- 2003-2007: European Learning Grid Infrastructure (ELeGI) – IST IP EU (FP6).
- 2003-2004: Learning Grid of Excellence Working Group (LeGE-WG) – IST STREP EU (FP6).

### TEAMS & RESEARCH GROUPS

- Since 2018: Member of the [LIRMM's Fado team](#) (Fuzziness, Alignments, Data & Ontologies).
- 2018-2019: Associated member (INRIA delegation) of [INRIA Sophia-Antipolis's Wimmics team](#) headed by F. Gandon (social & formal semantics on the Web, linked data).
- 2015-2018: Member of Musen's lab and until 2017 of Dumontier's lab at Stanford [BMIR](#) (medical informatics, knowledge representation and semantic Web (Protégé & BioPortal)).
- 2010-2018: Member of the [LIRMM's Smile team](#) (multi-agent systems, Web science, service-oriented computing, ontologies, serious games, simulation).
- 2007-2010: Member of Musen's lab and the NCBO team at Stanford [BMIR](#) (medical informatics, knowledge representation and semantic Web (Protégé & BioPortal)).
- March 2006: Associated member (internship) of the Open University's [KMI](#) group.
- 2003-2007: Member of the LIRMM's Kayou team (multi-agent systems, constraints, Web, Grid, service-oriented computing, ontologies, collaborative learning).

### RESEARCH TOPICS

Ontologies & vocabularies, Ontology repositories, Ontology-based services, Semantic Web, Semantic annotation, Biomedical Informatics, Ontology alignment, Metadata, Linked Open Data, Knowledge representation, Data integration, Information Retrieval, Text mining, Service-oriented computing, Web Science, Distributed systems, Multi-Agents Systems, Web 2.0. Applications to biomedicine/health and agronomy/food/plant/biodiversity.

## FUNDED GRANTS (AS LEADER)

Reported in Table 1, Section I.3, page 11.

## CURRENT RESEARCH ACTIVITY

Described within the manuscript.

## PAST RESEARCH ACTIVITY

- *NCBO project & postdoctoral research:* Within NCBO I worked on semantic annotation of biomedical data with biomedical ontologies. I actively contributed to the [NCBO BioPortal](#) web application well used in the biomedical community. I designed an **ontology-based annotation workflow**. This workflow embeds different components (e.g., concept recognition tool, semantic expansion algorithms) in order to leverage the knowledge represented in ontologies and **facilitate biomedical data integration**. Based on this workflow, I conceptualized, designed, developed and experiment three research applications: (i) the [NCBO Annotator](#), an ontology-based web service that can be used by the life sciences community to tag their data automatically with ontology concepts; (ii) the [NCBO Resource Index](#), a database of open biomedical resources annotated and indexed with ontology concepts (20+ resources and 200+ ontologies at that time) which can be used to search and integrate data; (iii) the [NCBO Recommender](#), a service which informs the user of the most appropriate ontologies relevant for their given dataset.
- *Doctoral research:* Situated at the **crossing of three important domains**: service-oriented computing (web service, components, business process, etc.), multi-agent systems (modeling, interaction, architecture) and Grid (resources sharing, Grid service, Grid computing). I proposed in my thesis a new vision for the concept of “service”, called **dynamic service generation**. This vision, based on interactions between agents (human or artificial) and relying of a Grid infrastructure, enabled dynamic construction of services based on the conversation between user & provider. Two important contributions were: (i) STROBE: an agent communication and representation model based on conversation contexts to enable interactive specification of agent capabilities; (ii) Agent-Grid Integration Language (AGIL): a grid-multiagent integrated model formalized with a description language which leverages the stateful and dynamic aspect of Grid services.
- *ELeGI project research:* I worked on a collaborative environment constructed over a Grid infrastructure based shared desktops. We experiment the environment with a community of chemists tackling the problem of **collaborative construction of an ontology**.

## COLLABORATIONS

Reported in, Section I.4, page 13. Collaborations before 2011 include:

- *2007-2010:* [NCBO collaborators](#) and community: Univ. of Colorado School of Medicine (L. Hunter), Univ. of California San Francisco (I. Sim), Medical College of Wisconsin (S. Twigger), Wright State Univ. (A. Sheth), Goal: leverage NCBO solutions within biomedical sciences scenarios.
- *2004-2006:* A. Krief's lab, Notre Dame de la Paix Univ., Namur, Belgium. collaborative construction of ontology for chemistry.
- *2003-2006:* Knowledge Media Institute (KMI), Open Univ., Milton Keynes, UK (E. Motta, J. Domingue, M. Eisenstadt). Goal: using agent approach for Grid services and collaboration.

## AWARDS & DISTINCTIONS

- Keynote speaker at 4<sup>th</sup> *Symposium on Information Management and Big Data* ([SIMBig 2017](#)).
- Shared best paper award at 6<sup>th</sup> *French Ontology Conference* ([JFO 2016](#)).
- Recipient of the EU Marie Curie-Sklodowska program (2016-2019).
- 1<sup>st</sup> Prize at the 2<sup>nd</sup> BD2K & 4<sup>th</sup> [Network of BioThings Hackathon](#) (Stanford, 2015)
- Holder of French ministry distinction, *Prime d'Excellence Scientifique* ([PES](#)) since 2013.
- Recipient of the French National Research Agency (ANR) Young Researcher program, 2012.
- Honorable mention award at 3<sup>rd</sup> *ACM International Conference on Web Science* ([WebSci 2011](#)).
- Selected in [Pr. Russ Altman's 2011 Year in Review](#) for journal article about biomedical ontology recommendation. *AMIA Translational Bioinformatics Summit* (AMIA-TBI 2011).





- [Semantic Web Challenge](#) 2010 winner (with the NCBO team) at 9<sup>th</sup> *International Semantic Web Conference (ISWC 2010)* with the NCBO Resource Index.

#### SOFTWARE DEVELOPMENT & TECHNOLOGY

- Since 2013, all development projects are maintained on GitHub:
  - <https://github.com/d2kab>
  - <https://github.com/sifrproject>
  - <https://github.com/agroportal>
  - <https://github.com/practikpharma>
- 2013-2018: Within the *SIFR* & *AgroPortal* projects:
  - Design of YAM-BIO a tool for ontology alignment with background knowledge resources (A. Annane's PhD project).
  - Design, development & maintenance of the SIFR BioPortal (<http://bioportal.lirmm.fr>) (French medical terminologies) & AgroPortal (<http://agroportal.lirmm.fr>) projects.
  - Design & development of the French Annotator and the NCBO Annotator+ both included within the SIFR BioPortal.
  - Design of ViewpointS, a graph-based system for collaborative knowledge representation and learning (G. Surroca's PhD project).
  - Design of BioTex for automatic extraction of biomedical terms from text (J. Lossio's PhD project).
  - Design & development (in collaboration with LGI2P) of semantic distances Web services.
- *BioPortal* (<http://bioportal.bioontology.org/>), a web repository of biomedical ontologies developed by NCBO. I actively contributed to the evolution and design to the core NCBO BioPortal services and participate in the support to the community.
- *NCBO Annotator, Resource Index & Recommender* (BioPortal URL + [/annotator](#), [/resources](#), [/recommender](#)). I was the main researcher (along with N. Shah, PhD, MD) and architect of these 3 services (prototyping, testing, evaluation, QA and deployment). I supervised 3 part time software developers working on these projects during 2 years.
- *STROBE model*, prototype implementation of the multi-agent model designed during my PhD project.
- Experimentation with the [Grid Shared Desktop](#) developed within the EleGI project.

#### SUPERVISION OF RESEARCH ACTIVITIES

- 2018-2020: Supervision of E. Abrahao (postdoc AgroPortal/Lingua) with K. Todorov & P. Neveu.
- 2017: Co-supervision of S. Zevio (MSc student, U. Montpellier) with S. Bringay & A. Tchechmedjiev.
- 2017: Supervision of C. Goehrs (MD & MSc. Student, U. of Bordeaux).
- 2016-2017: Supervision of A. Abdaoui (postdoc PractiKPharma).
- 2016-2018: Co-supervision of A. Tchechmedjiev (postdoc PractiKPharma) with S. Bringay.
- 2016: Co-supervision of S. Eholié (MSc student, U. of Nantes) with S. Bringay & M-D. Tapi-Nzali.
- 2015-2018: **Co-supervision of A. Annane, PhD candidate**, (cotutelle, Eiffel fellow) with Z. Bellashene & F. Azouaou (ESI Algeria)) on ontology alignment (SIFR & PractiKPharma).
- 2015: Co-supervision of C. El Ghandour & M. Serhani (MSc students, U. Montpellier) with J-A. Lossio on prototyping BioTex in SIFR BioPortal.
- 2015-2018: Supervision & management of A. Toulet (research engineer, AgroPortal project).
- 2015-2017: Supervision & management of V. Emonet (research engineer, SIFR project).
- 2015: Supervision of J. Diener (research engineer, IBC project) with P. Larmande.
- 2014: Co-supervision of P. Burc and O. Duploux (MSc students, U. Montpellier) with S. Harrispe (LGI2P, Nîmes) on semantic distances.
- 2014: Co-supervision of L-H. Méric (eng. student, IMT St Etienne) with P. Lemoisson and G. Surroca.
- 2014: Supervision of S. Melzi (MSc student, U. Montpellier).
- 2014: Co-supervision of A. Dia (MSc student, U. G. Berger, Senegal) with P. Lemoisson and G. Surroca.



- 2013: Co-supervision of K. Cauchois (MSc. student, U. Rouen) with S. Darmoni (CHU Rouen) on exporting HeTOP's content to OWL.
- 2013-2017: **Co-supervision of G. Surroca, PhD candidate**, with P. Lemoisson and S.A. Cerri, on graph-based social/semantic data knowledge representation with ViewpointS.
- 2013: Supervision of K. Bouarech, (MSc student, U. Montpellier).
- 2012-2015: **Co-supervision J-A. Lossio-Ventura, PhD candidate**, with M. Roche and M. Teisseire on biomedical terminology extraction (SIFR project).
- 2010: Co-supervision of R. Castro & B. Paiva (MSc students, U. Montpellier), with S.A. Cerri (collaboration Stanford-LIRMM) on semantic distances and web service composition.
- 2010: Supervision of T. Tenneneti (MSc student, Stanford) on concept recognition.
- 2009: Co-supervision of A. Ghazvinian (MSc student, Stanford) with N. Noy on ontology alignment.
- 2009: Co-supervision of G. Parai (MSc student, Stanford) with N. Shah on lexicon building.
- 2009: Co-supervision of N. Bhatia (MSc student, Stanford) with N. Shah on concept recognition.
- 2006: Co-supervision of F. Duvert (MSc students, U. Montpellier), with S.A. Cerri agent-grid ontology.
- 2005: Co-supervision of a 3-student-group (BSc. students, U. Montpellier) with R. Colleta on web service and constraint programming.
- 2005: Co-supervision of a 3-student-group (BSc. students, U. Montpellier) with S.A. Cerri on STROBE and MadKit.

## PROFESSIONAL RESPONSABILITIES

### MISSIONS & EXPERTISE

- **Project proposal reviewing** for French ANR (\*3) and US NIH (\*1).
- 2013: **French-US bioinformatics collaboration committee** member, supervised by A. Viari (INRIA) for the Ministries of Higher Education & Research and Foreign Affairs.
- 2011-2015: Member of the expert pool of the French Ministry of Higher Education & Research for evaluating research & development tax credit (French CIR and JEI).
- Article reviewing activity for 12 international journals, 19 international workshops or conferences & 8 national workshops or conferences. Detailed hereafter.

### OTHER RESPONSIBILITIES

- 2012-2015: Member of UM2 council for Information and Communication Technologies in Education (TICE). Representative for Polytech Montpellier.
- 2012-2015: Head of [Polytech Montpellier iPad for students project](#). I 'lead' a group of 70 teachers interested in pedagogical innovations using ICT and iPad, in and out of the classroom.
- 2012-2015: Responsible of the last year of the "Informatics & Gestion" curriculum at Polytech Montpellier Engineering School (eq. Master degree).
- 2004-2005: Elected representative of computer science PhD students at LIRMM. Interesting activity to understand the organization and operation of a research lab.

### PROGRAM CHAIRING AND ORGANIZATION

- Co-chair of *Semantics for Food, Agriculture, Environment and Nutrition* workshop ([SemFAEN 2018](#)) at Semantics 2018, Sept. 2018, Vienna, Austria.
- Co-session chair *Semantics for biodiversity and ecosystem research* at [ICEI 2018](#).
- Co-program chair and organization committee of 2<sup>nd</sup> *International Workshop on Semantics for Biodiversity* ([S4BIODIV 2017](#)) at ISWC 2017, Nov. 2017, Vienna, Austria. ~30 participants.
- Organization of the [AgroHackathon](#) series (in June 2016 and July 2017). ~15-30 participants.
- Participation to the organization (with S. Bringay) of 27<sup>èmes</sup> *Journées francophones d'Ingénierie des Connaissances* ([IC 2016](#)), June 2016, Montpellier, France, ~100 participants.
- Co-program chair (with D. Cassagne) of the "return of experience" track of the *French ICT in Education Conference* ([TICE 2014](#)), Nov. 2014, Beziers, France. ~100 participants.





- Local chair (with F. Scharffe) of 10<sup>th</sup> *Extended Semantic Web Conference* ([ESWC 2013](#)), May 26-30 2013, Montpellier, France. ~350 participants.
- Co-program chair and organization committee of 1<sup>st</sup> *International Workshop on Semantics for Biodiversity* ([S4BIODIV 2013](#)) at ESWC 2013, May 2013, Montpellier, France. ~40 participants
- 2010-2013: Organizer of the group of interest [Web Science Montpellier](#) Meetup and organization of 1<sup>st</sup> [Web Science Montpellier](#) Meetup workshop, in Montpellier, France, May 13<sup>th</sup> 2011. 25 participants.
- Participation to the organization of local workshops (OTM 2006 & ALCAA 2004).

#### ARTICLE REVIEWING

[Data Intelligence](#) (open access), [Semantic Web Journal](#) (IOS Press), [Applied Ontology](#) (IOS Press), [Bioinformatics](#) (Oxford Journals), [BMC Bioinformatics](#) (BioMed Central), Journal of [Web Semantics](#) (Elsevier), [Knowledge-Based Systems](#) (Elsevier), Journal of [Biomedical Informatics](#) (Elsevier), Journal of [Biomedical Semantics](#) (BioMed Central), [IMIA Year Book](#) (Schattauer), French [Technique et Science Informatique](#) (Hermès), French [Revue d'Epidémiologie et de Santé Publique](#) (Elsevier), [Service Oriented Computing and Applications](#) journal (Springer), Grid Computing and Multi-Agent Systems journal (Serials Publications)

#### INTERNATIONAL PROGRAM COMMITTEES

- 21<sup>st</sup>, 27<sup>th</sup>-28<sup>th</sup> Int. World Wide Web Conference ([WWW 2018-2019](#), [WWW 2012](#) (Demo track))
- 16<sup>th</sup>-17<sup>th</sup> Int. Semantic Web Conference ([ISWC 2017-2018](#)).
- 14<sup>th</sup>, 16<sup>th</sup> European Semantic Web Conference ([ESWC 2017, 2019](#)).
- European Federation for Information Technology in Agriculture, Food & Environment ([EFITA 2017](#)).
- 1<sup>st</sup> Language, Data and Knowledge conference ([LDK 2017](#))
- 1<sup>st</sup>-5<sup>th</sup> Int. Symposium on Information Management & Big Data ([SIMBig 2014-2018](#)).
- 11<sup>th</sup>-18<sup>th</sup> BioOntologies SIG ([BioOntologies 2009-2017](#)).
- 8<sup>th</sup>-11<sup>th</sup> Semantic Web Applications and Tools for Life Sciences ([SWAT4LS 2015-2016-2018](#)).
- 11<sup>th</sup> & 12<sup>th</sup> African Research in Computer Science and Applied Mathematics ([CARI 2014-2016](#))
- 1<sup>st</sup>-2<sup>nd</sup> Int. Workshop on Semantics for Biodiversity ([S4BIODIV 2013, 2017](#)).
- 1<sup>st</sup> Computational Semantics in Clinical Text ([CSCT 2013](#)) workshop
- 4<sup>th</sup> Int. Conference on Web Science ([WebSci 2012](#)).
- 1<sup>st</sup> & 2<sup>nd</sup> Int. Workshop on Web Science & Information Exchange in Medical Web ([MedEx 2010-2011](#)).
- 9<sup>th</sup>, 11<sup>th</sup> & 13<sup>th</sup> Int. Conference on Intelligent Tutoring ([ITS 2008, ITS 2014, ITS 2012](#)).
- 4<sup>th</sup> & 5<sup>th</sup> Int. KES Symposium on Agents & MAS Technologies & Applications ([AMSTA 2010-2011](#)).
- Workshop on Ontology Repositories for the Web ([SERES 2010](#)).
- 1<sup>st</sup> Int. Workshop on User-generated Services ([UGS 2009](#)).
- Workshop Extending Database Technology for Life Sciences workshop ([EDTLS 2009](#)).
- Int. Workshop on Service-Oriented Computing: Agents, Semantics, and Engineering ([SOCASE 2009](#)).

#### NATIONAL PROGRAM COMMITTEES

- Atelier *Web des données* ([AWD 2019](#)).
- Workshop Knowledge Engineering & Health ([IA & Santé 2018](#), [SIIM 2015 & 2017](#), [ICSanté 2012-2016](#)).
- Workshop sources & data integration in agriculture, food, environment ontologies ([IN-OVIVE 2017](#))
- 6<sup>èmes</sup> *Journées francophones sur les Ontologies* ([JFO 2016](#)).
- 24<sup>èmes</sup>-29<sup>èmes</sup> *Journées francophones d'Ingénierie des Connaissances* ([IC 2013-2018](#)).
- 1<sup>er</sup> Atelier *Ontologies et Jeux de Données pour évaluer le web sémantique* ([OJD 2012](#)).
- 1<sup>er</sup>-3<sup>èmes</sup> Atelier *Quantité et Robustesse pour le Web de données* ([QetR 2011-2013](#)).
- 1<sup>er</sup> Atelier *Extraction des Connaissances et Contextualisation* ([ExCoco 2011](#)).
- 7<sup>ème</sup> *Colloque Agents Logiciels, Coopération, Apprentissage, Activité* (ALCAA 2004).

#### DETAILED SEMINARS & INVITED PRESENTATIONS



17 presentations on [SlideShare](#), (cumulating ~9400 views).

- LIRMM Scientific Day, December 2018 (invited by P. Poignet).
- RDA France 1<sup>st</sup> National Day, JNSO 2018, December 2019 (invited by F. Genova).
- INRIA's Wimmics Seminar, November 2018 (invited by F. Gandon).
- PhenoHarmonIS workshop, May 2018 (invited by E. Arnaud).
- RDA 11<sup>th</sup> Plenary, IGAD pre-meeting, March 2018 (invited by I. Subirats).

- EUDAT Conference Semantic Working Group, January 2018 (invited by Y. Le Franc).
- Keynote INIST 'Ingénierie des Connaissances' Series, December 2017 (invited by C. Francois).
- IC-Foods Conference, November 2018 and 2017 (invited by M. Lange).
- Keynote at SIMBig 2017, September 2017 (invited by J-A. Lossio).
- GDR SemanDiv, July 2017 (invited by E. Garnier).
- French Minister – DSSIS (réunion serveurs multi-terminologiques), June 2017 (invited by B. Séroussi).
- BMIR Research in progress colloquium, Mai 2016 (invited by M. Musen).
- Protégé group meeting, Stanford Univ., April 2016 (invited by T. Tudorache).
- Dumontier's lab group meeting, Stanford Univ., January & November. 2016 (invited by M. Dumontier).
- Keynote at the French RISE 2015 workshop, Rennes, France, June 2015 (invited by C. Roussey).
- Protégé group meeting, Stanford Univ., April 2015 (invited by T. Tudorache).
- Forum TIC's, Mons, Belgium. April 2015 (invited by B. Champagne).
- LIGI2P Science & Society seminar, Nimes, France. March 2015 (invited by S. Harispe).
- CENTAL team at UC Louvain, Belgium, Dec. 2014 (invited by C. Fairon).
- Réseau IN-OVIVE, INRA, Montpellier, Oct. 2014 (invited by P. Neveu).
- IBC Scientific day, Montpellier, May 2014 (invited by O. Gascuel).
- SPIM team at INSERM Paris, June 2011 (invited by M-C. Jaulent).
- LIM team at Rennes Univ., April 2011 (invited by O. Dameron).
- CISMef team at Rouen School of Medicine, March 2011 (invited by S. J. Darmoni).
- Research seminar on ICT & Health, LIRMM, Montpellier, February 2011.
- EXMO team at INRIA Grenoble, France, March 2010 (invited by J. Euzenat).
- Smile team at LIRMM, UM2, France, February 2009 (invited by S. A. Cerri).
- EDELWEISS team at INRIA Sophia, France, January 2009 (invited by F. Gandon).
- Talk at the NCBO Developer Conference, Stanford Univ., USA, Dec. 2007.
- Intelligent Interactive Distributed Syst. group, Vrije Univ., Amsterdam (invited by F. Brazier). May 2007.
- LIRMM's Informatics department day, UM2, France. July 2005.
- Protégé group meeting, Stanford Univ., CA, USA (invited by M. Crubezy). June 2005.
- E-LeGI WP6 (Work Package 6) seminar, LIRMM, UM2, France. June 2004.
- Computer Science PhD students seminar, LIRMM, UM2, France. January 2004.
- Talk within the GT MFI (Groupe de Travail Modèles Formels de l'Interaction) working group, LIP6, Université Paris 6, France. December 2003.
- Social Informatics seminar, LIRMM, UM2, France. June 2003.

## PUBLICATIONS SUMMARY

---

My complete list of publications is described below or [online](#) or in the [HAL database](#). I try to publish in open access journals (gold open access), but when not the case, a **PDF is always available** for every document (green open access) on HAL. Other incomplete listings include: [Google Scholar](#), [Microsoft Academic](#), [ResearchGate](#), [DBLP](#), [PubMed](#), [CiteSeer](#).

The first author is the “main” author. The last author is generally the supervisor. All publications (84) or communications (26) have been peer-reviewed (if not explicitly mentioned), including:

- **23 journal** (6 as first author, 5 as second author, 3 as last author), 21 international conference, 15 workshop, 20 national (French) and 2 dissertations.
- 60 are international publications; most have been written in a collaborative context; more than 2/3 have been written by person(s) under my (co)supervision.

Overall my publications cumulate **~2290 citations** as of Google Scholar (January 2019); including 772 for publications as first author. I have published in multiple domains & contexts:

- **Biomedical Informatics**: 1 recent article in *Journal of Biomedical Informatics* (Elsevier, IF 3.23), 4 articles in *BMC Bioinformatics* (IF 3.45, CORE A) **cumulating 255 citations**, 1 in *Nucleic Acids Research* (Oxford, IF 10.16) with **764 citations**, 2 application notes in *Bioinformatics* (Oxford, IF 7.31), 2 articles (one with **253 citations**) at *AMIA Symposiums* which is one of the best place to publish in this field. 2 articles (one with **82 citations**) in *BMC Biomedical Semantics* (IF 2.41).
- **Semantic Web**: 2 articles & 3 posters/demos (all cumulating **158 citations**) in *International Semantic Web Conference* (CORE A), the main conference in the domain. Plus, the 1<sup>st</sup> prize at the 2010 Semantic Web

Challenge and a corresponding publication (**111 citations**) in *Journal of Web Semantics* (Elsevier, IF 2.76). One awarded paper at Web Science Conference.

- NLP, text mining & information retrieval: 4 conferences or workshop articles related to text mining and language in biomedicine (LREC'16, *PolTAL'14*, *IDEAS'14*, *JADT'14*, *LBM'13*). One article in *Information Retrieval* (Springer, IF 0.80) and in *Knowledge Discovery in Bioinformatics* (IGI Global) both cumulating **53 citations**).
- Agronomy: multiple poster-demos and workshop papers recently published in this new field of application. One article as 1<sup>st</sup> author in *Computers and Electronics in Agriculture* (Elsevier, IF 2.5). One article as 1<sup>st</sup> author in *Computers and Electronics in Agriculture* (Elsevier, IF 2.5). Two group articles in prestigious journals: *Database* (Oxford Academic, IF 3.98) and *PLoS One* (PLoS, 2.7).
- Distributed systems: 1 article in the reference journal for the topic of agent-grid integration, *Multiagent and Grid systems* (IOS Press, CORE B) as well as 1 article in the *Int. Workshop on Service-Oriented Computing: Agents, Semantics, and Engineering*. Plus 1 article in *Applied Artificial Intelligence* (Taylor & Francis, IF 0.65, CORE B) with **24 citations**.
- French conferences: Such as Journées francophones d'Ingénierie des Connaissances, or d'Informatique médicale or du Traitement Automatique des Langues Naturelles, or sur les Systèmes Multi-Agents », or des Ontologies or de Recherche d'Information et Applications. 9 French publications (over 20) are direct French versions of English papers; others are usually preliminary work.

## SUMMARY OF TEACHING ACTIVITIES

- **9 years of various academic teaching** (~1400h ~TD) to different kind of students of mixed levels. Described in specific section.
- Teacher at [Polytech Montpellier Engineering School](#). My teaching activities were paused from 2015 to 2019 during my mobility and the return phase of my Marie Curie project.
- 2012-2015: [Polytech Montpellier iPad for students](#) project. I run a working group of 70 teachers interested in pedagogical innovation using ICT and iPad, in and out of the classroom.
- Preparation of lectures/tutorials/technical work, evaluation tasks (exam, corrections, jury), projects and internships management, administrative responsibilities. Some classes in English from 2010 to 2012.
- 2012: One full series of 8 lectures given to Polytech students available on video on [iTunesU: Internet Application and Interoperability](#) (AIOP).
- *Lectures*: [Structure and Interpretation of Computer Programs](#), introduction to algorithmic and programming with Scheme/Maple, [French Informatics and Internet Certificate](#) (Open/MS Office, e-learning platforms, etc.), Internet languages (HTML, Java/Javascript, PHP, etc.), [Computer Architecture](#) (representation, CPU/Memory, MIPS language), Algorithmic & Programming (ADA, basic algorithmics, data structures), [Internet Application and Interoperability](#) (Web application architectures, Web technologies, XML, Web services, J2EE, .NET), [Semantic Web](#) (Ontologies 101, technologies & languages, applications).
- *Internship supervision*: technical BSc. (mathematics & computer science), MSc students in computer science.

## TECHNICAL SKILLS

- Programming languages: functional/applicative (Lisp, Scheme) or object-oriented (Java) or imperative (Ada, Maple). Some knowledge of MIPS assembler.
- Java & JEE framework technologies (JDBC, Spring, Eclipse).
- Service Oriented Architectures and Web applications. Web services in SOAP/WSDL (Axis) & REST (RestLet).
- Biomedical terminologies and ontologies (SNOMEDCT, MeSH, UMLS, OBO) as well as Semantic Web technologies (RDF/OWL/SKOS/SPARQL).
- Database systems (SQL), good experience with MySQL/JDBC and information system modeling language (UML, BPMN).
- Web languages & technologies (XML, HTML, Javascript, CSS, PHP/MySQL, JSON).
- Distant learning / e-learning platforms e.g., WebCT, Claroline, Moodle.
- MadKit multi-agent platform (developed within the SMILE team at LIRMM).



## PERSONNAL TOPICS

---

- Experiences in several associations (student, sportive, social ones). Summer jobs from 1996 to 2002 in agriculture and wineries.
- Music, travelling (several trips in Europe, America, South America and Asia.), reading (novel and press).
- Rock climbing (indoor/outdoor) and other outdoor sports (mountaineering, ice-climbing, hiking, etc.).

## LANGUAGES

---

- French: Mother tongue.
- English: Very good (school & working knowledge), lived 6 years the USA.
- Spanish: A few skills (learnt at school).
- Strong international orientation of work (publications & thesis manuscript written in English, international PhD defense jury, international postdoc).

## REFEREES

---

- Pr. Stefano A. Cerri, University of Montpellier – LIRMM – [cerri@lirmm.fr](mailto:cerri@lirmm.fr)
- Pr. Mark A. Musen, Stanford University – BMIR – [musen@stanford.edu](mailto:musen@stanford.edu)
- Pr. Nigam H. Shah, Stanford University – BMIR – [nigam@stanford.edu](mailto:nigam@stanford.edu)
- Pr. Michael N. Huhns, University of South Carolina – CIT – [huhns@sc.edu](mailto:huhns@sc.edu)

# LIST OF PUBLICATIONS

<b>Journal</b>	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]	19
<b>International Conference</b>	[24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]	21
<b>Serie</b>	[45]	1
<b>Workshop</b>	[46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]	15
<b>National (French) Conf.</b>	[61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80]	20
<b>Editor</b>	[81, 82]	2
<b>Dissertation</b>	[83, 84]	2
<b>Poster &amp; Demonstration</b>	[85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110]	26
<b>Report</b>	[111, 112, 113, 114, 115, 116]	6

## JOURNAL

---

- [CJ1] Pierre Monnin, Joël Legrand, Graziella Husson, Patrice Ringot, Andon Tchechmedjiev, **Clement Jonquet**, Amedeo Napoli, and Adrien Coulet. PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, IN PRESS, 2019.
- [CJ2] Andon Tchechmedjiev, Amine Abdaoui, Vincent Émonet, Stella Zevio, and **Clement Jonquet**. SIFR Annotator: Ontology-Based Semantic Annotation of French Biomedical Text and Clinical Notes. *BMC Bioinformatics*, 19:405–431, December 2018.
- [CJ3] Aravind Venkatesan, Gildas Tagny, Nordine El Hassouni, Imene Chentli, Valentin Guignon, **Clement Jonquet**, Manuel Ruiz, and Pierre Larmande. Agronomic Linked Data: a knowledge system to enable integrative biology in Agronomy. *PLoS One*, 13(11):e0198270, November 2018.
- [CJ4] Lisa Harper, Jacqueline Campbell, Ethalinda KS Cannon, Sook Jung, Dorrie Main, Monica Poelchau, Ramona Walls, Carson Andorf, Elizabeth Arnaud, Tanya Berardini, Clayton Birkett, Steve Cannon, James Carson, Bradford Condon, Laurel Cooper, Nathan Dunn, Chris Elsik, Andrew Farmer, Stephen Ficklin, David Grant, Emily Grau, Nic Herndon, Zhi-Liang Hu, Jodi Humann, Pankaj Jaiswal,

- Clement Jonquet**, Marie-Angélique Laporte, Pierre Larmande, Gerard Lazo, Fiona McCarthy, Naama Menda, Christopher Mungall, Monica Munoz-Torres, Sushma Naithani, Rex Nelson, Daureen Nesdill, Carissa Park, James Reecy, Leonore Reiser, Lacey-Anne Sanderson, Taner Sen, Margaret Staton, Sabarinath Subramaniam, Marcela Karey Tello-Ruiz, Victor Unda, Deepak Unni, Liya Wang, Doreen Ware, Jill Wegrzyn, Jason Williams, and Margaret Woodhouse. AgBioData Consortium Recommendations for Sustainable Genomics and Genetics Databases for Agriculture. *Database*, page bay088, September 2018.
- [CJ5] **Clement Jonquet**, Anne Toulet, Biswanath Dutta, and Vincent Emonet. Harnessing the power of unified metadata in an ontology repository: the case of AgroPortal. *Data Semantics*, pages 1–31, August 2018.
- [CJ6] Amina Annane, Zohra Bellahsene, Faïçal Azouaou, and **Clement Jonquet**. Building an effective and efficient background knowledge resource to enhance ontology matching. *Web Semantics*, 51:51–68, August 2018.
- [CJ7] Juan Antonio Lossio-Ventura, Jiang Bian, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. A novel framework for biomedical entity sense induction. *Biomedical Informatics*, 84:31–41, August 2018.
- [CJ8] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Soumia Melzi, Jitendra Jonnagaddala, and **Clement Jonquet**. Enhanced Functionalities for Annotating and Indexing Clinical Text with the NCBO Annotator+. *Bioinformatics*, page 3, January 2018.
- [CJ9] Esther Dzale Yeumo, Michael Alaux, Elizabeth Arnaud, Sophie Aubin, Ute Baumann, Patrice Buche, Laurel Cooper, Robert P. Davey, Richard A. Fulss, **Clement Jonquet**, Marie-Angélique Laporte, Pierre Larmande, Cyril Pommier, Vassilis Protonotarios, Carmen Reverte, Rosemary Shrestha, Imma Subirats, Aravind Venkatesan, Alex Whan, and Hadi Quesneville. Developing data interoperability through standards: a wheat community use case. *F1000 Research*, 6(1843), December 2017.
- [CJ10] **Clement Jonquet**, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzale’Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, and Pierre Larmande. AgroPortal: an ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144:126–143, January 2018.
- [CJ11] Philippe Lemoisson, Guillaume Surroca, **Clement Jonquet**, and Stefano A. Cerri. ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph. *Knowledge and Learning*, 12(2):119–145, May 2018.
- [CJ12] Marcos Martinez-Romero, **Clement Jonquet**, Martin J. O’Connor, John Graybeal, Alejandro Pazos, and Mark A. Musen. NCBO Ontology Recommender 2.0: An Enhanced Approach for Biomedical Ontology Recommendation. *Biomedical Semantics*, 8(21), June 2017.
- [CJ13] Juan-Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction: overview and a new methodology. *Information Retrieval, Special issue on Medical Information Retrieval*, 19(1):59–99, August 2015.
- [CJ14] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Towards a mixed approach to extract biomedical terms from text corpus. *Knowledge Discovery in Bioinformatics*, 4(1):15, 2014.
- [CJ15] **Clement Jonquet**, Paea LePendu, Sean Falconer, Adrien Coulet, Natalya F. Noy, Mark A. Musen, and Nigam H. Shah. NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. *Web Semantics*, 9(3):316–324, September 2011. 1st prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC’10, Shanghai, China.
- [CJ16] Christophe Roeder, **Clement Jonquet**, Nigam H. Shah, William A. Baumgartner Jr, and Lawrence Hunter. A UIMA Wrapper for the NCBO Annotator. *Bioinformatics*, 26(14):1800–1801, May 2010.
- [CJ17] **Clement Jonquet**, Mark A. Musen, and Nigam H. Shah. Building a Biomedical Ontology Recommender Web Service. *Biomedical Semantics*, 1(S1), June 2010. Selected in Pr. R. Altman’s 2011 Year in Review at AMIA TBI.
- [CJ18] Nigam H. Shah, Nipun Bhatia, **Clement Jonquet**, Daniel L. Rubin, Annie P. Chiang, and Mark A. Musen. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, 10(9:S14), September 2009.
- [CJ19] Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, and Mark A.

- Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(web server):170–173, May 2009.
- [CJ20] Nigam H. Shah, **Clement Jonquet**, Annie P. Chiang, Atul J. Butte, Rong Chen, and Mark A. Musen. Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. *BMC Bioinformatics*, 10(2:S1), February 2009.
- [CJ21] **Clement Jonquet**, Pascal Dugenie, and Stefano A. Cerri. Agent-Grid Integration Language. *Multiagent and Grid Systems*, 4(2):167–211, 2008.
- [CJ22] Pascal Dug  nie, Philippe Lemoisson, **Clement Jonquet**, and Monica Crub  zy. The Grid Shared Desktop: a bootstrapping environment for collaboration. *Advanced Technology for Learning, Special issue on Collaborative Learning*, 3(4):241–249, 2006.
- [CJ23] **Clement Jonquet** and Stefano A. Cerri. The STROBE model: Dynamic Service Generation on the Grid. *Applied Artificial Intelligence, Special issue on Learning Grid Services*, 19(9-10):967–1013, October-November 2005.

---

#### INTERNATIONAL CONFERENCE

- [CJ24] Biswanath Dutta, Anne Toulet, Vincent Emonet, and **Clement Jonquet**. New Generation Metadata vocabulary for Ontology Description and Publication. In E. Garoufallou, S. Virkus, R. Siatri, and D. Koutsomiha, editors, *11th Metadata and Semantics Research Conference, MTSR’17*, volume 755 of *Communications in Computer and Information Science*, pages 173–185, Tallinn, Estonia, November 2017. Springer.
- [CJ25] Philippe Lemoisson, Guillaume Surroca, **Clement Jonquet**, and Stefano A. Cerri. ViewpointS: When Social Ranking Meets the Semantic Web. In V. Rus and Z. Markov, editors, *30th International Florida Artificial Intelligence Research Society Conference, FLAIRS’17*, pages 329–334, Marco Island, FL, USA, May 2017. AAAI Press.
- [CJ26] Sol  ne Eholi  , Mike-Donald Tapi-Nzali, Sandra Bringay, and **Clement Jonquet**. MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. In A. Paschke, A. Burger, A. Splendiani, M.S. Marshall, and P. Romano, editors, *9th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS’16*, page 10, Amsterdam, The Netherlands, December 2016.
- [CJ27] Amina Annane, Zohra Bellahsene, Faical Azouaou, and **Clement Jonquet**. Selection and Combination of Heterogeneous BK to Enhance Biomedical Ontology Matching. In E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, editors, *20th International Conference on Knowledge Engineering and Knowledge Management, EKAW’16*, volume 10024 of *Lecture Notes in Artificial Intelligence*, pages 19–33, Bologna, Italy, November 2016. Springer.
- [CJ28] Guillaume Surroca, Philippe Lemoisson, **Clement Jonquet**, and Stefano A. Cerri. Subjective and generic distance in ViewpointS: an experiment on WordNet. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS’16*, number 11, page 6, Nimes, France, June 2016. ACM.
- [CJ29] Amina Annane, Vincent Emonet, Faical Azouaou, and **Clement Jonquet**. Multilingual Mapping Reconciliation between English-French Biomedical Ontologies. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS’16*, number 13, page 12, Nimes, France, June 2016. ACM.
- [CJ30] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Automatic Biomedical Term Polysemy Detection. In *10th International Conference on Language Resources and Evaluation, LREC’16*, pages 23–28, Portoroz, Slovenia, May 2016. European Language Resources Association.
- [CJ31] Guillaume Surroca, Philippe Lemoisson, **Clement Jonquet**, and Stefano A. Cerri. Preference Dissemination by Sharing ViewpointS : Simulating Serendipity. In *7th International Conference on Knowledge Engineering and Ontology Development KEOD’15*, volume 2, pages 402–409, Lisbon, Portugal, November 2015.
- [CJ32] Soumia Melzi and **Clement Jonquet**. Scoring semantic annotations returned by the NCBO Annotator. In A. Paschke, A. Burger, P. Romano, M.S. Marshall, and A. Splendiani, editors, *7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS’14*, volume 1320 of *CEUR Workshop Proceedings*, page 15, Berlin, Germany, December 2014. CEUR-WS.org.
- [CJ33] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Yet Another Ranking Function for Automatic Multiword Term Extraction. In A. Przepiorkowski and M.

- Ogrodniczuk, editors, *9th International Conference on Natural Language Processing, PolTAL'14*, volume 8686 of *Lecture Notes in Artificial Intelligence*, pages 52–64, Warsaw, Poland, September 2014. Springer.
- [CJ34] Julien Grosjean, Lina F. Soualmia, Khedidja Bouarech, **Clement Jonquet**, and Stefan J. Darmoni. An Approach to Compare Bio-Ontologies Portals. In C. Lovis, B. Séroussi, A. Hasman, L. Pape-Haugaard, O. Saka, and S.K. Andersen, editors, *26th International Conference of the European Federation for Medical Informatics, MIE'14*, volume 205 of *Studies in Health Technology and Informatics*, pages 1008–1012, Istanbul, Turkey, September 2014. IOS Press.
- [CJ35] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Integration of Linguistic and Web Information to Improve Biomedical Terminology Extraction. In A-M. Almeida, J. Bernardino, and E. F. Gomes, editors, *18th International Database Engineering & Applications Symposium, IDEAS'14*, pages 265–269, Porto, Portugal, July 2014. ACM.
- [CJ36] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction. In *5th International Symposium on Languages in Biology and Medicine, LBM'13*, pages 45–49, Tokyo, Japan, December 2013. Database Center for Life Science.
- [CJ37] **Clement Jonquet**, Paea LePendu, Sean M. Falconer, Adrien Coulet, Natalya F. Noy, Mark A. Musen, and Nigam H. Shah. NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. In C. Bizer and D. Maynard, editors, *Semantic Web Challenge, 9th International Semantic Web Conference, ISWC'10*, page 8, Shanghai, China, November 2010. 1st prize.
- [CJ38] Paea LePendu, Natalya F. Noy, **Clement Jonquet**, Paul R. Alexander, Nigam H. Shah, and Mark A. Musen. Optimize First, Buy Later: Analyzing Metrics to Ramp-up Very Large Knowledge Bases. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *9th International Semantic Web Conference, ISWC'10*, volume 6496 of *Lecture Notes in Computer Science*, pages 486–501, Shanghai, China, November 2010. Springer.
- [CJ39] Gautam K. Parai, **Clement Jonquet**, Rong Xu, Mark A. Musen, and Nigam H. Shah. The Lexicon Builder Web service: Building Custom Lexicons from two hundred Biomedical Ontologies. In *American Medical Informatics Association Annual Symposium, AMLA'10*, Washington, DC, USA, November 2010.
- [CJ40] Amir Ghazvinian, Natasha F. Noy, **Clement Jonquet**, Nigam H. Shah, and Mark A. Musen. What Four Million Mappings Can Tell You about Two Hundred Ontologies. In A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, *8th International Semantic Web Conference, ISWC'09*, volume 5823 of *Lecture Notes in Computer Science*, pages 229–242, Washington DC, USA, November 2009. Springer.
- [CJ41] **Clement Jonquet**, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMLA-TBI'09*, pages 56–60, San Francisco, CA, USA, March 2009.
- [CJ42] **Clement Jonquet**, Mark A. Musen, and Nigam H. Shah. A System for OntologyBased Annotation of Biomedical Data. In A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, editors, *International Workshop on Data Integration in the Life Sciences, DILS'08*, volume 5109 of *Lecture Notes in BioInformatics*, pages 144–152, Evry, France, June 2008. Springer.
- [CJ43] Stefano A. Cerri, Monica Crubézy, Pascal Dugénie, **Clement Jonquet**, and Phillippe Lemoisson. The Grid Shared Desktop for CSCL. In P. Cunningham and M. Cunningham, editors, *eChallenges 2006 Conference*, volume 3 of *Information and Communication Technologies and the Knowledge Economy*, pages 1493–1499, Barcelona, Spain, October 2006. IOS Press.
- [CJ44] **Clement Jonquet** and Stefano A. Cerri. i-dialogue: modeling agent conversation by streams and lazy evaluation. In *International Lisp Conference, ILC'05*, pages 219–228, Stanford University, CA, USA, June 2005.

---

## SERIE

- [CJ45] **Clement Jonquet**, Marc Eisenstadt, and Stefano A. Cerri. Learning agents and Enhanced Presence for generation of services on the Grid. In P. Ritrovato, C. Allison, S.A. Cerri, T. Dimitrakos, M. Gaeta, and S. Salerno, editors, *Towards the Learning GRID: advances in Human Learning Services*, volume 127 of *Frontiers in Artificial Intelligence and Applications*, pages 203–213. IOS Press, November 2005.

- [CJ46] Andon Tchechmedjiev and **Clement Jonquet**. Enrichment of French Biomedical Ontologies with UMLS Concepts and Semantic Types for Biomedical Named Entity Recognition Through Ontological Semantic Annotation. In *Workshop on Language, Ontology, Terminology and Knowledge Structures, LOTKS'17*, number W17-7007, page 8, Montpellier, France, September 2017. ACL.
- [CJ47] **Clement Jonquet**. Challenges for ontology repositories and applications to biomedicine & agronomy. In J.L. Lossio-Ventura and H. Alatrasta-Salas, editors, *4th Annual International Symposium on Information Management and Big Data, SIMBig'17*, volume 2029 of *CEUR Workshop Proceedings*, pages 25–37, Lima, Peru, September 2017. Keynote Speaker Paper.
- [CJ48] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, and **Clement Jonquet**. ICD10 Coding of Death Certificates with the NCBO and SIFR Annotator(s) at CLEF eHealth 2017 Task 1. In *Working Notes of CLEF eHealth Evaluation Lab*, volume 1866 of *CEUR Workshop Proceedings*, page 16, Dublin, Ireland, September 2017.
- [CJ49] Pierre Monnin, **Clement Jonquet**, Joel Legrand, Amedeo Napoli, and Adrien Coulet. PGxO: A very lite ontology to reconcile pharmacogenomic knowledge units. In *Network Tools and Applications in Biology Workshop, NETTAB'17*, Preprints, page 4, Palermo, Italy, October 2017. PeerJ. Peer reviewed by NETTAB'17 PC.
- [CJ50] **Clement Jonquet**, Anne Toulet, and Vincent Emonet. Two years after: a review of vocabularies and ontologies in AgroPortal. In *International Workshop on sources and data integration in agriculture, food and environment using ontologies, IN-OVIVE'17*, page 13, Montpellier, France, July 2017. EFITA.
- [CJ51] **Clement Jonquet**, Vincent Emonet, and Mark A. Musen. Roadmap for a multilingual BioPortal. In J. Gracia, J.P. McCrae, and G. Vulcu, editors, *4th Workshop on the Multilingual Semantic Web, MSW'15*, volume 1532 of *CEUR Workshop Proceedings*, pages 15–26, Portoroz, Slovenia, June 2015.
- [CJ52] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. SIFR project: The Semantic Indexing of French Biomedical Data Resources. In J.A. Lossio-Ventura and H. Alatrasta-Salas, editors, *1st International Symposium on Information Management and Big Data, SIMBig'14*, volume 1318 of *CEUR Workshop Proceedings*, pages 58–61, Cusco, Peru, September 2014.
- [CJ53] Julien Grosjean, Lina F. Soualmia, Khedidja Bouarech, **Clement Jonquet**, and Stefan J. Darmoni. Comparing BioPortal and HeTOP: towards a unique biomedical ontology portal? In *2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBIO'14*, page 11, Granada, Spain, April 2014.
- [CJ54] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Biomedical Terminology Extraction: A new combination of Statistical and Web Mining Approaches. In E. Nee, J-M. Daube, M. Valette, and S. Fleury, editors, *12th International Workshop on Statistical Analysis of Textual Data, JADT'14*, pages 421–432, Paris, France, June 2014.
- [CJ55] **Clement Jonquet**, Nigam H. Shah, and Mark A. Musen. Prototyping a Biomedical Ontology Recommender Service. In *Bio-Ontologies: Knowledge in Biology, SIG, ISMB-ECCB'09*, pages 65–68, Stockholm, Sweden, July 2009.
- [CJ56] Pascal Dugénie, **Clement Jonquet**, and Stefano A. Cerri. The Principle of Immanence in GRID-Multiagent Integrated Systems. In R. Meersman, Z. Tari, and P. Herrero, editors, *4th International Workshop On Agents and Web Services Merging in Distributed Environments, AWeSOMe'08, OTM 2008 Workshops*, volume 5333 of *Lecture Notes in Computer Science*, pages 98–107, Monterrey, Mexico, November 2008. Springer.
- [CJ57] **Clement Jonquet**, Pascal Dugénie, and Stefano A. Cerri. Service-Based Integration of Grid and Multi-Agent Systems Models. In R. Kowalczyk, M.N. Huhns, M. Klusch, Z. Maamar, and Q.B. Vo, editors, *International Workshop on Service-Oriented Computing: Agents, Semantics, and Engineering, SOCASE'08*, volume 5006 of *Lecture Notes in Computer Science*, pages 56–68, Estoril, Portugal, May 2008. Springer.
- [CJ58] Frédéric Duvert, **Clement Jonquet**, Pascal Dugénie, and Stefano A. Cerri. AgentGrid Integration Ontology. In R. Meersman, Z. Tari, and P. Herrero, editors, *2nd International Workshop on Agents, Web Services and Ontologies Merging, AWeSOMe'06*, volume 4277 of *Lecture Notes in Computer Science*, pages 136–146, Montpellier, France, November 2006. Springer.
- [CJ59] **Clement Jonquet** and Stefano A. Cerri. Agents Communicating for Dynamic Service Generation. In *1st International Workshop on Grid Learning Services, GLS'04*, pages 39–53, Maceio, Brazil, September 2004.

- [CJ60] Stefano A. Cerri, Marc Eisenstadt, and **Clement Jonquet**. Dynamic Learning Agents and Enhanced Presence on the Grid. In *3rd International LeGE-WG Workshop: Grid Infrastructure to Support Future Technology Enhanced Learning*, Berlin, Germany, December 2003. Electronic Workshops in Computing.

---

NATIONAL (FRENCH) CONFERENCE

---

- [CJ61] **Clement Jonquet**. Maîtriser une technologie de gestion des ontologies et vocabulaires en France : défis et enjeux. In *SemWebPro Conference*, page 2, Paris, France, November 2018.
- [CJ62] Fabienne Kettani, Stéphane Schneider, Sophie Aubin, Robert Bossy, Claire François, **Clement Jonquet**, Andon Tchechmedjiev, and Anne Toulet Claire Nédellec. Projet Visa<sup>TM</sup> : l'interconnexion OpenMinTeD – AgroPortal – ISTE<sup>X</sup>, un exemple de service de Text et Data Mining pour les scientifiques français. In Sylvie Rawnez, editor, *29<sup>èmes</sup> Journées Francophones d'Ingénierie des Connaissances, IC'18, Poster Session*, pages 247–249, Nancy, France, July 2018.
- [CJ63] Amine Abdaoui, Andon Tchechmedjiev, William Digan, Sandra Bringay, and **Clement Jonquet**. French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français. In *4<sup>ème</sup> Symposium sur l'Ingénierie de l'Information Médicale, SIIM'17*, page 10, Toulouse, France, November 2017.
- [CJ64] Anne Toulet, Vincent Emonet, and **Clement Jonquet**. Modèle de métadonnées dans un portail d'ontologies. In G. Diallo and O. Kazar, editors, *6<sup>èmes</sup> Journées Francophones sur les Ontologies, JFO'16*, Bordeaux, France, October 2016. Best paper award.
- [CJ65] **Clement Jonquet**, Amina Annane, Khedidja Bouarech, Vincent Emonet, and Soumia Melzi. SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *16<sup>th</sup> Journées Francophones d'Informatique Médicale, JFIM'16*, page 16, Genève, Suisse, July 2016.
- [CJ66] Solène Eholié, Mike Donald Tapi Nzali, Sandra Bringay, and **Clement Jonquet**. MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein. In *2<sup>ème</sup> Atelier sur l'Intelligence Artificielle et la Santé*, page 7, Montpellier, France, June 2016.
- [CJ67] Amina Annane, Vincent Emonet, Faical Azouaou, and **Clement Jonquet**. Réconciliation d'alignements multilingues dans BioPortal. In Nathalie Pernelle, editor, *27<sup>èmes</sup> Journées Francophones d'Ingénierie des Connaissances, IC'16*, number 18, page 12, Montpellier, France, June 2016.
- [CJ68] **Clement Jonquet**, Esther Dzalé-Yeumo, Elizabeth Arnaud, and Pierre Larmande. AgroPortal: a proposition for ontology-based services in the agronomic domain. In *3<sup>ème</sup> atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du Vivant et de l'Environnement, IN-OVIVE'15*, page 5, Rennes, France, June 2015.
- [CJ69] Guillaume Surroca, Philippe Lemoisson, **Clement Jonquet**, and Stefano A. Cerri. Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité. In *26<sup>èmes</sup> Journées Francophones d'Ingénierie des Connaissances, IC'15*, page 12, Rennes, France, June 2015.
- [CJ70] Juan-Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Prédiction de la polysémie pour un terme biomédical. In E. Gaussier, editor, *12<sup>ème</sup> Conférence en Recherche d'Information et Applications, CORLA'15*, pages 437–452, March, Paris, France 2015.
- [CJ71] **Clement Jonquet** and Mark A. Musen. Gestion du multilinguisme dans un portail d'ontologies: étude de cas pour le NCBO BioPortal. In C. Roche, R. Costa, and E. Coudyzer, editors, *Terminology & Ontology : Theories and applications Workshop, TOTh'14*, page 2, Brussels, Belgium, December 2014.
- [CJ72] **Clement Jonquet**, Christophe Fiorio, Philippe Papet, Stéphanie Belin-Mejean, Claudine Pastor, and 'Cellule iPad des enseignants de Polytech Montpellier'. REX : Innovation pédagogique via l'utilisation de tablettes numériques à Polytech Montpellier. In D. Cassagne and C. Jonquet, editors, *9<sup>ème</sup> conférence des Technologies de l'Information et de la Communication pour l'Enseignement, TICE'14, Session Retour d'Expérience (REX)*, pages 97–106, Béziers, France, November 2014.
- [CJ73] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. Extraction automatique de termes combinant différentes informations. In Brigitte Bigi, editor, *21<sup>ème</sup> Traitement Automatique des Langues Naturelles, TALN'14*, volume 2, pages 407–412, Marseille, France, July 2014.
- [CJ74] Guillaume Surroca, Philippe Lemoisson, **Clement Jonquet**, and Stefano A. Cerri. Construction et évolution de connaissances par confrontation de points de vue: prototype pour la recherche d'information

scientifique. In Catherine FaronZucker Catherine Roussey, editor, *25èmes Journées Francophones d'Ingénierie des Connaissances, IC'14*, page 12, Clermont-Ferrand, France, Mai 2014.

- [CJ75] **Clement Jonquet**, Adrien Coulet, Nigam H. Shah, and Mark A. Musen. Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical. In S. Despres, editor, *21èmes Journées Francophones d'Ingénierie des Connaissances, IC'10*, pages 271–282, Nîmes, France, June 2010.
- [CJ76] **Clement Jonquet**, Nigam H. Shah, and Mark A. Musen. Un service Web pour l'annotation sémantique de données biomédicales avec des ontologies. In M. Fieschi, P. Staccini, O. Bouhaddou, and C. Lovis, editors, *13èmes Journées Francophones d'Informatique Médicale, JFIM'09*, volume 17 of *Informatique et Santé*, Nice, France, April 2009.
- [CJ77] **Clement Jonquet**, Pascal Dugénie, and Stefano A. Cerri. Intégration orientée service des modèles Grid et Multi-Agents. In V. Chevrier and M-P. Huget, editors, *14èmes Journées Francophones sur les Systèmes Multi-Agents, JFSMA'06*, pages 271–274, Annecy, France, October 2006. Hermès.
- [CJ78] **Clement Jonquet** and Stefano A. Cerri. Agents as Scheme Interpreters: Enabling Dynamic Specification by Communicating. In *14th Congrès Francophone AFRIF-AFLA de Reconnaissance des Formes et Intelligence Artificielle, RFLA'04*, volume 2, pages 779–788, Toulouse, France, January 2004.
- [CJ79] **Clement Jonquet** and Stefano A. Cerri. Apprentissage issu de la communication pour des agents cognitifs. In J-P. Briot and K. Ghédira, editors, *11èmes Journées Francophones sur les Systèmes Multi-Agents, JFSMA'03*, pages 83– 87, Hammamet, Tunisia, November 2003. Hermès.
- [CJ80] **Clement Jonquet** and Stefano A. Cerri. Cognitive Agents Learning by Communicating. In *Colloque Agents Logiciels, Coopération, Apprentissage et Activité Humaine, ALCAA'03*, pages 29–39, Bayonne, France, September 2003.

---

#### EDITOR

- [CJ81] Alsayed Algergawy, Naouel Karam, Friederike Klan, and **Clement Jonquet**, editors. *Proceedings of the 2nd International Workshop on Semantics for Biodiversity, S4BioDiv'17*, volume 1933 of *CEUR Workshop Proceedings*, Vienna, Austria, October 2017.
- [CJ82] Pierre Larmande, Elizabeth Arnaud, Isabelle Mougenot, **Clement Jonquet**, Thérèse Libourel, and Manuel Ruiz, editors. *Proceedings of the 1st International Workshop on Semantics for Biodiversity, S4BioDiv'13*, Montpellier, France, May 2013.

---

#### DISSERTATION

- [CJ83] **Clement Jonquet**. *Dynamic Service Generation: Agent interactions for service exchange on the Grid*. PhD thesis, University Montpellier 2, Montpellier, France, November 2006.
- [CJ84] **Clement Jonquet**. Communication agent et interprétation Scheme pour l'apprentissage au méta-niveau. Master thesis, University Montpellier 2, Montpellier, France, June 2003.

---

#### POSTER & DEMONSTRATION

- [CJ85] Nordine El Hassouni, Manuel Ruiz, Anne Toulet, **Clement Jonquet**, and Pierre Larmande. The Agronomic Linked Data (AgroLD) project. In *European conference dedicated to the future use of ICT in the agri-food sector, bioresource and biomass sector, EFITA'17, demonstration session*, page 257, Montpellier, France, July 2017.
- [CJ86] **Clement Jonquet**, Anne Toulet, Vincent Emonet, and Pierre Larmande. AgroPortal: an ontology repository for agronomy. In *European conference dedicated to the future use of ICT in the agri-food sector, bioresource and biomass sector, EFITA'17, demonstration session*, page 261, Montpellier, France, July 2017.
- [CJ87] **Clement Jonquet**, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzale'Yeumo, Vincent Emonet, Valeria Pesce, and Pierre Larmande. AgroPortal: an open repository of ontologies and vocabularies for agriculture and nutrition data. In Ben Schaap, editor, *GODAN Summit Open Data Research Symposium on Agriculture and Nutrition, GODAN'16*, New York, NY, USA, September 2016.
- [CJ88] **Clement Jonquet**, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzale'Yeumo, Vincent Emonet, John Graybeal, Mark A. Musen, Cyril Pommier, and Pierre Larmande. Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In P. Jaiswal and R. Hoehndorf, editors, *7th*

*International Conference on Biomedical Ontologies, ICBO'16, Demo Session*, volume 1747 of *CEUR Workshop Proceedings*, page 3, Corvallis, Oregon, USA, August 2016.

- [CJ89] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. A Way to Automatically Enrich Biomedical Ontologies. In *19th International Conference on Extending Database Technology, EDBT'16, Poster Session*, number 305, page 2, Bordeaux, France, March 2016. OpenProceedings.org.
- [CJ90] **Clement Jonquet**, Esther Dzalé-Yeumo, Elizabeth Arnaud, Pierre Larmande, Anne Toulet, and Marie-Angélique Laporte. AgroPortal : A Proposition for Ontology-Based Services in the Agronomic Domain. In *23rd Plant & Animal Genome Conference, poster session*, page P0343, San Diego, USA, January 2016.
- [CJ91] Soumia Melzi and **Clement Jonquet**. Representing NCBO Annotator results in standard RDF with the Annotation Ontology. In A. Paschke, A. Burger, P. Romano, M.S. Marshall, and A. Splendiani, editors, *7th International Semantic Web Applications and Tools for Life Sciences poster session, SWAT4LS'14*, volume 1320 of *CEUR Workshop Proceedings*, page 5, Berlin, Germany, December 2014. CEUR-WS.org.
- [CJ92] Aravind Venkatesan, Pierre Larmande, **Clement Jonquet**, Manuel Ruiz, and Patrick Valduriez. Facilitating efficient knowledge management and discovery in the Agronomic Sciences. In *4th Plenary Meeting of the Research Data Alliance*, Amsterdam, The Netherlands, September 2014.
- [CJ93] Juan Antonio Lossio-Ventura, **Clement Jonquet**, Mathieu Roche, and Maguelonne Teisseire. BIOTEX: A system for Biomedical Terminology Extraction, Ranking, and Validation. In M. Horridge, M. Rospocher, and J. Ossenbruggen, editors, *13th International Semantic Web Conference, Demonstration, ISWC'14*, volume 1272 of *CEUR Workshop Proceedings*, pages 157–160, Riva del Garda, Italy, October 2014.
- [CJ94] **Clement Jonquet**, Christophe Fiorio, Philippe Papet, Stéphanie Belin-Mejean, and 'Cellule iPad des enseignants de Polytech Montpellier'. Scénarios pédagogiques numériques via l'utilisation de l'iPad par et pour les étudiants de Polytech Montpellier. In T. Karsenti, editor, *2ème Sommet iPad en éducation*, page 1, Montreal, Canada, May 2014. CRIFPE.
- [CJ95] Emmanuel Castanier, **Clement Jonquet**, Soumia Melzi, Pierre Larmande, Manuel Ruiz, and Patrick Valduriez. Semantic Annotation Workflow using BioOntologies. In *Workshop on Crop Ontology and Phenotyping Data Interoperability*, Montpellier, France, April 2014. CGIAR.
- [CJ96] **Clement Jonquet**. BioPortal : ontologies et ressources de données biomédicales à portée de main. In L. Tamine, S. Darmoni, and L. Soualmia, editors, *1ère édition du Symposium sur l'Ingénierie de l'Information Médicale, SIIM'11, Démos*, Toulouse, France, June 2011.
- [CJ97] Patricia L. Whetzel, **Clement Jonquet**, Cherie H. Youn, Michael Dorf, Ray Fergerson, Mark A. Musen, and Nigam H. Shah. The NCBO Annotator: OntologyBased Annotation as a Web Service. In Barry Smith, editor, *International Conference on Biomedical Ontology, Software demonstration session*, pages 302–303, Buffalo, NY, USA, July 2011.
- [CJ98] Patricia L. Whetzel, Nigam H. Shah, Natasha F. Noy, **Clement Jonquet**, Cherie H. Youn, Paul R. Alexander, Michael Dorf, and Mark A. Musen. Ontology-based Tools to Enhance the Curation Workflow. In *4th International Biocuration Conference, Poster Session*, Tokyo, Japan, October 2010.
- [CJ99] Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, **Clement Jonquet**, Adrien Coulet, Cherie Youn, Michael Dorf, and Mark A. Musen. Ontology-based Web Services for Semantic Applications. In *Bio-Ontologies: Semantic Applications in Life Sciences, SIG, Poster session, ISMB'10*, Boston, MA, USA, July 2010.
- [CJ100] Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, **Clement Jonquet**, Adrien Coulet, Nicholas B. Griffith, Cherie H. Youn, Michael Dorf, and Mark A. Musen. Ontology Web Services for Semantic Applications. In *Pacific Symposium on Biocomputing, Poster presentations, PSB'10*, Hawaii, USA, January 2010.
- [CJ101] Nigam H. Shah, Natasha F. Noy, **Clement Jonquet**, Adrien Coulet, Patricia L. Whetzel, Nicholas B. Griffith, Cherie H. Youn, Benjamin Dai, Michael Dorf, and Mark A. Musen. Ontology Services for Semantic Applications in Health and Life Sciences. In *American Medical Informatics Association Annual Symposium, Demonstrations, AMLA'09*, Washington DC, USA, November 2009.
- [CJ102] **Clement Jonquet**, Nigam H. Shah, Cherie H. Youn, Chris Callendar, MargaretAnne Storey, and Mark A. Musen. NCBO Annotator: Semantic Annotation of Biomedical Data. In *8th International Semantic Web Conference, Poster and Demonstration Session, ISWC'09*, Washington DC, USA, November 2009.
- [CJ103] Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Cherie H. Youn, Chris Callendar, Adrien Coulet, Daniel L. Rubin, Barry Smith, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: Ontologies and

- Integrated Data Resources at the Click of the Mouse. In B. Smith, editor, *International Conference on Biomedical Ontology, ICBO'09*, page 197, Buffalo, NY, USA, July 2009.
- [CJ104] Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Cherie H. Youn, Adrien Coulet, Chris Callendar, Daniel L. Rubin, Barry Smith, Margaret-Anne Storey, Christopher G. Chute, and Mark A. Musen. BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. In *Bio-Ontologies: Knowledge in Biology, SIG, Poster session, ISMBECCB'09*, Stockholm, Sweden, July 2009.
- [CJ105] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Cherie H. Youn, Daniel L. Rubin, and Mark A. Musen. BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse. In *17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB'09) and the 8th European Conference on Computational Biology (ECCB'09), Poster session*, Stockholm, Sweden, July 2009.
- [CJ106] Patricia L. Whetzel, Natalya F. Noy, Nigam H. Shah, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Cherie H. Youn, Michael J. Montegut, Daniel L. Rubin, Margaret-Anne Storey, Chris G. Chute, and Mark A. Musen. BioPortal: A Web Repository for Biomedical Ontologies and Data Resources. In *3rd International Biocuration Conference, Poster presentations*, page 97, Berlin, Germany, April 2009.
- [CJ107] Simon N. Twigger, Jennifer Smith, Rajni Nigam, **Clement Jonquet**, and Mark A. Musen. Billion Data Points Trapped in International Data Repository Daring rescue Planned! In *3rd International Biocuration Conference, Poster presentations*, page 34, Berlin, Germany, April 2009.
- [CJ108] Patricia L. Whetzel, Natasha F. Noy, Nigam H. Shah, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Michael J. Montegut, Daniel L. Rubin, Cherie H. Youn, and Mark A. Musen. BioPortal: A Web Repository for Biomedical Ontologies and Ontology-indexed Data Resources. In *Pacific Symposium on Biocomputing, Poster presentations, PSB'09*, page 90, Hawaii, USA, January 2009.
- [CJ109] Mark A. Musen, Nigam H. Shah, Natasha F. Noy, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, James Buntrock, **Clement Jonquet**, Michael Montegut, and Daniel L. Rubin. BioPortal: Ontologies and Data Resources with the Click of a Mouse. In *American Medical Informatics Association Annual Symposium, Demonstrations, AMLA'08*, pages 1223–1224, Washington DC, USA, November 2008.
- [CJ110] Natalya F. Noy, Nigam H. Shah, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, **Clement Jonquet**, Michael Montegut, Daniel L. Rubin, Cherie Youn, and Mark A. Musen. BioPortal: A Web Repository for Biomedical Ontologies and Data Resources. In C. Bizer and A. Joshi, editors, *7th International Semantic Web Conference, Poster and Demonstration Session, ISWC'08*, volume 401 of *CEUR Workshop Proceedings*, Karlsruhe, Germany, October 2008. CEURWS.org.

## REPORT

---

- [CJ111] Valeria Pesce, Jeni Tennison, Lisette Mey, **Clement Jonquet**, Anne Toulet, Sophie Aubin, and Panagiotis Zervas. A Map of Agri-food Data Standards. *F1000 Research*, Technical Report 7-177, Global Open Data for Agriculture and Nutrition (GODAN), February 2018. Not peer reviewed.
- [CJ112] Amina Annane, Zohra Bellahsene, Faical Azouaou, and **Clement Jonquet**. YAM-BIO – Results for OAEI 2017. System paper, LIRMM, University of Montpellier, Montpellier, France, November 2017. Ontology Alignment Evaluation Initiative 2017 Campaign.
- [CJ113] **Clement Jonquet**, Mark A. Musen, and Nigam H. Shah. Help will be provided for this task: Ontology-Based Annotator Web Service. Research report BMIR2008-1317, Stanford University, CA, USA, May 2008.
- [CJ114] Nigam H. Shah, **Clement Jonquet**, and Mark A. Musen. Ontrez project report. Research report BMIR-2007-1289, Stanford University, CA, USA, November 2007.
- [CJ115] **Clement Jonquet** and Stefano A. Cerri. Characterization of the Dynamic Service Generation concept. Research report 06007, University Montpellier 2, France, February 2006.
- [CJ116] **Clement Jonquet**. A framework and ontology for Semantic Grid services: an integrated view of WSMF and WSRF. May 2005. Unpublished draft research report, University Montpellier 2 and KMi, Open University, May 2005.

[CJ-UR1] Anne Toulet, Biswanath Dutta, **Clement Jonquet**, Assessing the Practice of Ontology Metadata: A Survey Result, *Non published Report*, Montpellier, France, June 2018.

[CJ-UR2] Amine Abdaoui, Andon Tchechmedjiev, William Digan, Sandra Bringay, **Clement Jonquet**, French ConText: a Publicly Accessible System for Detecting Negation, Temporality and Experienter in French Clinical Notes, *Biomedical Informatics*, under review - JBI-18-620 - 2nd round, 2018.

[CJ-UR3] Amina Annane, Zohra Bellahsene, Façal Azouaou, **Clement Jonquet**, Using Background Knowledge to Enhance Ontology Matching: a Survey, *Semantic Web*, Rejected - 1525-2737, then 1662-2874, to be resubmitted, 2019.

## (FRENCH) DETAILS DES ACTIVITES D'ENSEIGNEMENT

### EXPERIENCE

- Depuis 2010 : Maitre de Conférences à l'Ecole Polytechnique de Montpellier, composante de l'Université de Montpellier (192h~TD par an). De 2010 à 2015, j'étais responsable de 2 modules en 3ème et 5ème année d'école d'ingénieur et je participais à d'autres enseignements. De 2012 à 2015, j'étais responsable de la 5ème (et dernière) année de la filière « Informatique et Gestion ». J'encadre également des stages de fin d'études et de projets industriels. Mes activités d'enseignement ont été en pause de 2015 à 2019 pendant ma mobilité et la phase de retour de mon projet Marie Curie.
- 2012-2015 : Coordination de la cellule des enseignants de Polytech qui s'intéressent à l'innovation pédagogique à l'aide des tablettes numériques (iPad). Nous avons décrit, testé et réalisé plusieurs scénarios d'utilisation des iPads dans la classe. Je m'occupais également d'une partie de la logistique du projet (1000 iPads).
- 2006-2007 : ATER (complet) à l'Université Montpellier 3. Mon expérience d'enseignement (192h~TD) dans une université différente (arts, lettres, langues, sciences humaines et sociales) a été très enrichissante. Elle m'a permis d'intégrer une autre équipe d'enseignement et de cotoyer un autre public que celui de l'Université Montpellier 2 (sciences et techniques). J'y exerçais également des responsabilités communes et administratives.
- 2003-2006 : Moniteur CIES à l'Université Montpellier 2. Le monitorat fut ma première expérience d'enseignement. En 3 ans (64h~TD/an), elle m'a permis d'exercer à petite échelle beaucoup des tâches de l'enseignant : préparation des cours/TD/TP, participation à l'évaluation (rédaction des sujets, corrections, jury), encadrement de projet et de stage, tâches administratives etc. J'ai aussi suivi un ensemble de formations (expression, théâtre, préparation, projets, etc.) proposées par le Centre d'Initiation à l'Enseignement Supérieur (CIES) de Montpellier.

### RÉCAPITULATIF

	Enseignement	Type	Heures (~TD)/an
(depuis 2013)	Web Sémantique	CM	6
(en 2011 seulement)	Algorithmique et programmation Ada/C	TD/TP	40
Maitre de Conférences Polytech Montpellier (depuis 2010)	Architecture des Ordinateurs	CM	18
		TD/TP	42
	Applications Internet et Interopérabilité	CM	18
		TD/TP	24
	Encadrement stage	-	30

	Encadrement PIFE	-	38
	Divers cours		4
	Divers encadrement	-	10
ATER Univ. Montpellier 3 (2006-2007)	C2i niveau débutant	CM/TD/TP	102
	C2i niveau avancé	CM/TD/TP	60
	Internet/Php/Javascript	CM/TD/TP	30
Moniteur CIES Univ. Montpellier 2 (2003-2006)	Programmation/Scheme/ évaluation	TD	100
		TP	40
	Programmation/Algorith- mique/Maple	TD	24
		TP	20
	Divers cours	CM	9
	Encadrement	-	15
<b>TOTAL :</b>			<b>192*7 ans ~ 1400h (~TD)</b>

## ENSEIGNEMENTS EFFECTUÉS

### ▪ 2013-2015 : **Module « Web Sémantique »**

Public : Ecole d'ingénieur « Informatique et Gestion », 5<sup>ème</sup> année.

Contribution : Cours inspiré de 2 tutoriaux (N. Noy & F. Gandon) et d'un listing d'applications du web sémantique. Gestion complète du module.

Objectifs : Le cours a pour objectif une introduction aux principes et technologies du web sémantique.

URL : <http://mon.univ-montp2.fr/claroline/course/index.php?cid=P1S904>

### ▪ 2010-2015 : **Module « Applications Internet et Interopérabilité »**

Public : Ecole d'ingénieur « Informatique et Gestion », 5<sup>ème</sup> année.

Contribution : Cours effectué en anglais (3 ans). ~300 transparents de cours. Gestion complète du module et des interventions extérieures sur J2EE et .NET.

Objectifs : Le cours a pour objectif la compréhension des architectures d'application Web. Une approche historique est suivie pour faire une revue des différents principes et modèles. Les technologies des applications Web et d'interopérabilité sont également présentées e.g., XML, J2EE, .NET, Web services, etc.

Série de cours sur iTunes.

URL : <http://mon.univ-montp2.fr/claroline/course/index.php?cid=P1S911>

### ▪ 2010-2015 : **Module « Architecture des Ordinateurs »**

Public : Ecole d'ingénieur « Informatique et Gestion », 3<sup>ème</sup> année.

Objectifs : Le cours a pour objectif la compréhension de l'architecture des ordinateurs afin d'acquérir les connaissances de base utiles à la compréhension des autres disciplines de l'informatique. L'accent est notamment mis sur les principes de codage des données et des instructions et sur le fonctionnement de la mémoire et de l'unité centrale de traitement.

Contribution : ~240 transparents de cours et 3 feuilles de TD. Gestion complète du module.

URL : <http://mon.univ-montp2.fr/claroline/course/index.php?cid=M513>

### ▪ 2011 : **Module « Algorithmique et programmation »**

Public : Ecole d'ingénieur « Informatique et Gestion », 3<sup>ème</sup> année.

Contribution : Intervention en TD/TP.

Objectifs : Compréhension des algorithmes comme une description précise et rigoureuse d'une suite d'opérations permettant d'obtenir, en un nombre fini d'étapes, la solution d'un problème. Type abstrait de données. Structure de données. La partie programmation aborde dans un premier temps le langage Ada puis le langage C.

URL : <http://mon.univ-montp2.fr/claroline/course/index.php?cid=PIG51P1S511>

Responsable : Christophe Fiorio – [fiorio@lirmm.fr](mailto:fiorio@lirmm.fr)

■ 2003&2004 : **Module « Introduction à la programmation avec Scheme »**

Public : 2<sup>ème</sup> année Deug MIAS (Mathématique, Informatique et Applications aux Sciences)

Objectifs : Ce module vise à introduire aux étudiants les concepts de base de l'abstraction procédurale, l'abstraction de données et des mécanismes d'évaluation (substitution, environnement, etc.) à l'aide d'un langage de programmation fonctionnel/applicatif, Scheme.

Contribution : Pour ce module, j'ai réalisé conjointement avec un collègue moniteur un ensemble de 8 nouvelles feuilles de TD et 6 nouvelles feuilles de TP ainsi que des encadrements de projets. Je me suis également occupé de l'organisation générale du module (équipe, réunions, réservation des salles, etc.).

Responsable : Stefano A. Cerri – [cerri@lirmm.fr](mailto:cerri@lirmm.fr)

■ 2005 : **Module « Introduction à l'algorithmique et à la programmation (Mapple) »**

Public : 1<sup>ère</sup> année Licence (sciences)

Objectifs : Ce module vise à introduire aux étudiants les concepts de base de la programmation (variable, affectation, structure de contrôle, etc.) à l'aide d'un langage algorithmique puis d'un langage de programmation impérative, Maple.

Contribution : Module pour lequel j'ai intégré une équipe d'enseignement importante, ce qui m'a fait découvrir d'autres aspects de l'enseignement. Les feuilles de TD/TP existaient déjà.

Responsable : Philippe Janssen – [janssen@lirmm.fr](mailto:janssen@lirmm.fr)

URL : <http://ens.math.univ-montp2.fr/SPIP/ULIN101>

■ 2006 : **« Certificat Informatique et Internet »** (C2i niveau débutant et avancé)

Public : tous niveaux/toutes filières (lettres & sciences sociales)

Objectifs : Le C2i est un certificat national qui atteste de la compétence et de la maîtrise des technologies de l'information et de la communication. L'enseignement effectué n'est pas de l'Informatique « pure », mais de l'initiation à l'outil Informatique, à la bureautique et à Internet (forum, mails, HTML, etc.). Cela m'a permis de me confronter à des aspects plus pédagogiques que techniques de l'enseignement. Le public ayant très peu d'expérience en informatique.

Contribution : Participation à l'amélioration d'un cours déjà existant. Contributions administratives et techniques.

Responsable : Patrice Séébold – [seebold@lirmm.fr](mailto:seebold@lirmm.fr)

URL : <http://www.univ-montp3.fr/miap/ens/info/index.html>

■ 2007 : **Module « Informatique de l'Internet »**

Public : 2<sup>ème</sup> année, Licence MASS (Mathématiques Appliquées aux Sciences Sociales)

Objectifs : Module dont l'objectif est d'introduire aux étudiants les langages et les concepts de l'Internet (HTML, Java/Javascript, PHP, etc.).

Contribution : Mise à jour et reprise d'éléments de cours existants.

Responsable : Joël Quinqueton – [jq@lirmm.fr](mailto:jq@lirmm.fr)

URL : <http://www.univ-montp3.fr/miap/ens/MASS/XLIN401/index.htm>

---

## INTERVENTIONS DIVERSES & ENCADREMENTS

- 2010-2015 : Encadrement de 12 stages de fin d'études des étudiants de Polytech Montpellier. Encadrement de 8 projets industriels. Participation à divers jurys.
- 2010-2012 : Cours en Master TIC et Santé (UM2 et Institut Telecom).
- 2006 : Cours en M2P Informatique de l'UM2 dans le module « Informatique Sociale ».
- 2004 : Cours en DEA Informatique de l'UM2 dans le module « Système Multi-Agents ».



# Chapter VII.

## Selected Publications



Shenandoah National Park

The following publications are included in the next pages.

### Journal

[CJ2] Andon Tchechmedjiev, Amine Abdaoui, Vincent Émonet, Stella Zevio, and **Clement Jonquet**. SIFR Annotator: Ontology-Based Semantic Annotation of French Biomedical Text and Clinical Notes. *BMC Bioinformatics*, 19:405–431, December 2018.

[CJ5] **Clement Jonquet**, Anne Toulet, Biswanath Dutta, and Vincent Emonet. Harnessing the power of unified metadata in an ontology repository: the case of AgroPortal. *Data Semantics*, pages 1–31, August 2018.

[CJ8] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Soumia Melzi, Jitendra Jonnagaddala, and **Clement Jonquet**. Enhanced Functionalities for Annotating and Indexing Clinical Text with the NCBO Annotator+. *Bioinformatics*, page 3, January 2018.

[CJ10] **Clement Jonquet**, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé-Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, and Pierre Larmande. AgroPortal: an ontology repository for agronomy. *Computers and Electronics in Agriculture*, 144:126–143, January 2018.

[CJ15] **Clement Jonquet**, Paea LePendou, Sean Falconer, Adrien Coulet, Natalya F. Noy, Mark A. Musen, and Nigam H. Shah. NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. *Web Semantics*, 9(3):316–324, September 2011. 1st prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC’10, Shanghai, China.

### Conference

[CJ41] **Clement Jonquet**, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMLA-TBI’09*, pages 56–60, San Francisco, CA, USA, March 2009.



SOFTWARE

Open Access



# SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes

Andon Tchechmedjiev<sup>1,3\*</sup>, Amine Abdaoui<sup>1</sup>, Vincent Emonet<sup>1</sup>, Stella Zevio<sup>1</sup> and Clement Jonquet<sup>1,2</sup>

## Abstract

**Background:** Despite a wide adoption of English in science, a significant amount of biomedical data are produced in other languages, such as French. Yet a majority of natural language processing or semantic tools as well as domain terminologies or ontologies are only available in English, and cannot be readily applied to other languages, due to fundamental linguistic differences. However, semantic resources are required to design semantic indexes and transform biomedical (text)data into knowledge for better information mining and retrieval.

**Results:** We present the SIFR Annotator (<http://biportal.lirmm.fr/annotator>), a publicly accessible ontology-based annotation web service to process biomedical text data in French. The service, developed during the *Semantic Indexing of French Biomedical Data Resources (2013–2019)* project is included in the SIFR BioPortal, an open platform to host French biomedical ontologies and terminologies based on the technology developed by the US *National Center for Biomedical Ontology*. The portal facilitates use and fostering of ontologies by offering a set of services –search, mappings, metadata, versioning, visualization, recommendation– including for annotation purposes. We introduce the adaptations and improvements made in applying the technology to French as well as a number of language independent additional features –implemented by means of a proxy architecture– in particular annotation scoring and clinical context detection. We evaluate the performance of the SIFR Annotator on different biomedical data, using available French corpora –Quaero (titles from French MEDLINE abstracts and EMEA drug labels) and CépiDC (ICD-10 coding of death certificates)– and discuss our results with respect to the CLEF eHealth information extraction tasks.

**Conclusions:** We show the web service performs comparably to other knowledge-based annotation approaches in recognizing entities in biomedical text and reach state-of-the-art levels in clinical context detection (negation, experienter, temporality). Additionally, the SIFR Annotator is the first openly web accessible tool to annotate and contextualize French biomedical text with ontology concepts leveraging a dictionary currently made of 28 terminologies and ontologies and 333 K concepts. The code is openly available, and we also provide a Docker packaging for easy local deployment to process sensitive (e.g., clinical) data in-house (<https://github.com/sifrproject>).

## Introduction

Biomedical data integration and semantic interoperability are necessary to enable translational research [1–3]. The biomedical community has turned to ontologies and terminologies to describe their data and turn them

into structured and formalized knowledge [4, 5]. Ontologies help to address the data integration problem by playing the role of common denominator. One way of using ontologies is by means of creating semantic annotations. An annotation is a link from an ontology concept to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the concept [6]. In ontology-based –or semantic– indexing, we use these annotations to “bring together” the data elements from the resources. Ontologies help to design semantic indexes of data that leverage the medical

\* Correspondence: [andon.tchechmedjiev@lirmm.fr](mailto:andon.tchechmedjiev@lirmm.fr)

<sup>1</sup>Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier, CNRS, 161, rue Ada, 34095 Montpellier cedex 5, France

<sup>3</sup>LGI2P, IMT Mines Ales, Univ Montpellier, Alès, France

Full list of author information is available at the end of the article



knowledge for better information mining and retrieval. Despite a large adoption of English in science, a significant quantity of biomedical data uses other languages, e.g., French. For instance, clinicians often use the local official administrative language or languages of the countries they operate in to write clinical notes. Besides the existence of various English tools, there are considerably less terminologies and ontologies available in French [7, 8] and there is a strong lack of related tools and services to exploit them. The same is true of languages other than English generally speaking [8]. This lack does not match the huge amount of biomedical data produced in French, especially in the clinical world (e.g., electronic health records).

In the context of the *Semantic Indexing of French Biomedical Data Resources* (SIFR) project ([www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)), we have developed the SIFR BioPortal [9], an open platform to host French biomedical ontologies and terminologies based on the technology developed by the US *National Center for Biomedical Ontology* (NCBO) [10, 11]. The portal facilitates the use and fostering of ontologies by offering a set of services such as search and browsing, mapping hosting and generation, rich semantic metadata description and edition, versioning, visualization, recommendation, community feedback. As of today, the portal contains 28 public ontologies and terminologies (+ two private ones, cf. Table 1), that cover multiple areas of biomedicine, such as the French versions of MeSH, MedDRA, ATC, ICD-10, or WHO-ART but also multilingual ontologies (for which only the French content is parsed) such as Rare Human Disease Ontology, OntoPneumo or Ontology of Nuclear Toxicity.

One of the main motivation to build the SIFR BioPortal was to design the SIFR Annotator (<http://biportal.lirmm.fr/annotator>), a publicly accessible and easily usable ontology-based annotation web service to process biomedical text and clinical notes in French. The annotator service processes raw textual descriptions, tags them with relevant biomedical ontology concepts, expands the annotations using the knowledge embedded in the ontologies and contextualizes the annotations before returning them to the users in several formats such as XML, JSON-LD, RDF or BRAT. We have significantly enhanced the original annotator packaged within the NCBO technology [12, 13], including the addition of scoring, score filtering, lemmatization, and clinical context detection; not to mention some enhancements have not been implemented only for French but have been generalized for the original English NCBO Annotator (or any other annotator based on NCBO technology) through a “proxy” architecture presented by Tchechmedjiev et al. [14]. A preliminary evaluation of the SIFR Annotator has shown that the web service matches the results of previously reported work in French, while

being public, of easy access and use, and turned toward semantic web standards [9]. However, the previous evaluation was of limited scope and new French benchmarks have since been published, which has motivated a more exhaustive evaluation of all the new capabilities mostly with the following corpora: (i) the Quaero corpus (from CLEF eHealth 2015 [15]) which includes French MEDLINE citations in (titles & abstracts) and drug labels from the European Medicines Agency, both annotated with UMLS Semantic Groups and Concept Unique Identifiers (CUIs); (ii) the CépiDC corpus (from CLEF eHealth 2017 [16]) which gathers French death certificates annotated with ICD-10 codes produced by the French epidemiological center for medical causes of death (CépiDC<sup>1</sup>). Additionally, the new contextualization features make SIFR Annotator the first general annotation workflow with a complete implementation of the ConText/NegEx algorithm for French [17]; evaluated on two types of clinical text as reported in a dedicated article (Abdaoui et al: French ConText: a Publicly Accessible System for Detecting Negation, Temporality and Experienter in French Clinical Notes, under review).<sup>2</sup>

The rest of the paper is organized as follows: The **Background** section presents related work pertaining to ontology repositories, semantic annotation tools, and knowledge-based approaches for French biomedical text information extraction. The **Implementation** section describes the SIFR BioPortal, the provenance of the ontologies as well as the architecture and implementation details of the SIFR Annotator and its generic extension mechanism. The **Results and Evaluation** section presents an experimental evaluation of the SIFR Annotator performance through three tasks (named entity recognition, death certificate coding as well as contextual clinical text annotation). The **Discussion** section analyses the merits and limits of our approach through a detailed error analysis and outlines future directions for the improvement of the SIFR Annotator.

## Background

### Biomedical ontology and terminology libraries

In the biomedical domain, multiple ontology libraries (or repositories) have been developed. The OBO Foundry [18] is a reference community effort to help the biomedical and biological communities build their ontologies with an enforcement of design and reuse principles, which has been a tremendous success. The OBO Foundry web application (<http://obofoundry.org>) is an ontology library which serves content to other ontology repositories, such as the NCBO BioPortal [10], OntoBee [19], the EBI Ontology Lookup Service [20] and more recently AberOWL [21]. None of these platforms are multilingual or focus on features

**Table 1** SIFR BioPortal semantic resources

Acronym	Name	Source/Group	Format	#Classes/#Individuals	#Props.	Linguality
MDRFRE	Dictionnaire médical pour les activités réglementaires en matière de médicaments	UMLS/UMLS	RRF	68,980	14	FTO
MSHFRE	Medical Subject Headings, version française	UMLS/UMLS	RRF	27,879	6	FTO
MTHMSTFRE	Terminologie minimale standardisée en endoscopie digestive	UMLS/UMLS	RRF	1700	1	FTO
STY	Réseau Sémantique UMLS	UMLS/UMLS	OWL	133	0	FTO
CIM-10	Classification Internationale des Maladies - 10ème révision	CISMeF/UMLS	OWL	19,853	0	FTO
WHO-ARTFRE	Terminologie des effets indésirables	CISMeF/UMLS	OWL	3483	0	FTO
CISP-2	Classification Internationale des Soins Primaires, deuxième édition	CISMeF/UMLS	OWL	745	4	FTO
CIF	Classification Internationale du Fonctionnement, du handicap et de la santé	CISMeF/UMLS	OWL	1496	2	FTO
SNMIFRE	Systematized Nomenclature of MEDicine, version française	CISMeF/UMLS	OWL	106,291	8	FTO
MEDLINEPLUS	MedlinePlus Health Topics	CISMeF/UMLS	OWL	849	2	FTO
ATCFRE	Classification ATC (anatomique, thérapeutique et chimique)	CISMeF/UMLS	OWL	5768	2	FTO
PDO	CFEF - Prenatal Diagnosis Ontology	LIMICS	OWL	802	0	FMO
ONTOLURGENCES	Ontologie des urgences	LIMICS	OWL	10,031	61	FMO
CCAM	Classification Commune des Actes Médicaux	CISMeF	OWL	9663	8	FOO
ONTOPNEUMO	Ontologie de la pneumologie française.	LIMICS	OWL	1153	22	FMO
TOP-MENELAS	Top ontologie de ONTOMENELAS	LIMICS	OWL	339	298	FMO
LPP	Liste des Produits et Prestations	AMELI/CISMeF	OWL	3746	4	FOO
NABM	Nomenclature des Actes de Biologie Médicale	AMELI/CISMeF	OWL	1055	3	FOO
INM	Ontologie des Interventions Non Médicamenteuses	CEPS/LIRMM	OWL	159	3	FOO
TRANSTHES	Thésaurus de la transfusion sanguine	INIST-CNRS/Loterre	SKOS	2033	0	FOO
MEMOTHES	Thésaurus Psychologie cognitive de la mémoire humaine	INIST-CNRS/Loterre	SKOS	772	0	FOO
BHN	Biologie Hors Nomenclature	LIRMM/CISMeF	OWL	1534	2	FOO
ONTOTOXNUC	Ontology of nuclear toxicity	CEA/LIMICS	OWL	650	0	FMO
HRDO	Ontologie des maladies rares humaines	INSERM/LIMICS	OWL	135,939	20	FMO
MUEVO	Vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein	LIRMM	SKOS	306	18	FOO
ONL-MR-DA	Ontologie des l'acquisition de jeux de données IRM	NEUROLOG	OWL	702	244	FOO
ONL-DP	Ontologie des traitements de jeux de données	NEUROLOG	OWL	541	220	FOO
ONL-CORE-MSA	Ontologie noyau des instruments pour l'évaluation des états mentaux	NEUROLOG	OWL	329	249	FOO
Average				13,661.2	46.2	
Total				387,623	1206	

pertaining to French [22].<sup>3</sup> Moreover, only BioPortal offers an embedded semantic annotation web service. Another resource for terminologies in biomedicine is the UMLS Metathesaurus [23] which contains six French versions of standard terminologies.

The NCBO BioPortal (<http://bioportal.bioontology.org>) [10], developed at Stanford, is considered now as the reference open repository for (English) biomedical ontologies that were originally spread out over the web

and in different formats. There are 690+ public semantic resources in this collection as of early 2018. By using the portal's features, users can browse, search, visualize and comment on ontologies both interactively through a web interface, and programmatically via web services. Within BioPortal, ontologies are used to develop an annotation workflow [13] used to index several biomedical text and data resources using the knowledge formalized in ontologies, to provide semantic search features and enhance

the information retrieval experience [24]. The NCBO BioPortal functionalities have been progressively extended over the last 12 years, and the platform has adopted semantic web technologies (e.g., ontologies, mappings, metadata, notes, and projects are stored in an RDF<sup>4</sup> triple store). NCBO technology [11] is domain-independent and open source. A BioPortal virtual appliance<sup>5</sup> embedding the complete code and deployment environment is available, allowing anyone to set up a local ontology repository and customize it. The NCBO virtual appliance is quite regularly requested by organizations that need to use services like the NCBO Annotator but have to process sensitive data in house e.g., hospitals. NCBO technology has already been adopted for different ontology repositories such as the MMI Ontology Registry and Repository [25], the Earth Sciences Information Partnership earth and environmental semantic portal (see <http://commons.esipfed.org/node/1038>). We are also working on AgroPortal [26], an ontology repository for agronomy.

As for French, the need to list and integrate biomedical ontologies and terminologies has been identified since the 2000s, more particularly within the Unified Medical Language for French (UMLF) [27] and VUMeF [28] (Vocabulaire Unifié Medical Francophone) initiatives, which aimed to reproduce or get closer to the solutions of the US National Library of Medicine such as the UMLS Metathesaurus [23]. The need to support unified and interrelated terminologies was identified by the InterSTIS project (2007–2010) [29]. This need was to serve the problem of semantic annotation of data. The main results of this project in terms of multi-terminological resources were:

- The SMTS portal based inter alia on ITM technology developed by Mondeca [30]. If SMTS is no longer maintained today, ITM still exists and is deployed by the company for its customers, in the field of health or otherwise.
- The Health Multiple Terminology Portal (HMTTP) [31] developed by the CISMef group, which later became HeTOP (Health Terminology / Ontology Portal – [www.hetop.eu](http://www.hetop.eu)) [32]. HeTOP is a multi-terminological and multilingual portal that integrates more than 50 terminologies or ontologies with French content (but only offers public access to 28 of them<sup>6</sup>). HeTOP supports searching for terms, accessing their translations, to identifying the links between ontologies and especially querying the data indexed by CISMef in platforms such as Doc-CISMef [33]. The added value of the portal clearly comes from the medical expertise of its developers, who integrate ontologies methodically one by one, produce translations of the terms and index (semi-manually) the data resources of the domain.

The philosophies of HeTOP and NCBO BioPortal are different even if they occupy the same niche. HeTOP's vision, similar to that of UMLS, is to build a “metathesaurus” so that each source ontology is integrated into a specific (and proprietary) model and is manually inspected and translated. Of course, this tedious work has the added value of a great wealth and confidence in the data integrated, but comes at the cost of a complex and long human process that does not scale to the number of health or biomedical ontologies produced today (similarly, the US National Library of Medicine can hardly keep pace with the production of biomedical ontologies for integration into UMLS). In addition, this content is difficult to export from the proprietary HeTOP information system, which does not offer publicly API or standard and interoperable format for easy retrieval (although, in the context of this work, several ontologies were exported by CISMef in OWL format thanks to a wrapper developed during the SIFR project). The vision of the NCBO BioPortal is different, it consists in offering an open platform, based on semantic web standards, but without integrating ontologies one by one in a meta model. The platform supports mechanisms for producing and storing alignments and annotations but does not create new content nor curate the content produced by others. The portal is not multilingual, but it offers a variety of services to users who want to upload their ontologies themselves or just reuse some already stored in the platform. For an exhaustive comparison of HeTOP and BioPortal annotation tools, we recommend reading [34].

Within the SIFR project, we were driven by a roadmap to (i) make BioPortal more multilingual [22] and (ii) design French-tailored ontology-based services, including the SIFR Annotator. We have reused NCBO technology to build the SIFR BioPortal (<http://bioportal.lirmm.fr>) [9], an open platform to host French biomedical ontologies and terminologies only developed in French or translated from English resources and that are not well served in the English-focused NCBO BioPortal. The SIFR BioPortal currently hosts 28 French-language ontologies (+ two privates) and comes to complement the French ecosystem by offering an open, generic and semantic web compliant biomedical ontology and health terminology repository.

#### Annotation tools for French biomedical data

One of the main use cases for ontology repositories is to allow the annotation of text data with ontologies [6], so as to make the formal meaning of words or phrases explicit (structured knowledge) through the formal structure of ontologies, which has numerous applications. One such application is semantic indexing, where text is indexed on the basis of annotated ontology concepts, in such a way as

to allow information retrieval and access through high level abstract queries, or to allow for semantically enabled searching of large quantities of text [35]. For example, when querying data elements, one may want to filter search results by selecting only elements that pertain to “disorders” by performing a selection through the relevant semantic annotations with UMLS Semantic Group [36] or Semantic Types [37]. In this article, we mainly focus on annotation tools for French biomedical data.<sup>7</sup>

Ontology-based annotation services often accompany ontology repositories. For instance, BioPortal has the NCBO Annotator [12, 13], OLS had Whatizit [38] and now moved to ZOOMA, and UMLS has MetaMap [39]. Similarly, since 2004, the CISMef group has developed several French automatic indexing tools based on a bag of words algorithm and a French stemmer. We can mention: (i) F-MTI (French Multi-Terminology Indexer) now property of Vidal, a French medical technology provider [40]. (ii) the ECMT (Extracteur de Concepts Multi-Terminologique – <http://ecmt.chu-rouen.fr>) web service, the core technology of which has been transferred to the Alicante company. As a quick comparison, ECMT does not allow to choose the ontology to use in the annotation process, offers only seven terminologies, and supports semantic expansion features (mappings, ancestors, descendants) only since v3 (released after the start of SIFR project). The web service does not follow semantic web principles, does not enforce the use of URIs and the public fronting API is limited to short snippets of text. However, both F-MTI and ECMT’s use of a more advanced concept matching algorithm based on natural language processing techniques (bag of words) is an advantage compared to the SIFR Annotator.

A quantitative evaluation of annotation performance is of critical importance to enable comparison to other state-of-the-art annotation systems. In the following, we shall review existing evaluation campaigns for French biomedical Named Entity Recognition (NER)<sup>8</sup> and a brief qualitative and quantitative comparison of participating systems.

Since 2015, the main venue for the evaluation of French biomedical annotation are the CLEF eHealth information extractions tasks [16, 41, 42]. In 2015 (Task1b) and 2016 (Task2), the objective was to perform biomedical entity recognition on the French-language Quaero corpus [15], which contains two sub-corpora: *EMEA* (European Medicines Agency), composed of 12 training drug notices and four test notices; and *MEDLINE* composed of 832 citation titles for training and of 832 titles for testing. The objective of the task was two-fold: 1) to annotate the input text with concept spans and UMLS Semantic Groups (called *plain entity recognition* or *PER*); 2) annotate previously identified entities with UMLs CUIs (called *normalized entity recognition* or

*NER*). The 2016 edition repeated the same task with a different subset of training documents (the training corpus of 2016 was the test corpus of 2015) and test sets. In 2016, there was also a second annotation task, where the aim was to annotate each line of a French death certificates corpus with ICD-10 diagnostic codes (the test corpus contains 31 k certificates and 91 k lines). The 2017 edition (task 2) kept only the death certificate annotation task, although corpora were proposed in both French and English.

The participating systems included a mixture of machine learning methods and knowledge-based annotation methods. In 2015, there were two knowledge-based systems, ERASMUS [43] and SIBM (CISMef) [44]. The ERASMUS system ranked first with a F1 score of over 75%; it used machine translation (concordance across two translation systems) to translate UMLS concept labels and definitions into French before applying an existing English biomedical concept recognition tool with supervised post-processing. The CISMef system was based on their ECMT annotation web service using a dictionary composed of concept labels from French biomedical ontologies from HeTOP (55 of them at that time, extended from the seven accessible in the public ECMT web service), and obtains variable evaluation results ranging from under 1% F1 score to 22% depending on the task and parameters of the evaluation (up to 65% approximate match F1-score). The other participating systems were mostly based on *conditional random fields* or classifier ensemble systems and ranked competitively with the ERASMUS system.

In 2016, ERASMUS and SIBM (CISMef) participated again [45, 46]. SIBM (CISMef) participated with an entirely different knowledge-based annotation system. Both SIBM and ERASMUS, along with BITEM, performed concept matching from the French subset of UMLS. The other participating systems were based on supervised machine learning techniques (*support vector machines*, *linear dirichlet allocation*, *conditional random fields*) but only participated for plain entity recognition. The ERASMUS system prevailed once more using the same approach as in 2015 with F1 scores comprised between 65 and 70% on PER and 47% and 52% for NER. The SIBM system from CISMef performed much better than in 2015 with F1 scores between 42 and 52% for PER and between 27 and 38% for NER depending on the task (up to 66% approximate match F1 score).

For both 2015 and 2016, knowledge-based systems tend to perform better than supervised systems, in particular ERASMUS’s machine translation approach. Supervised systems are only competitive against plain entity recognition, they are otherwise outclassed, likely due to the relatively small amount of training data available. Systems relying only on French terminologies

(mostly every system except ERASMUS) tend to be at a disadvantage, as the coverage of corpus by French labels is low, given that the corpus was built by bilingual annotators that did not restrict themselves to French labels and used CUIs to annotate sentences independently of the existence of a label in French for those CUIs in UMLS. This limitation also concerns the SIFR Annotator which uses only French terminologies; we will discuss later how we address this bias in our evaluation.

In 2016, for the death certificate annotation task, the ERASMUS system prevailed, but this time using an information retrieval indexing approach (Solr indexing + search on lines) with over 84% F1 score. Follow, ERIC-ECSTRA (a supervised system) [47], SIBM, LIMSI (information retrieval approach, [48]) and BITEM (pattern matching between dictionary and text).

In 2017, there were a total of seven systems, including our generic SIFR Annotator; comparison results are reported in the [Results](#) section of this article. Among the seven systems, six were knowledge-based. LITL [49] used a Solr index to create a term index from the provided dictionaries and a rule-based matching criterion based on index searches. We (LIRMM) [50] used the SIFR Annotator with an additional custom terminology generated from the provided dictionaries. Mondeca [51] also used the dictionaries along with a GATE annotation workflow [52] to match codes to sentences. SIBM [53], dropping the ECMT-based system, matched terms with multiple level (word, phrase) fuzzy matching and an unsupervised candidate ranking approach (for disambiguation), similarly to WBI [54] that used a Solr index and fuzzy search to match candidates along followed by supervised candidate ranking.

Most of CLEF eHealth's French information extraction approaches were specific to the evaluation tasks. While they are interesting to push the state-of-the-art and obtain the best performance within a competitive context, their general usefulness outside of the task is limited. The custom systems implemented to best fit the tasks are not easily generalizable for use outside of the competition as independent, open and generic systems. In 2015 and 2016, only SIBM used a generic approach not specific to the benchmark. In 2017, SIBM switched to a task-specific approach and SIFR Annotator was the only open and generic approach, and which is available as an open web service independently of the competition. In this article, we report on how we exploited the task as a means of evaluating and mitigating the shortcoming of the SIFR Annotator in order to implement or identify improvements to the annotation service generalizable to any application of biomedical semantic annotation.

The CLEF eHealth 2017 Task 1 also included a reproducibility track, where participants could submit instructions to build and run their systems and evaluate the

reproducibility of each other's experiments. Four participating systems partook in this exercise (KFU, LIRMM, the unofficial LIMSI and UNIPD, another non-official participant). The evaluation consisted of allocating a maximum of 8 h per system to replicate the results and to fill in an evaluation survey by reporting difficulties and observations. Our SIFR Annotator system produced results with under 1% difference in precision, recall of F1 score compared to our official submission. While our CLEF eHealth experiments were performed in a sandboxed and controlled environment (clean instance of SIFR Annotator with only the terminologies needed for the evaluation), we decided to instruct reproducing teams how to use our online production SIFR Annotator for the reproduction to demonstrate the robustness of the platform and its ease of access/usability. The reproduction was successful and led to an accurate reproduction of the sandboxed results within less than an hour for reproducing teams.

## Implementation

### Building the SIFR BioPortal

#### Terminology/ontology acquisition

Porting an ontology-based annotation tool to another language in only half of the work. Beyond specific matching algorithms, one of the main requirements is to gather and prepare the relevant ontologies and terminologies used in the annotation process. Indeed, the ontologies offer thematic coverage, lexical richness and relevant semantics. However, ontologies and terminologies in biomedicine are spread out over the Web, or not yet publicly available; they are represented in different formats, change often and frequently overlap. In building the SIFR BioPortal and Annotator our vision was to embrace semantic web standards and promote openness and easy access. The list of ontologies and terminologies currently available in the SIFR BioPortal is available in Table 1. Hereafter, we describe each of the sources:

- Our first source of semantic resources is the UMLS Metathesaurus, which contains six French terminologies, translations of their English counterparts. For instance, the MeSH thesaurus is translated and maintained in French by INSERM (<http://mesh.inserm.fr>) and new releases are systematically integrated within the UMLS Metathesaurus. We used the NCBO-developed umls2rdf tool (<https://github.com/ncbo/umls2rdf>) to extract three of these sources in RDF format and load them in our portal.<sup>9</sup> These sources are regularly updated when they change in the UMLS.
- Our second source of French terminologies is the CISMef group, which in France is the most important actor to import and translate medical

terminologies. During the SIFR project, the group developed an OWL extractor for the HeTOP platform which can be used to produce an OWL version of any resource integrated by CISMef within HeTOP. 11 of the SIFR BioPortal terminologies have been produced with this converter and rely on CISMef for updates, URI providing and dereferencing.

- Our third source of ontologies is the NCBO BioPortal. Indeed, multilingual biomedical ontologies that contain French labels are generally uploaded to the NCBO BioPortal by their developers. We automatically pulled the ontology sources into the SIFR BioPortal and display/parse only the French content in our user interface and backend services (including the SIFR Annotator dictionary). By doing so, the NCBO BioPortal remains the main entry point for such ontologies –for English use cases– while SIFR BioPortal serves the French content of the same ontologies and links back to the mother repository. Ontology developers do not have to bother about the SIFR BioPortal as the source of information for ontology metadata and new versions remains the NCBO BioPortal.
- Finally, direct users or institutions are the last source of ontologies and terminologies in the SIFR BioPortal. The resources concerned are semantic resources developed only in French that are either not included in HeTOP or not offered by CISMef. Indeed, such use-cases are outside the scope of CISMef with their HeTOP platform and adding new ontologies to HeTOP involves a lengthy administrative process. Therefore, the SIFR BioPortal fills this need for the French biomedical ecosystem by offering an open and generic platform on which uploading a resource is quick and obvious and automatically comes to complete the SIFR Annotator dictionary. For instance, the CNRS's Scientific and Technical Information Department helps scientists in adopting semantic web standards for their standardized terminologies used for instance in literature indexing. The Loterre project ([www.loterre.fr](http://www.loterre.fr)) offers multiple health related SKOS vocabularies for which the SIFR BioPortal is another point of dissemination and automatic API access.

### **Portal content and ontology curation**

Within the SIFR BioPortal, semantic resources are organized in groups. Groups associate ontologies from the same project or organization for better identification of their provenance. For instance, we have created a group for all the ontologies of the LIMICS research group, imported from the NCBO BioPortal, or being a

translation of an English UMLS source. The SIFR BioPortal has the capability (inherited from the NCBO BioPortal) to classify concepts based on CUIs and Semantic Types from UMLS. For instance, it enables the SIFR Annotator to filter out results based on a certain Semantic Types of Semantic Groups (as described later). For the three terminologies within the UMLS group directly extracted from the UMLS Metathesaurus format (MDREFRE, MSHFRE, MTHMSTFRE) the CUI and Semantic Type information provided by the Metathesaurus were correctly available. However, for most of the six other ontologies in the UMLS group, produced by CISMef in OWL format (CIM-10, SNMIFRE, WHOART-FRE, MEDLINEPLUS, CISP-2, CIF), the relevant UMLS identifiers (CUI & TUI) were missing or improperly attached to the concepts. We therefore enriched them to reconcile their content with UMLS concepts and Semantic Type identifiers [55]. For this, we used a set of previously reconciled multilingual mappings [56] made through a combination of matching techniques to associate concept codes between French terminologies and their English counterparts in UMLS.

All in all, the SIFR BioPortal contains now 10 ontologies with UMLS interoperability among a total of 28. Since we relied on retrieving and normalizing existing mappings, we could only enrich ontologies that were in UMLS to begin with, however, we are working on integrating a generalized reconciliation feature that would automatically align terminologies submitted to SIFR BioPortal with the UMLS Metathesaurus. In addition, SIFR BioPortal includes an interlingual mapping feature that allows interlinking with equivalent ontologies in English. There are currently nine French terminologies with interportal mappings to NCBO BioPortal [56]. In a broader multilingual setting, the UMLS Metathesaurus, for some resources such as MeSH, is a de-facto multilingual pivot that allows linking annotations with concepts across languages and to generate inter-portal mappings. As with any multilingual pivot structure, care must be taken when dealing with ambiguous multilingual labels that may be an important source of noise if more than two languages are involved.

There are numerous practical and tedious technical issues with any efforts to integrate biomedical ontologies in an open ontology repository. Heterogeneous ontologies often contain many inconsistencies and “incorrect” constructs which often show up when put together in the same platform. For instance:

- Inconsistent concept hierarchy (multiple roots, no hierarchy, no root concept);
- Non-compliance with best practice standards (especially semantic web standards);
- Use of heterogeneous and non-standard properties.

Moreover, ontologies, although they may be available online, often do not define clear licensing information, which prevents their diffusion on any ontology library. Lengthy investigations to find the authors (or authority organization) of the ontologies and then to negotiate licensing terms are often required before a resource can be hosted in the SIFR BioPortal. In certain cases, the semantic resource is accessible (user interface & web services) but not downloadable.

Despite the numerous challenges facing such an endeavor, SIFR BioPortal, across all the ontologies indexed in the repository, currently represents the largest open French-language biomedical dictionary/term repository,<sup>10</sup> with over 380 K concepts and around twice that number of terms. Enabling the SIFR Annotator service to use additional ontologies is as simple as uploading them to the portal (the indexing and dictionary generation are automatic) and take only a few minutes. Table 1 summarizes some statistics about the repository's content in terms of size and general characteristics of the semantic resources.

On the subject of licencing of the resources, two of the four terminologies directly extracted from UMLS are subjected to UMLS license terms and are not directly downloadable from SIFR BioPortal. They are available for people that do have UMLS licenses, although our system doesn't directly interface with the UMLS license server.

For the other ontologies and terminologies, access rights have been discussed to allow us to make them openly available when relevant. Often, resources within SIFR are loaded by their developer directly. We encourage our contributors to unambiguously assign a specific license to their ontology or terminology (and provide the technical means to capture this information). In addition, there are some private ontologies that are not visible to the public, any user can add such ontologies for their private needs and access is granted only by the user who submitted the ontology.

It is important to note that regardless of licensing, the non-private resources can always be used for annotation i.e., their identifiers (URI, CUI) can be used to annotate text sent to the Annotator.

### SIFR Annotator Workflow & Features

The SIFR Annotator allows annotating text supplied by users with ontology concepts. It uses a dictionary composed of a flat list of terms built from the concept and synonym labels from all the ontologies and terminologies uploaded in the SIFR BioPortal. The SIFR Annotator is built on the basis of the NCBO Annotator [12, 13] which is included in the NCBO virtual appliance. We have customized the original service for French but also developed new language independent features. In the

following, we describe the complete SIFR Annotator workflow (including new and preexisting functionalities). The Annotator is meant to be accessed through a REST API but there is also a user interface that serves as a demonstrator and that allows a full parametrization (Fig. 1).

The SIFR Annotator mainly relies on Mgrep [57] as concept recognizer. Although experiments have been carried out –both by NCBO and us– to swap the underlying concept recognizer with another (MetaMap, Alvis, Mallet, UniTex), Mgrep is still the default recognizer. It uses a simple label matching approach but offers a fast and reliable (precision) matching that enables its use in real-time high load web services. Mgrep and/or the NCBO Annotator have been evaluated [58–61] on different English-language datasets and usually perform very well in terms of precision e.g., 95% in recognizing disease names [62]. A comparative evaluation of MetaMap [39] and Mgrep within NCBO Annotator was made in 2009 [12] when the NCBO Annotator was first released. There are, however, no evaluations of Mgrep on French text.

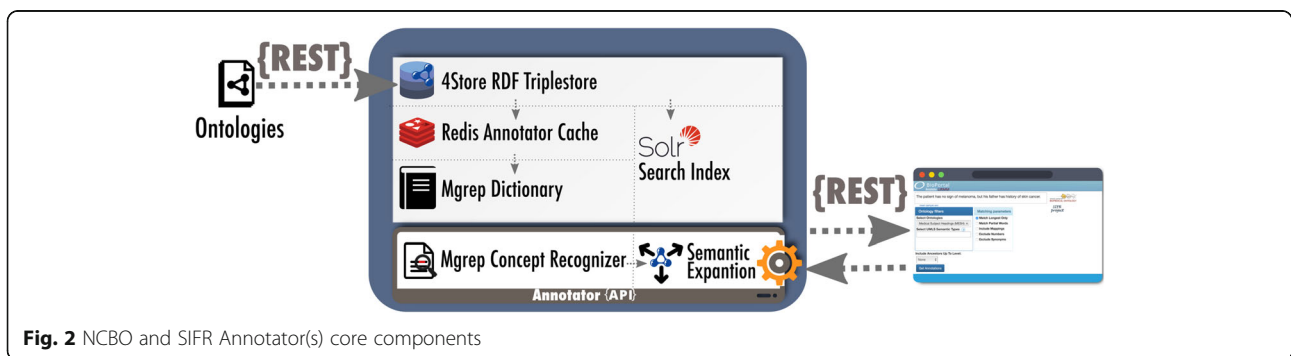
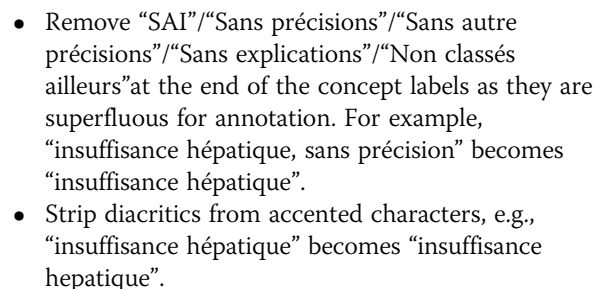
The architecture of the NCBO and SIFR Annotator(s) is described in Fig. 2. When ontologies are submitted to the corresponding repository, they are loaded in a 4Store RDF triplestore and indexed in an Apache Solr search index. Subsequently, the labels of concepts (main labels and alternative labels) are cached within a Redis table, and thereafter used to generate a dictionary for the Mgrep concept recognizer. During annotation, the concepts that have been matched to the text undergo semantic expansion (mappings and hierarchy). The process and associated features are detailed hereafter with a running example to illustrate the steps more concretely.

### Dictionary creation

The dictionary consisting of all the terms harvested from the ontologies is a central component of the concept recognizer. Mgrep works with a tab-separated dictionary file containing unique identifiers for each term as well as the term to match themselves. If terms are duplicated among multiple ontologies, they will be repeated inside the Mgrep dictionary.

When a new ontology is uploaded and parsed by the SIFR BioPortal concept labels and synonyms are indexed (using Solr) and cached (using Redis) for respectively faster retrieval and to build the dictionary. For features such as lemmatization another custom lemmatized dictionary is also produced and used depending on the annotation options selected.

For instance, the MSHFRE concept D001943<sup>11</sup> with preferred label “Tumeurs du sein” and three synonyms will correspond to the following entries in the default dictionary:



- Separate individual clauses from conjunctive sentences (split on by coordinating conjunctions), e.g., “absence congénitale de la vessie et de l’urètre” becomes “absence congénitale de la vessie” and “absence congénitale de l’urètre”.
- Normalize punctuation (replace by spaces).
- Remove parenthesized or bracketed precisions, e.g., “myopathie myotubulaire (centro-nucléaire)” becomes “myopathie myotubulaire”.

Our experiments have shown that recall increases with such heuristics, while precision decreases. Given that splitting labels increases noise, the heuristics are currently deactivated by default. For example, the dictionary entry:

77366455283 **Troubles généraux et anomalies au site d'administration**

Would be split as follows after the application of the heuristics:

77366455283 **Troubles généraux au site d'administration**

77366455283 **anomalies au site d'administration**

Possibly generating false positive annotations.

The NCBO Annotator is developed and maintained by the NCBO and does not easily support quick add-ons. To extend the original Annotator’s architecture without modifying the original application, we developed a proxy web service that can run independently and extend the service by pre-processing inputs and post-processing outputs, as we will discuss further in Section “Generalization to the any NCBO-like Annotator”. Figure 3 describes the extended SIFR Annotator workflow, where the blue frame represents the core components from Fig. 2. The main steps of the workflow are described in more detail hereafter.

### Text/query preprocessing

When a query is sent to the SIFR Annotator, it first performs some preprocessing on the parameters to implement some of the extended features e.g., lemmatizing the text. At this stage, some parameters are intercepted

and others are rewritten to be forwarded. For example, Semantic Groups are expanded into appropriate Semantic Types that are then handled by the original core Annotator components. For instance, to annotate the text “diagnostic de cancer du sein précoce” with MeSH and Meddra and with concepts belonging to the ‘disorders’ Semantic Group, one will make the following request to SIFR Annotator:

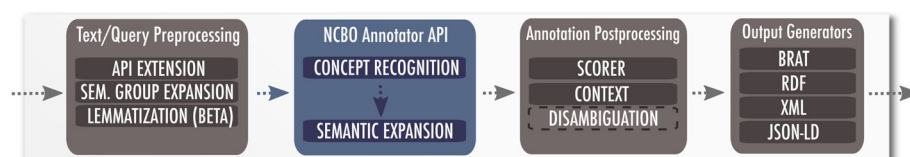
```
text = "diagnostic de cancer du sein précoce"
ontologies = "MSHFRE,MDRFRE"
semantic_groups = DISO.12
```

During this step, the latest parameter will be transformed into a list of Semantic Types (T020,T190,T049,T019,T047,T050,T033,T037,T048,T191,T046,-T184) for “disorder” that are handled by the original annotator web service (described hereafter).

### Core annotator components

At this step the original core components inherited from the NCBO technology are called:

- Concept recognition. The text is first passed to the concept recognizer, by default Mgrep, along with the previously generated dictionary. Mgrep, returns an annotation with the following information: concept identifier and the substring of the text corresponding to the matched token with its start-end offsets (from the beginning of the text in number of characters). The Annotator then retrieves the information (particularly URIs) of each annotating concept in the Solr index in order to generate a significant response to the users. Concept recognition can be parameterized with:
  - match\_longest\_only = true. Keeps the longest annotation spans, among overlapping annotations. For example, if we annotate “cancer du sein”, this parameter will discard the individual “sein” and “cancer” annotations.
  - match\_partial\_words = true. Enables matching concepts that correspond to substrings in tokens. For example, for the text “système



**Fig. 3** Proxy service architecture implementing the SIFR Annotator extended workflow. During preprocessing, parameters are handled and text can be lemmatized, before both are sent to the core annotator components. During annotation postprocessing, scoring and context detection are performed. Subsequently, the output is serialized to the requested format

cardiovasculaire”, we would match the concept “vasculaire” when this option is enabled.

Other secondary parameters are available (e.g., stop words, minimum token length, inclusion/exclusion of synonyms).<sup>13</sup>

- Annotation filtering. The SIFR Annotator can filter annotations by UMLS Semantic Types and UMLS Semantic Groups for resources with concepts enriched with such information; typically, those from the UMLS group.

○ semantic\_types = [list\_of\_TUIs],  
semantic\_groups = [list\_of\_SemGroups]<sup>14</sup>

For instance, a pharmacogenomics researcher doing a study, may restrict the annotations to the types ‘disorders’ and ‘chemicals & drugs’ to investigate the effect of adverse drug reactions.

- Semantic expansion. Direct annotations identified within the text are then expanded using the hierarchical structure of ontologies as well as mappings between them. For example: an is-a transitive closure component traverses an ontology parent-child hierarchy to create new annotations with parent concepts. For instance, if a text is annotated with a concept from HRDO, such as mélanome, this component generates a new annotation with the term Tumeur/néoplasie, because HRDO provides the knowledge that a melanoma is a kind of neoplasm/tumor. Similarly, the mapping component will create additional annotations with ontology concepts mapped to the previously matched annotating concepts. This functionality allows to “expand” the lexical coverage of an ontology by using alignments with more lexically rich ontologies. Or it enables the SIFR Annotator to use the semantics of other ontologies while returning annotations with solely the user selected target ontologies. For instance:

?text=Néoplasme malin\_&longest\_only=true  
&expand\_mappings=true  
&expand\_class\_hierarchy=true  
&class\_hierarchy\_max\_level=1

In this example, “Néoplasme malin” directly matches only in SNMIFRE, however the SNMIFRE concept maps to 7 other ontologies through mappings (CUI mappings from UMLS and user-contributed mappings). This means that if we need to use, for instance, MeSH (MSHFRE) as an annotation target, the mappings will enable us to perform concept recognition with the full richness of the labels of equivalent concepts through said mappings, while returning only annotations with MeSH concepts to the user.

The UMLS Metathesaurus, for some resources such as MeSH is a de-facto multilingual pivot that allows expanding annotations with concepts across languages.

As with any multilingual pivot structure, care must be taken when dealing with ambiguous multilingual labels that may be an important source of noise.

### Annotation Postprocessing

Annotations resulting from concept recognition and semantic expansion are post-processed –expanded, filter or enriched. Clinical context detection and scoring are two examples of annotation enrichment, while score-threshold and Semantic Group filtering are examples of filtering operations.

- Scoring. When doing ontology-based indexing, the scoring and ranking of the results become crucial to distinguish the most relevant annotations within the input text. For instance, one may assume a term repeated several times will be of higher importance. Higher scores reflect more important or relevant annotations. However, this feature is not included in the NCBO Annotator.<sup>15</sup> In the SIFR Annotator, we have implemented and evaluated a new scoring method allowing to rank the annotations and enabling to use such scores for better indexing of the annotated data. By using a natural language processing-based term extraction measure, called C-Value [63], we were able to offer three relevant scoring algorithms which use frequencies of the matches and positively discriminate multi-words term annotations. This work is reported and evaluated in Melzi et al. [63]. We also implemented a thresholding feature that allows to prune annotations based on absolute or relative score values<sup>16</sup>:

○ score = [cvalue, valueh, old] allows to select the scoring method.

○ score\_threshold = [0–9] + sets an absolute score cut-off threshold. Annotations with lower scores are discarded.

○ confidence\_threshold = [0..100] sets a relative cut-off threshold on the score density function for the distribution of annotation scores returned by the annotator.

- Clinical context detection. When annotating clinical text, the context of the annotated clinical conditions is crucial: Distinguishing between affirmed and negated conditions (e.g., “no sign of cancer”); whether a condition pertains to the patient or to others (e.g., family members); or temporality (is a condition recent or historical or hypothetical). NegEx/ConText, is one of the best performing and fastest (open-source) algorithms for clinical context detection in English medical text [64, 65]. NegEx/ConText is based on lexical cues (trigger terms) that modify the default status of medical conditions

appearing in their scope. For instance, by default the system considers a condition affirmed, and marks it as negated only if it appears under the scope of a negation trigger term. Each trigger term has a pre-defined scope either forward (e.g., “denies”) or backward (e.g., “is ruled out”), which ends by a colon or a termination term (e.g., “but”). Although an implementation of NegEx was available for French [66], we extended it to the complete ConText algorithm by methodologically translating and expanding the required trigger terms. We integrated NegEx/ConText in SIFR Annotator, which is now a unique open ontology-based annotation service that both recognize ontology concepts and contextualize them. This work is reported and evaluated in detail in Abdaoui-et-al.; however, we briefly report performance evaluation in Section “[Clinical Context Detection Evaluation](#)”. Here is an example where all three context features are enabled:

```
?text=Le patient n'a pas le cancer, mais son père a des
antécédents de mélanome
&negation=true
&experiencer=true
&temporality=true
&semantic_groups=DISO
```

#### Output generators

Finally, the workflow generates the final JSON-LD output or converts it to different formats (e.g., BRAT). NCBO Annotator supports JSON-LD and XML outputs, but while JSON-LD is a recognized format, it is not sufficient for many annotation benchmarks and tasks, especially in the semantic web and natural language communities. SIFR Annotator adds support for standard linguistic annotation formats for annotation (BRAT and RDF) and task-specific output formats (e.g., CLEF eHealth/Quaero). The new output formats allow us to produce outputs compatible with evaluation campaigns and in turn to evaluate the SIFR Annotator. Moreover, they enable interoperability with various existing annotation standards.

For instance, in order to generate the output for the Quaero evaluation, one may use:

```
?text=cancer_du_poumon
&semantic_groups=DISO
&format=quaero
```

#### Generalization to the any NCBO-like annotator

In order to generalize the features developed for French in the SIFR BioPortal to annotators in other BioPortal appliances, we have adopted a *proxy*<sup>17</sup> architecture (presented previously), that allows the implementation of features on top of the original REST API, thereby

extending it through an intermediary web-service. The advantage of such an architecture is that a proxy instance can be seamlessly pointed to any running BioPortal instance. We have set-up this technology to port new features to the original BioPortal service and offer an NCBO Annotator+ [14] and to the AgroPortal [26]. Hereafter is an example of an annotation request on an English sentence sent to the NCBO Annotator+ using the extended features enabled by the proxy architecture:

```
http://services.biportal.lirmm.fr/ncbo_annotatorplus/
?text=The patient has no sign of melanoma but his
father had history of skin cancer.
&ontologies=MESH
&longest_only=true
&negation=true
&experiencer=true
&temporality=true
&score=cvalue
&semantic_groups=DISO
```

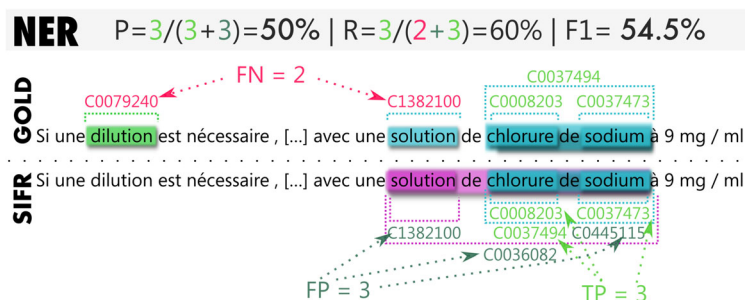
#### Results and evaluation

In this section we shall present and analyze our evaluation of SIFR Annotator on three tasks. The first is biomedical named entity recognition and normalization (using the Quaero corpus from CLEF eHealth 2015), the second is ICD-10 diagnostic coding of death certificates (using the CépiDC corpus from CLEF eHealth 2017) and the third is a summary of the evaluation for the context detection features of SIFR Annotator (negation, temporality, experiencer). We evaluate each feature independently: the purpose of the two first evaluations is to gauge how the SIFR Annotator performs for concept recognition; while the third evaluation assess the accuracy of our French adaptation of ConText.

#### Annotation of MEDLINE titles and EMEA notices with UMLS concepts and semantic groups

As discussed in Section “[Annotation Tools for French Biomedical Data](#)”, the only French biomedical named entity recognition openly available corpora come from the CLEF eHealth information extraction tasks. The CLEF eHealth NER tasks from 2015 and 2016 tasks are based on subsets of the Quaero corpus [15]. We evaluate the ability of SIFR Annotator to identify entities and annotate them with UMLS Semantic Groups (Plain Entity Recognition or PER evaluation) and CUIs (Normalized Entity Recognition or NER evaluation) on the subset of the Quaero corpus comparable to the results of CLEF eHealth 2015 Task 1 (training corpus in Quaero).

Figure 4 illustrates the objective of the PER evaluation task and Fig. 5 that of the NER evaluation tasks (and their score calculation). The example is an actual sample from the results produced by SIFR Annotator and



**Fig. 4** Illustration of the PER annotation task and the score computation. Entities in PER are identified by their character offsets (begin and end from the start of the text) and by their UMLS Semantic Group

illustrates some of the limitations of the evaluation. In Plain Entity Recognition, some entities are not contained in the semantic resources of the SIFR BioPortal (dilution), some entities are recognized properly, but are categorized in a different Semantic Group due to ambiguity (for “solution”, both classifications (CHEM, OBJC) are often correct but the gold standard keeps only one), some entities are recognized by SIFR Annotator but are not contained in the gold standard (although they could or should like, “solution de chlorure de sodium” in the example, which is the longest possible match).

For the normalized entity annotation with CUIs, if the entity and its Semantic Group are wrong, a false positive is generated, even if the CUI is actually correct (e.g., “solution” C1282100). Which is likely to lead to overall reductions in precision compared with the PER evaluation.

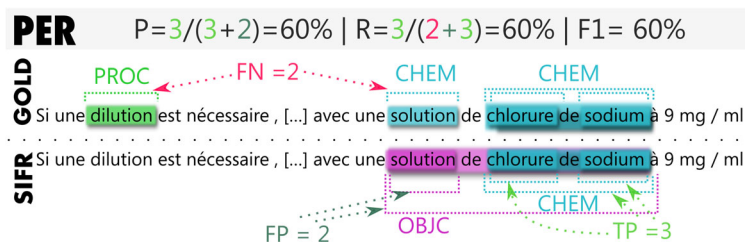
Additionally, the SIFR Annotator may identify several valid CUIs, although the gold standard always expects a single one (non-exhaustive annotation). For example, the software annotates “chlorure de sodium” with C0037494 and C0445115. The former is what the gold standard expects, the CUI for the chemical solution, while the latter is the CUI for the pharmaceutical preparation (normal saline), which is a correct answer that counts as a false positive.

### Construction Biases & Production of the adapted Quaero Corpus

As previously mentioned, one important bias of Quaero, is that it uses UMLS meta-concepts identified by CUIs irrespective of whether or not a French label exists in the UMLS. We have seen that this had a strong influence on the results and constitutes an advantage for systems using machine translation (ERASMUS in particular).

By reconciling UMLS concepts and Semantic Type information inside the French terminologies offered by CISMef [55], we have mitigated this issue by greatly extending the coverage of the “French UMLS”; but the problem still remains.

Because the SIFR Annotator does not use machine translation, in order to obtain a fairer and more significant evaluation, we produced a pruned version of the Quaero gold-standard by filtering out all manual annotations made with CUIs for which there are no French labels in any of the 10 ontologies of the UMLS group in SIFR BioPortal. If all CUIs for a text span are removed, then the whole annotation is removed from the corpus. Table 2 presents the statistics of the original corpus compared to that of the adapted corpus. The script used to generate the subset of the corpus along with the list of CUIs used for the filtering will be made available on github.



**Fig. 5** Illustration of the NER annotation task and the score computation. In NER, we annotate entities found in PER with one or more CUIs

**Table 2** Number of CUIs expected between the gold standard annotations in the Quaero corpus and the adapted Quaero corpus

	Quaero		Adapted Quaero	
	EMEA Dev	MEDLINE Dev	EMEA Dev	MEDLINE Dev
CUIs (uniq.)	2261 (526)	2978 (1843)	1733 (425)	2465 (1477)
	EMEA Test	MEDLINE Test	EMEA Test	MEDLINE Test
CUIs (uniq.)	2203 (474)	3093 (1907)	1710 (388)	2606 (1544)
	EMEA Train	MEDLINE Train	EMEA Train	MEDLINE Train
CUIs (uniq.)	2695 (651)	2995 (1861)	2279 (541)	2491 (1488)

For the uniq. Statistic, only the first occurrence of a CUI is counted. In MEDLINE, each document is a title of 10–15 word forms on average, while EMEA documents are full notices with several hundred word forms each

### Experimental Protocol & Parameters Tuning

We now present the experimental protocol used for the evaluation of SIFR Annotator on the EMEA and MEDLINE sub corpora of Quaero (original and adapted) on both the Plain Entity Recognition [PER] and the Normalized Entity Recognition [NER] annotations tasks, along with a description of the parameter tuning process. We present the baseline annotation setting along with two post-annotation disambiguation heuristics.

In the baseline setting, we used the “quaero” output format of the SIFR Annotator which produces a BRAT output format compliant with the evaluation scripts for the task. The parameters of SIFR Annotator used for the baseline annotation were the following:

- match\_longest\_only = false as the gold dataset annotates both multi-word terms and their constituents.
- match\_patial\_words = false as there are no such annotations possible in this task.
- negation = false, temporality = false, experiencer = false as the tasks does not require contextual annotations.
- semantic\_groups = {}, semantic\_types = {}, as all Semantic Types are found in the gold annotations.
- We used all the 10 terminologies in the UMLS group within the SIFR BioPortal.

Depending on the type of text we are annotating, using all 10 UMLS terminologies may not be ideal as some may not correspond to the data and thus create annotation noise (false positives). In the present evaluation, the EMEA and MEDLINE sub corpora contain very different types of text (citation titles vs. drug notices), which justifies the need of finding the best subset of ontologies. To that end, we performed a grid search over all combinations of terminologies (we evaluated a total of  $\sum_{k=1}^{10} \binom{10}{k} = 1023$  combinations) by scoring the resulting annotations on each of the dev sub-corpora.<sup>18</sup>

Once the optimal combination is found for both MEDLINE and EMEA, we evaluated the performance of

the baseline annotation and of two post-annotation disambiguation heuristics on the test and training corpora for both the original Quaero corpus and the adapted Quaero corpus. We report on the actual values of the optimal target ontology lists prior to the evaluation results in the next section.

Because the Quaero corpus was constructed considering the UMLS Metathesaurus as a unique semantic resource and given that the nature of the SIFR Annotator is to consider UMLS as a group of 10 terminologies, we can already predict a shortcoming of SIFR Annotator with regard to task performance. In UMLS, one concept from a particular source, may be tagged with more than one CUI and consequently to more than one Semantic Group, inevitably creating ambiguities when multiple sources are used together. This is a well-known constraint/limitation when using UMLS [23]. Most of the 10 UMLS source terminologies in SIFR BioPortal have concepts with multiple Semantic Groups and/or CUIs, whereas Quaero gold standard annotations used only one, which will predictably lead to an ambiguity problem. Additionally, given that an entity and its Semantic Group must be correct in PER before the CUI annotation in NER is counted as correct (as shown at the beginning of Section “Error Analysis”), we expect a decrease in precision, while recall should stay the same between PER and NER, similarly to all systems participating in CLEF eHealth 2015 Task 1 [Hypothesis 1].

Additionally, we can expect SIFR Annotator to perform better in terms of recall on the adapted Quaero corpus and thus a higher overall F1 score [Hypothesis 2].

### Disambiguation heuristics

One way of mitigating the effect of the hypothesized compound effect of the ambiguity is to attempt to find a heuristic that avoids the ambiguity altogether at the potential expense of either precision or recall. We evaluated two heuristics:

- [DAA – Discard Ambiguous Annotations] If we favor precision over recall, then a strategy is to remove ambiguity altogether by discarding any annotations belonging to several Semantic Groups. This strategy will likely reduce recall as some of the discarded annotations could be true positives [Hypothesis 3].
- [DBP – Distribution Based Prioritization] If we favor recall over precision, then another strategy is to disambiguate the Semantic Groups by keeping the most likely group as estimated with regards to the development corpus. In other words, we learn a frequency-based ranking of Semantic Groups and always keep only the best ranking Semantic Group. Statistically, in many cases as far as word sense

disambiguation is concerned, the most frequent sense of a word is correct a majority of times depending on the degree of ambiguity. The most frequent sense heuristic is typically used as a strong baseline in word sense disambiguation studies [67, 68]. Although in the case of Semantic Groups, the frequencies are not contextualized and thus will not impact as well as in a typical word sense disambiguation task, we expect some improvement in precision for PER and NER [Hypothesis 4].

## Results

First, optimal parameters for both EMEA and MEDLINE are:

- Set of ontologies. This parameter is independent from the DAA and DBP heuristics.
- Ranking of Semantic Groups based on their frequency distribution in the development corpus (DBP heuristic).

Both parameters remain the same for the baseline on the full and adapted corpora.

Table 3 summarizes the optimal values of the parameters estimated on the Quaero development corpus.

We then ran the annotation for PER and NER, on EMEA and MEDLINE on the full and adapted Quaero training corpora with the baseline setting and with the two heuristics. Table 4 summarizes the results in terms of Precision (P), Recall (R), F1 measure, and provides the average and median values for other CLEF eHealth 2015 Task 1 participants to which we may compare our results.

**PER evaluation** The baseline approach for PER obtains slightly better results ( $F1 = 57.2\%$ ) on the EMEA corpus compared to MEDLINE ( $F1 = 52.9\%$ ) which can probably be explained by the fact that each title in MEDLINE pertains to a broad range of medical topics whereas EMEA is only about medication notices. The former necessarily offers a more diverse distribution of Semantic Groups, which is more difficult to identify.

The DAA heuristic does not consistently lead to better results than the baseline. For EMEA, the performance is lower than the baseline ( $-5.7\%P$ ,  $-0.1\%R$ ,  $-2.4\%F1$ ), while for MEDLINE, it significantly improves the

baseline results ( $+9.1\%P$ ,  $+0\%R$ ,  $+4\%F1$ ). This seems to invalidate Hypothesis 3, as recall is unaffected. In EMEA, where there is less ambiguity, the heuristics tend to delete annotations where there was at least one correct Semantic Group annotation, which leads to lower precision, while for MEDLINE, it is more likely to delete annotations where none of the Semantic Groups are correct. With the DBP heuristic, there is a consistent increase in both P and R across EMEA ( $+6.8\%P$ ,  $+4.5\%R$ ,  $+5.4\%F1$ ) and MEDLINE ( $+7.2\%P$ ,  $+5\%R$ ,  $+6\%F1$ ), which validated Hypothesis 4, although there is also a reliable increase in recall.

Compared to CLEF eHealth 2015 Task 1 participants, on the EMEA sub-task, our system, in its best configuration, would rank 4th with regard to participating systems ( $-4.3\%$  compared to the system ranked right before and  $+8.5\%$  ahead of the following system) behind ERASMUS, IHS-RD and Watchdogs. Among those systems, the two with which we are methodologically comparable, ERASMUS and Watchdogs, both used some kind of machine translation approaches. IHS-RD (as well as the system right after ours) used supervised machine learning.

On the MEDLINE sub-task, SIFR Annotator would rank 2nd with regard to participating systems, only behind ERASMUS ( $+7.3\%$ ), but before IHS-RD ( $-6.5\%$ ) and Watchdogs. This can be explained by the fact that MEDLINE has a set of more diverse expected Semantic Group annotations, while EMEA mostly contains CHEM and DISO, which means the entropy of the Semantic Group distribution is higher, which makes it more difficult to use a supervised machine learning. ERASMUS and SIFR Annotator being knowledge-based, they suffer much less from the increased entropy of the expected Semantic Group distribution. The advance of ERASMUS can mainly be explained by a richer dictionary enabled by the translation approach.

**NER evaluation** As expected, due to the added difficulty of the NER task compared to PER, annotation performance is significantly lower. The drop (between  $-16.9\%$  F1 and  $23.2\%$  F1) is similar on average for all participating systems, which validates Hypothesis 1.

The relative effect of DAA and DBP is the same in PER and in NER, meaning that the ranking between the baseline and the two heuristics remains the same in NER than it was in PER.

**Table 3** Estimated optimal parameters

	EMEA	MEDLINE
Optimal set of ontologies	MSHFRE, CIM-10, MDRFRE, SNMIFRE, CISP-2, CIF, ATCFRE	MSHFRE, MDRFRE, SNMIFRE, MEDLINE+, CIF, CISP-2, ATCFRE
Frequency ranking for Semantic Groups for DBP heuristic	CHEM, DISO, LIVB, PROC, ANAT, PHYS, OBJC, GEOG, DEVI, PHEN	DISO, PROC, ANAT, CHEM, LIVB, PHYS, DEVI, PHEN, GEOG, OBJC

**Table 4** Results on the Quaero Training for PER and NER

	Plain Entity Recognition [PER]						Normalized Entity Recognition [NER]					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	EMEA			EMEA adapted			EMEA			EMEA adapted		
BSL	64.0	51.7	57.2	63.1	59.3	61.2	49.8	30.9	37.8	48.6	35.1	40.8
DAA	58.3	51.6	54.8	57.5	59.3	58.4	45.0	30.7	36.2	44.0	34.8	38.8
DBP	<b>70.8</b>	<b>56.2</b>	<b>62.6</b>	<b>69.2</b>	<b>64.0</b>	<b>66.7</b>	<b>54.21</b>	<b>31.0</b>	<b>39.4</b>	<b>54.1</b>	<b>35.36</b>	<b>42.8</b>
Avg.	58.7	47.3	51.1	Not Available			33.3	46.0	34.7	Not Available		
Med.	73.1	55.9	61.3				19.1	56.5	25.2			
	MEDLINE			MEDLINE adapted			MEDLINE			MEDLINE adapted		
BSL	57.5	49.0	52.9	55.2	55.8	55.5	44.0	<b>30.5</b>	36.0	43.8	<b>35.5</b>	39.2
DAA	<b>67.9</b>	49.0	56.9	<b>62.2</b>	55.8	60.2	<b>52.9</b>	<b>30.5</b>	<b>38.7</b>	<b>52.7</b>	<b>35.5</b>	<b>42.4</b>
DBP	64.7	<b>54.0</b>	<b>58.9</b>	62.0	61.1	<b>61.5</b>	49.5	30.4	37.6	49.25	35.4	41.2
Avg.	53.3	39.6	44.0	Not Available			32.1	46.1	34.0	Not Available		
Med.	64.9	40.0	48.7				29.5	59.0	22.8			

Evaluation on both the EMEA and MEDLINE sub corpora for the original Quaero corpus and our adapted Quaero corpus. For the original corpora, we report on the average and median results of the systems participating in CLEF eHealth 2015 Task 1. Values in bold correspond to the best results in each category

For EMEA, the DBP heuristic performs best (39.4% F1), while for MEDLINE, DAA performs best (42.4% F1) due to a reduction in precision. This effect is understandable as the heuristics affect only the Semantic Group annotations and do not influence the FP and FN ratio in NER.

With regard to the ranking in CLEF eHealth 2015 Task 1, fewer systems participated. Without explanation, the IHS-RD system that outperformed us on EMEA in PER, completely fails to annotate with CUIs with a F1 score of less than 1%. We rank second after the ERASMUS system (– 25%) by far, however, SIFR Annotator is also much better than the other systems. Only HIT-W1 gets a F1 score above 1%, but SIFR Annotator is significantly ahead with + 17.6% (Supervised CRF with an UMLS sense inventory). The failure of supervised systems that did not use UMLS as a sense inventory is normal, given the small amount of training data compared to the millions of possible CUI annotation from UMLS and the label ambiguity. ERASMUS and SIFR Annotator do not suffer from this drawback. Despite the translation aspect, the better performance of ERASMUS is due to their superior coverage in PER but also because they annotate with UMLS CUIs directly as a target, while SIFR Annotator annotates with source concepts that are more ambiguous with regard to CUIs (we annotate many CUIs, while ERASMUS annotates only one as the task expects).

**Evaluation with the adapted Quaero corpus** The overall effect of the adapted Quaero corpus on the results of the PER task is to slightly lower precision and significantly increased recall, which increases the F1 score, on average by + 3.9% on EMEA and by + 2.8% on MEDLINE. The overall effect on the NER

task is similar but with a lower magnitude of change. The relative effects of the heuristics remain unchanged for both PER and NER. The adaptation of the corpus mostly has the expected effect of increasing the recall and thus the F1 score by a few points (Hypothesis 2). The decrease in precision indicates that on average the entities kept in the adapted corpus are more ambiguous in terms of CUIs compared to the full corpus. If we could evaluate all participating systems on the adapted corpus, we would expect that it does not affect the performance of translation-based systems, while there would be a consistent increase in the recall of systems that do not rely on translation approaches. This would likely bridge much of the gap with ERASMUS, while likely remaining second.

#### Annotation of death certificates with ICD-10 codes

The objective of CLEF eHealth 2017 Task 2 [69] is to annotate death certificates with ICD-10 codes both in French and in American English. We chose to participate in the task in order to evaluate the performance of SIFR Annotator for French and the NCBO Annotator for English. Here, we only present the results for the French corpus and point to the system paper [50] for additional details. Let us first describe the task and the French corpus, followed by a presentation of the additional semantic sources used (SKOS dictionary) and of the algorithm that maps concept ICD-10 concept URLs to ICD codes.

#### Task and Corpus description

A corpus of French death certificates from CépiDC was provided: a training corpus of 65,844 documents and

195,204 lines,<sup>19</sup> a development corpus of 27,851 document and 80,900 lines and a test corpus of 31,683 documents and 91,954 lines. The corpora are digitized versions of actual death certificates filled in by clinicians. Although the punctuation is not always correct or present, in the corpus, each document is already segmented in lines (as per the standard international death certificate model) which for the most part only contain single sentences.

The French corpus was provided in both an aligned and a raw format. We only report on the performance for the aligned corpus as our approach leads to similar results for both. The raw format provides two files, a CausesBrutes file and an Ident file. The former contains semicolon separated values for the Document identifier (DocID), the year the certificated was coded (YearCoded), the line identifier (LineID), the raw text as it appears in the certificate (RawText), an interval type during which the condition occurred (IntType - seconds, minutes, hours, weeks, years) and an interval value (IntValue). The Ident file contains a document identifier, the year the certificate was coded, the gender of the person, the code for the primary cause of death, the age and the location of death. Here is an example:

```
DocID ; YearCoded; LineID; RawText; IntType;
IntValue
161477; 2014 ; 1 ; INSUFFISANCE RESPIRATOIRE
AIGUE; 3; 5
161477; 2014 ; 2 ; PNEUMOPATHIE D
INHALATION; 3; 5
161477; 2014 ; 5 ; PSYCHOSE
CHRONIQUE;NULL;NULL
```

```
DocID; YearCoded; Gender; Age; LocationOfDeath
93715; 2014 ; 2 ; 80; 2
```

The performance on the task was reported as Precision, Recall and F1 score for the whole corpus and for the sub-corpus of deaths from external causes (a subset of ICD-10 codes), which are much harder to determine automatically. The baseline system produced by the organizers used conditional code frequencies estimated from the training data to select the most likely code for a death certificate line.

### Dictionary construction

SIFR BioPortal already contained the French ICD-10<sup>20</sup> (CIM-10) reference terminology. This OWL version was originally produced by the CISMef team from an automatic export from the HeTOP server [32]. However, the purpose of ICD-10 is to serve as a general-purpose

reference to code medical acts, and not to be directly used for text annotation and, especially not in a particular clinical task such as death certificate coding. Indeed, from our experiments, using the original CIM-10 alone for annotation leads to a F1 score below 10%.

For the French corpus, a set of dictionaries was provided by C  piDC that give a standardized description text of each of the codes that appear in the corpora. Additionally, the data from the aligned training and development corpora could also be used to enrich the lexical terms of ICD-10. In order to use these dictionaries within the SIFR Annotator, we had to encode them using a format accepted by SIFR BioPortal, which includes RDFS, OWL, SKOS, OBO or RRF (UMLS format). In this case, the ideal choice in terms of standardization, potential reusability and simplicity was to use SKOS (Simple Knowledge Organization System) a W3C Recommendation specialized for vocabularies and thesaurus. Thus, we produced a SKOS dictionary called CIM-10 DC based on the French dictionaries and aligned corpus.<sup>21</sup>

We set out in this construction process by first defining the appropriate schema to represent the SKOS dictionaries. We chose to use the same URIs as concepts identifiers for the skos:Concept than for the corresponding owl:Class in the available CIM-10 terminology, which allows our dictionaries to be fully aligned with the original terminologies they enrich (from the perspective of ontology alignment). Each of the CIM-10 codes was represented by a skos:Concept. The URIs are composed of a base URI and a class identifier that represents the CIM-10 codes, in the following format: “[A-Z][0–9][0–9]\.?[0–9]?” (e.g., G12.1 or A10).<sup>22</sup>

We first built a code index, that associated to each code to the list of labels retrieved from the DiagnosisText field in the dictionary; and then add text from the RawText and StandardText fields from the corpus (associated to codes through the ICD-10 field in the corpus file). For each code concept, the C  piDC dictionaries contained multiple labels. In order to follow SKOS specification, we had to select a preferred name automatically (skos:prefLabel) and assign the other labels as alternative labels (skos:altLabel), which has no consequence for annotation. The selection heuristic took the shortest label that does not contain three or more consecutive capital letter (likely an acronym).

An important issue when building the SKOS dictionaries was to assign ambiguous labels (i.e., identical labels which correspond to different codes). Indeed, those labels create ambiguity in the annotations and leads to better recall at the price of a low precision. For example, the label “choc septique” was present as preferred label or synonyms for 58 different codes. Our “ontological” approach posits that the same label should not be

assigned to the same label, and yet ICD codes are not ontology concepts, but diagnostic codes, which shows a limit of semantic annotation approaches for such tasks, as opposed to machine learning systems that do not suffer from the same drawback.

We had to implement a selection heuristic to determine the most suitable code to which the label should be bound. Taking inspiration from the idea of the most frequent sense baseline often used in Word Sense Disambiguation tasks, we adopted a heuristic that assigns ambiguous labels to the most frequent code only (just like in the first evaluation). We use the training corpus to estimate the frequencies of use of the codes (gold standard annotations) so that when a label can belong to several codes, we can sort the codes by frequency and chose the most frequent code (MFC).

#### Mapping algorithm between concept URIs and ICD-10 codes

Given that we used the SIFR Annotator, besides manually curating the created SKOS dictionaries, the final step to obtaining a working system for the task was to write a complete workflow to<sup>23</sup>:

1. Read the corpus in the raw or aligned formats;
2. Send the text to the SIFR Annotator REST API with the right ontologies and annotation parameters and retrieve the annotations produced;
3. Apply post-annotation heuristics to reduce ambiguity;
4. Produce the output in the right raw or aligned format.

We have used only the “RawText” information of both the aligned and raw datasets. We did not use any other information/features such as age or gender contained in the files. The evaluation run performed the annotation with the longest\_only parameter activated on a local instance of the SIFR Annotator with CIM-10 and the SKOS dictionary we produced as target ontologies. We implemented two post-annotation heuristics:

- *Most Frequent Code.* If a particular line was annotated with several codes, we only keep the most frequent code based on the code distribution of the training corpus.
- *Code Frequency Cutoff.* We calculate a normalized probability distribution of the codes that annotate a particular line and only keep the codes below a cumulative probability threshold.

However, both heuristics led to a stark reduction in recall without leading to a satisfactory increase in

precision to compensate and thus ended up lowering the overall F1 scores, which is why we did not activate them for our participation in the task.

#### Results

13 runs have been submitted by 9 teams to the French raw evaluation. Seven runs have been submitted by five teams to the French aligned evaluation. Table 5 presents the results obtained by our SIFR Annotator against the average and median results of the runs submitted to the evaluation task.

The SIFR Annotator results are exactly the median value of all the results with the raw dataset, but slightly under the median value for the aligned datasets (all causes). Indeed, teams that have used other information from the aligned dataset probably performed better than the SIFR Annotator here. Regarding the external causes, we obtain better precision and F1 than the average and median results submitted to the challenge.

The other systems that participated on the French Raw task can be divided in three categories: supervised machine learning (TUC, LIMSI), information retrieval models (IMS-UNIPD, LITL) and annotation approaches (SIFR Annotator, SIBM, Mondeca). The official results and ranking only include SIBM, LITL, SIFR Annotator and TUC (with faulty submitted results). The unofficial systems include LIMSI, UNIPD, TUC (corrected) and Mondeca. The SIFR Annotator was ranked third on all causes and second on external causes behind the SIBM system. The SIBM system is significantly ahead (> 20 + %) as it is the only system to perform code disambiguation. The difference with the second system (LITL) and ours is only of + 0.1%, hardly a significant difference. Had LIMSI run officially with their supervised system, they would have been first (82.5% F1), followed by SIBM and then the corrected TUC system (between 66.6 and 66.7% F1) and UNIPD (between 44.1 and 53.7%).

The performance of SIFR Annotator is somewhat lower than for a typical entity recognition task, because of the significant ambiguity (the same label can correspond to several different classes (here ICD-10 codes) found in the dictionaries provided with the task and in turn in our SKOS dictionary. This highlights that such a focused and specific text mining task is most likely

**Table 5** Results for ICD-10 coding of death certificates for the French Raw Evaluation

	All Causes			External Causes		
	P	R	F1	P	R	F1
SIFR	54.1	48.0	50.9	44.3	36.7	40.1
Avg.	47.5	35.8	40.6	36.7	24.7	29.2
Med.	54.1	41.4	50.8	44.3	28.3	37.6

We present P, R, F1 on all causes (all ICD-10 codes) and on external causes

better suited for machine learning approaches. However, despite of their limitations, the NCBO and SIFR Annotators obtained median results, respectively on French and English, when compared to the performance of all the participating systems. Therefore, considering the other discussed advantages, we believe they are two services that can help in a wide class of text mining or annotation problems, but of course not for all.

### Clinical context detection evaluation

As described among the features of the SIFR Annotator, there is a module for contextualizing annotations (Negation, Experiencer, Temporality) based on the ConText algorithm [65]. We adapted the algorithm to French and enriched existing translation efforts. We evaluated the French ConText on a sub-corpus of death certificates from the CLEF eHealth Task 1 corpus (6 sentences for experiencer, 150 for temporality, 1030 for negation) and on a clinical corpus from the European Hospital Georges Pompidou (630 lines for experiencer, 475 lines for temporality, and 400 lines for negation). French ConText implementation & evaluation are described in another communication; hereafter, we briefly summarize the main results.<sup>24</sup>

We reported an evaluation of the SIFR Annotator with F1 scores between 83.7 & 86.3% for negated concepts (better by more than 5% of previously reported results adapting NegEx to French), F1 88.9% and 91.7% for the detection of historical entities and between 79.2 and 90.9% for concepts pertaining to an experiencer other than the patient. The results are on-par with other state-of-the-art approaches (NegEx for negation, machine learning, etc.), independently from the concept recognition performance. Please consult the full evaluation in the article for more details.

### Discussion

In this section we discuss the results of the three evaluations and explain some of the shortcomings of SIFR Annotator by reviewing typical errors made in the annotation process. Some of the limitations are task-specific, while others are more general. We shall then draw some perspectives for future improvements.

#### Error analysis

In order to further improve our open web-service, we performed a detailed error analysis on the results of the two evaluation tasks from CLEF eHealth so as to be able to identify future direction for improvement. We reviewed and categorized the main errors in terms of False Positives and False negatives and give concrete examples from both tasks.

#### PER annotation errors

We extracted a list of 50 random errors from the outputs on the full Quaero corpus and looked at their causes in detail (Table 6).

Among the false positives, one of the most frequent cause of errors is the production of annotations that were not in the gold standard. Given that the creation of the gold standard is subjective in terms of the entities chosen to be annotated by the experts [15],<sup>25</sup> such errors are caused because of the exhaustive automatic annotation performed, which is a positive characteristic for any annotation system. Without medical expertise, by looking at a subset of these annotations, we could obviously conclude that many of them were not actual errors but indeed missing annotations in the corpus. Such omissions constitute a bias playing against knowledge-based approaches, when the set of ontologies used to compile the dictionary is richer than what human annotators considered when building the gold standard. Conversely, machine learning approaches, trained directly on a subset of the annotated corpus will not encounter this problem, but on the other hand will not have the capability of generalizing on unseen text.

The other frequent false positive error, is when SIFR Annotator only annotates a concept partially i.e., annotates the individual words with separate concepts, but not the whole expected concept. The gold standard always annotates both multi-word terms and the individual constituents. The SIFR Annotator almost always get the individual words right but not the multi-word terms.

In the example given in Table 5, the label “signes du système nerveux central” (or a simplified/tokenized version of it) does not exist in the French UMLS terminologies. The corresponding preferred label of actual corresponding concept (matching Semantic Group and CUI) is: “signes et symptômes divers du système nerveux central” which means that human expertise was required to infer that the text corresponds to a broader concept, which is very hard to reproduce for the SIFR Annotator.

Such errors could be remedied by enriching the original terminologies and ontologies (or the compiled dictionary) with more alternative labels. As previously mentioned in Section “[Terminology/Ontology Acquisition](#)”, we are already working on this but have observed mitigated results where the gain in recall does not match the loose in precision for the moment.

The third most common cause of false positives is an incorrect Semantic Group annotation.<sup>26</sup> For example, in some instances, we annotated with DISO (Disease), when it should be ANAT (Anatomy). Despite fixing some incoherent Semantic Type assignments in the source terminologies in the UMLS, the inevitable solution is to equip the SIFR Annotator with a multi-level (class, UMLS concept, type, group) disambiguation

**Table 6** PER annotation error analysis

	Description	Example	% in EMEA (14 FP & 36 FN)	% in MEDLINE (15FP & 35 FN)
FP	Annotation with a concept that was not covered in the gold standard	"évaluant la douleur"/Proc. (i.e., "pain evaluation") matched but not in gold standard.	<b>10</b>	10
	Partial annotation on some but not all of the expected tokens	"système nerveux central" recognized instead of "signes du système nerveux central" (spelling)	<b>10</b>	<b>12</b>
	Incorrect Semantic Group annotation	"rein" (kidney) annotated with DISO. instead of ANAT. Generates both an FP and an FN.	8	8
	Concept missing from the French ontologies in the portal	Expected annotation: "canaux" (canals), but the SIFR Annotator dictionary only contains "canal, sai" (canal unspecified), which cannot match	<b>34</b>	12
FN	Morphosyntactic variation	Expected annotation "sériques" (an adjectivation of sérum) as ANAT, whereas the ontology label is "sérum" (the noun).	18	<b>26</b>
	Formulation different from concept labels (synonym, paraphrase)	Expected annotation "flacon" (vial), while the ontology concept label read "bouteille" (bottle).	14	22
	Incorrect Semantic Group	"rein" (kidney) annotated with DISO instead of ANAT. Generates both an FP and an FN.	6	10
	Unrecognized acronym or medical abbreviation	The gold standard expects "SNM" to be annotated with DISO, while the ontologies only contain "syndrome malin des neuroleptiques".	2	0

Performed on 50 uniformly sampled errors on EMEA and MEDLINE obtained with the baseline method. The two most common causes are highlighted in bold

module. More generally, beyond ambiguity related to UMLS, the SIFR Annotator obviously suffers from ambiguity between the general usage of a word and its medical usage (e.g., cold).

Among false negatives, one of the most common causes of error is morphosyntactic variation (18%) or a different formulation of the labels compared to the text (14%), meaning variations of the word due to differing grammatical roles (plurals, conjugations, etc.) or a different formulation for complex concept labels. This limitation is inherent to the concept recognizer, Mgrep, that does not deal with such variations (see "canaux" example in Table 5). We are exploring two possible solutions to the problem:

- We have developed a beta lemmatization feature in the SIFR Annotator that is not yet properly evaluated. However preliminary tests indicate that it would fix morphosyntactic recognition errors significantly.
- We are developing an alternate concept recognizer robust to morphosyntactic variations and to reformulation of complex expression (based on stem indexing of the words of ontology labels and word-embedding matching), although the operational integration is not mature enough to permit a production-level evaluation like we have gone here.

A common error producing false negatives (34%) is the absence of a concept from the ontologies (with the adequate Semantic Type and CUI), which is mitigated to some extent with the adapted Quaero corpus as we remove CUIs that do not exist in French sources. In such

cases, knowledge-based approaches are indeed intrinsically limited by their ability to recognize only entities that have been captured into knowledge inside ontologies or terminologies first.

Among the less-frequent causes of false negatives, we have ambiguous Semantic Group annotations that are the main cause of incorrect Semantic Groups in false positives already covered above. We thus come back to the same idea of a multi-level disambiguation approach as the best potential mitigation.

#### NER errors

Any of the PER errors above also lead to errors in the NER task as per the construction of the task itself along with additional errors caused by the finer grain annotation:

- (E1) The expected CUIs are present in the SIFR Annotator results, but there are additional CUI annotations, which generates TPs for the expected CUIs and FPs for the others.
- (E2) None of the CUI annotations match the expected CUIs, which leads to TNs being generated for the expected CUIs and FPs for the generated CUI annotations.

At least one CUI was found for all entities identified in PER. In EMEA, E1 corresponds to 40% errors and E2 corresponds to 60% of errors, while in MEDLINE, the proportion is 50/50. In the case of E1, a disambiguation of the multiple concepts returned by the SIFR Annotator would be an effective solution to the problem, as

previously mentioned for ambiguous Semantic Groups annotations in PER. The main cause for E2 errors is that the expert annotators did not annotate with all possible CUIs but picked one CUI among many possibilities. Therefore, the SIFR Annotator might return more specific or more general concept, which are not incorrect but which result from different annotation perspectives.

#### Death certificate coding errors

Similarly, we sampled 200 false positives and false negatives from the best runs of the SIFR Annotator on the French aligned development dataset and proceeded to manually determine the causes of the errors (Table 7).

The most frequent types of error are the following (see examples in Table 7):

- (79%) Errors because of missing synonyms that cannot be matched at a string-match level.
- (16%) Morphosyntactic or lexical variation (e.g., accent, dash, comma, spelling). The errors due to morphosyntactic variation (and more general concept annotation due to a partial match) have the same cause that similar errors in the PER and NER evaluations and their possible solutions are the same: an alternative concept recognizer. The mapping expansion mechanism in SIFR Annotator could tackle such an issue, but there are very few mappings to and from CIM-10 at the moment. All phenomena that are common in reality but not captured as synonyms by the source ontologies will not be recognized properly.
- (2.5%) Annotations were made with a more specific code (i.e., child in ICD-10 hierarchy) compared to the gold standard, often because of a partial match within a phrase.
- (2.5%) Errors caused by implicit semantic information that requires medical knowledge to identify. In both examples in Table 7, the code to identify is very general and the text does not really convey the coding explicitly; perhaps other fields in

the data or in the knowledge of the experts helped them to code this death certificate meaningfully. This issue can hardly be remedied in the context of the SIFR Annotator as it is a process at a higher order of complexity than merely performing concept annotations (complex semantic inference).

#### Limitations & future prospects

The purpose of the SIFR Annotator, and originally of the NCBO Annotator [13, 24], was not to beat task-specific state-of-the-art annotation systems. The goal was to offer generic but quite accurate workflow directly connected to their respective ontology repository. The concrete advantages of the services come from: (i) the size and variety of their dictionaries coming from ontologies, (ii) their availability as a web service that can be easily included in any semantic indexing workflow, and finally (iii) their adoption of a semantic web vision that strongly encourages using dereferenceable URIs that can then be reused to facilitate data integration and semantic interoperability. One should also note that the semantic expansion step (which uses the mappings between ontologies and the *is\_a* hierarchies to generate additional annotations) as well as the post-processing of the annotations (which scores and contextualizes the annotations) are interesting exclusive features that are evaluated neither with the Quaero corpus nor in CLEF eHealth 2017 task 1.

That being said, the main limitations we can draw from our evaluations and from the error analyses from the perspective of annotation tasks are the following:

- The concept recognition component (Mgrep) used in SIFR BioPortal is limited in some aspects compared to current state-of-the-art, however, it offers significant advantages in a few contexts. Mgrep favors precision over recall and has been shown to almost always outperform MetaMap [12]. Moreover, Mgrep is agnostic with regard to the annotating resources, while many other systems are coupled with the UMLS Metathesaurus only (e.g.,

**Table 7** Most frequent SIFR Annotator errors for the death certificate coding task at CLEF eHealth 2017

Error	Example	Percent
Formulation different from synonym labels for expected concept	"arrêt respiratoire" (R09.2) not identified in "arrêt cardio respiratoire" or "détresse cardiorespiratoire."	79%
Morphosyntactic variation	"Arrêt respiratoire" (R09.2) not identified in text "arrêt réspiratoire" due to incorrect diacritic.	16%
Annotation with a more general code (higher in the concept hierarchy)	"coma d'origine indéterminée et arrêt respiratoire progressif" matched with a more specific code, while the gold standard expects "arrêt respiratoire" (R09.2)	2.5%
Correct annotation dependent of detecting implicit semantic information	Code I10 "hypertension essentielle (primitive)" is hard to identify from "TC suite à une chute avec épilepsie séquellaire et tr cognitifs" as expected in the gold standard. Code R68.8 "autres symptômes et signes généraux précisés" was not identified within the text "atteinte polyviscérale diffuse."	2.5%

MetaMap). Tools using more advanced NLP techniques (fuzzy matching, syntactic analysis, language model-based matching) can lead to equally precise annotations with an increased recall, but at the cost of execution speed. The main disadvantages of Mgrep are: simple string matching; closed-source and difficult to improve upon. Mgrep was chosen regardless of limitations because precision is more important than recall (for biomedical annotation) and in a production setting, the speed of the matching is of the utmost importance.<sup>27</sup> Since we cannot contribute to Mgrep, the best course of action is the development of a new concept recognition component. Such a development is already underway and under active testing, for a potential release date in late 2018.

- The ontological resources publicly available for French are limited compared to resources for English and much work may be done to release new public ontologies and to engineer new ontologies for domains not covered by existing resources. Even since the inception of the SIFR project, this has been a major goal and an active effort, much more is needed. We are for instance collaborating with pharmacologists to build a comprehensive and legally recognized resource for medication and drugs in French that is interoperable with international ATC codes. We are also actively incorporating new terminologies and ontologies in the SIFR BioPortal. In the future we also plan to automatically enrich any semantic resources in the repository with Semantic Types using machine learning in order to continue to offer annotations at different level of granularity even for ontologies that have never been integrated in the UMLS.
- The SIFR BioPortal is a multi-ontology approach where all labels belong to a single dictionary, which leads to annotation ambiguities at different granularities (concepts, CUIs, Semantic Types or Groups). The SIFR Annotator therefore requires a multi-level disambiguation module, as previously discussed.

Besides those limitations, the SIFR Annotator has significant advantages that are not highlighted in the evaluation tasks. One advantage is the ability to exploit the hierarchy, to obtain an annotation of a text at different levels of semantic granularity, which in turn can be effectively exploited for indexing large amounts of biomedical or clinical data. Annotations of terms with higher level parents allows to capture a very broad thematic semantic information, and can be exploited for text classification, while more specific annotations can be used for general purpose indexing or for knowledge extraction.

Another advantage of SIFR BioPortal and Annotator is the ability for users to contribute mappings between ontologies. Mappings correspond to explicit equivalence relations between ontology concepts. The original BioPortal infrastructure supports the loading of explicit mappings between ontologies contained in the repository but also automatically generates mappings based on class labels, URIs or CUIs. Those mappings can be used for annotation. For example, to annotate with one target ontology (e.g., ICD-10 for diagnostic coding), while still benefiting from the labels and alternative labels accessible through mappings during the concept recognition phase.

SIFR BioPortal additionally supports interportal mappings that can refer to ontologies in NCBO-like ontology repository. In previous work, we have reconciled and uploaded in the SIFR BioPortal 228 K French/English interportal mappings for UMLS ontologies between SIFR and NCBO BioPortal [70]. In a multilingual context, in the future we could, for instance, annotate French text with English concepts (or vice versa) in order to generate comparable corpora indexes across languages (an invaluable resource for cross-lingual text mining and information retrieval).

Adapting the BioPortal technology to Spanish is a possible future extension of the SIFR Annotator technology. Not only does Spanish already have numerous medical ontologies and terminologies, but the potential impact for clinical text annotations that are interoperable between Spanish and English is extremely significant, especially in the context of the linguistic landscape in the United States, where Spanish speaking communities are an important demographic. As an example, such an adaptation would allow English-speaking doctors to access the essential information found in Spanish language clinical health records, when treating Spanish speaking patients. We are in the process of identifying relevant partners to concretize such project.

## Conclusions

We presented the development and evaluation of SIFR Annotator, a semantic free-text annotation service for French made available in the SIFR BioPortal ontology repository, based on technology from NCBO BioPortal. We adapted the technology for the French language and extended the original features to be more suitable for multi-level annotation of clinical text and other possible scenarios.

We have shown the SIFR Annotator web service is comparable, in terms of quality and annotation performance to other knowledge-based annotation approaches in the two presented tasks, although the task objectives were not directly compatible with our annotation approach.<sup>28</sup> We believe that SIFR Annotator can help in a wide range of text mining or annotation problems, but of course not universally. We also highlighted the

shortcomings of our SIFR Annotator tool and proposed some possible solutions for their mitigation in future technical evolutions of the service.

Our work on SIFR Annotator, is not limited to French, however, the technical efforts have mainly been focused on decoupling the architecture from English and for allowing an easy adaptation to other languages. Although our target language is French, we have made some of our new features also available for English [14] and we believe our efforts and experience would facilitate deployment of new instance of BioPortal and its Annotator in other language (especially roman language or linguistically close to French/English) after minor configuration and adjustments. Such an adaptation does not dispense from the gargantuan task of gathering and engineering ontologies in other languages, but it gives a platform to make the efforts meaningful.

SIFR BioPortal has become the largest generic and open—with publicly access resources, code and related data—French-language biomedical ontology and terminology repository in France. In turn, SIFR Annotator is today the richest French language open annotator web service (competing annotators are either not available or closed-source online services). We are currently developing several partnerships in France to use SIFR Annotator within hospitals (CHRU Nancy, George Pompidou European Hospital in Paris) or for large-scale annotation efforts (e.g., to annotate the corpus of course of the French national medicine curriculum in the SIDES 3.0 project).

## Availability and requirements

**Project name:** SIFR Annotator

**Web application:** <http://biportal.lirmm.fr/annotator>

**Project home page:** <http://www.lirmm.fr/sifr>

**Code repository:** <http://github.com/sifrproject>

**NCBO codebase:** [https://github.com/sifrproject/ncbo\\_annotator](https://github.com/sifrproject/ncbo_annotator)

Proxy: <https://github.com/sifrproject/annotators>

**Operating system(s):** The Web application is platform independent. An easy local deployment procedure is available using Docker to process sensitive (e.g., clinical) data in-house (<https://github.com/sifrproject/docker--compose-biportal>). This works on Linux.

**Programming language:** Ruby 2.3 (NCBO codebase) + Java 8 (Proxy)

**Other requirements:** When deploying manually: Rails 4, Tomcat 8, Redis, Memcached, MySQL, Apache HTTP Sever + Phusion passenger. When deploying with Docker, a Linux system, Docker, Docker Compose.

**License:** Stanford NCBO code based is Licensed as BSD-2. LIRMM's modification to codebase and Proxy's implementation is open source (License not yet determined).

## Endnotes

<sup>1</sup>Centre d'épidémiologie sur les causes médicales de décès, Unité Inserm US10, (<http://www.cepidc.inserm.fr>)

<sup>2</sup>Article currently under review (second round) in Journal of Biomedical Informatics (JBI-17-745).

<sup>3</sup>NCBO BioPortal hosts some non-English ontologies, most of the time (but not always) as “views” of their English counterparts.

<sup>4</sup>The Resource Description Framework (RDF) is the W3C language to described data. It is the backbone of the semantic web. SPARQL is the corresponding query language. By adopting RDF as the underlying format, an ontology repository based on NCBO technology can easily make its data available as linked open data and queryable through a public SPARQL endpoint. To illustrate this, the reader may consult the Link Open Data cloud diagram (<http://lod-cloud.net>) that since 2017 includes ontologies imported from the NCBO BioPortal (most of the Life Sciences section).

<sup>5</sup>[www.bioontology.org/wiki/index.php/Category:NCBO\\_Virtual\\_Appliance](http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance)

<sup>6</sup>Not to be confused with the 28 terminologies in SIFR BioPortal.

<sup>7</sup>Please note that NCBO Annotator is always present among the systems being compared in various reviews [58, 71, 72].

<sup>8</sup>Named Entity Recognition (or entity extraction) is the process of locating and categorizing named entities in text. We consider semantic annotation, i.e., the process of locating concepts previously defined in ontologies into text, as a subtask of NER.

<sup>9</sup>For WHOFRE and ICPCFRE we had another version and LNC-FR-FR has not been included yet because it has a specific format.

<sup>10</sup>CISMeF's repository is larger, but the content is not publicly accessible.

<sup>11</sup><https://goo.gl/rccsJi>

<sup>12</sup><https://goo.gl/5mJmgv>

<sup>13</sup>You may refer to the API documentation: <http://data.biportal.lirmm.fr/documentation>.

<sup>14</sup>The filtering by Semantic Type was available in the core Annotator components, however, we extended this feature to Semantic Groups which are themselves defined by grouping Semantic Types [14].

<sup>15</sup>Since BioPortal 4.0 (end of 2013), the scoring has been removed from the NCBO Annotator. In our work we have re-implemented the original score and offered two better ones.

<sup>16</sup>For instance, with an absolute score threshold of 3.1: <https://goo.gl/yKe8gY>

<sup>17</sup>Proxy in the sense of the architectural software design principle applied to microservice architectures, not to be confused with an HTTP Proxy, a tool to secure external internet access from within a closed network.

<sup>18</sup>All parameter estimation efforts are performed on the original Quaero corpus with the aforementioned baseline parameters.

<sup>19</sup>Each death certificate is a document and contains exactly 4 lines.

<sup>20</sup><http://bioportal.lirmm.fr/ontologies/CIM-10>

<sup>21</sup>We are currently interacting with the CéPIDC to potentially publicly release the dictionary we've build for third parties.

<sup>22</sup>The corresponding URI in CIM-10 and thus in CIM-10 DC is <http://chu-rouen.fr/cismef/CIM-10#G12.1>, where <http://chu-rouen.fr/cismef/CIM-10#> is the base URI and G12.1 the code identifier.

<sup>23</sup>The evaluation program for the death certificate coding task was implemented in Java, in the same repository as for the named entity recognition evaluation on the Quaero corpus.

<sup>24</sup>A first version of the system has been published in a French peer-reviewed workshop [17]. A more complete evaluation and system description is currently under review (second round) for JBI.

<sup>25</sup>Annotators were supplied with entity-pre-annotations but were given free rein to delete annotations or add new entity annotations. The paper reports that a significant number of entities were added after the pre-annotation process, however, exhaustivity is difficult to achieve.

<sup>26</sup>The 10% breakdown into 2% incorrect single semantic group annotations and 8% ambiguous annotations with multiple semantic groups.

<sup>27</sup>An experimental recognizer based on Mallet and AlvisNLP have been built in the context of a past hackathon but proved incompatible with a production-grade annotation system due to their slow processing speed.

<sup>28</sup>It is important to note our system was not specifically tailored for these tasks and that performance will highly vary depending of the data to annotate and the ontologies targeted.

## Acknowledgements

We would like to thank the US National Center for Biomedical Ontology at Stanford University, for their assistance with the NCBO Annotator, the CépiDC and CLEF eHealth organizers (Aurélié Névél in particular) for their authorization to use the corpora (Quaero and 2017 Task 1), the European Hospital Georges Pompidou for the evaluation of ConText on a clinical corpus, as well as all users of the SIFR Annotator that guide us in its development. We would also like to thank CISMef, an early partner of the SIFR project, for providing numerous health terminologies.

## Funding

This work was funded by the Semantic Indexing of French biomedical Resources (SIFR – [www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) and PractiKPharma projects (<http://practikpharma.loria.fr>) that received funding from the French National Research Agency (grant ANR-12-JS02-01001 and ANR-15-CE23-0028) as well as by the European H2020 Marie Skłodowska-Curie action (agreement No 701771), the University of Montpellier and the CNRS.

## Availability of data and materials

- Data for NER Evaluation: Quaero corpus available at <https://quaerofrenchmed.limsi.fr> under the GNU Free Documentation License. Evaluation programs and sub-corpus generators available at <https://github.com/twktheainur/bpannotatoreval>.
- Data for CLEF eHealth evaluation: CépiDC data not publicly available due to privacy constraints (clinical text), access to data possible upon signature of a data use agreement as per the official instructions for the task <https://sites.google.com/site/clefehealth2017/task-1>.
- Data for Clinical Context evaluation: Clinical corpus not publicly available as it is based on actual electronic health records from the European Hospital Georges Pompidou. Our experiments were produced in-house with partnership with an hospital team.

## Authors' contributions

AT contributed to the development of the proxy architecture, the implementation of the most of the additional features (except scoring) offered in SIFR Annotator, he devised the experimental protocol for the Biomedical NER evaluation and for CLEF eHealth 2017 evaluation, performed the evaluation experiments, wrote a significant portion of the article. AA performed the adaptation of ConText/NegEx to French, the enrichment of the trigger terms and the evaluation of the feature. He also devised the dictionary construction heuristics and participated in the CLEF eHealth 2017 evaluation efforts. VE participated in the development of SIFR BioPortal in general and SIFR Annotator in particular and provided engineering support. SZ performed preliminary experiments on Quaero and on the CLEF eHealth 2017 corpus and undertook the error analysis for both evaluation tasks. CJ is the PI of the SIFR and co-PI of PractiKPharma ANR projects and recipient of the H2020 Marie Skłodowska-Curie action, he is the original creator of NCBO Annotator and the progenitor of SIFR BioPortal and SIFR Annotator. He directed all the research efforts and participated in the writing of the article significantly. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

CISMef/SIBM (see related work section) were an early partner of the SIFR project.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier, CNRS, 161, rue Ada, 34095 Montpellier cedex 5, France. <sup>2</sup>Center for Biomedical Informatics Research (BMIR), Stanford University, 1265 Welch Rd, Stanford, CA 94305, USA. <sup>3</sup>LGI2P, IMT Mines Ales, Univ Montpellier, Alès, France.

Received: 20 June 2018 Accepted: 10 October 2018

Published online: 06 November 2018

## References

1. Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. In: AMIA Annual Symposium Proceedings. Washington D.C: AMIA; 2006. p. 106–110.
2. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the semantic web. *BMC Bioinformatics*. 2007;8:S2. <https://doi.org/10.1186/1471-2105-8-S2>.
3. Drolet BC, Lorenzi NM. Translational research: understanding the continuum from bench to bedside. *Transl Res*. 2011;157:1–5. <https://doi.org/10.1016/j.trsl.2010.10.002>.
4. Blake JA. Bio-ontologies—fast and furious. *Nat Biotechnol*. 2004;22:773–4.
5. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinform*. 2008;9:75–90.

6. Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, et al. Semantic annotation for knowledge management: requirements and a survey of the state of the art. *Web Semant Sci Serv Agents World Wide Web*. 2006;4:14–28.
7. Névél A, Grosjean J, Darmoni SJ, Zweigenbaum P. Language Resources for French in the Biomedical Domain. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al., editors. 9th International Conference on Language Resources and Evaluation, LREC'14. Reykjavik, Iceland: European Language Resources Association; 2014. p. 2146–51.
8. Névél A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. *J Biomed Semantics*. 2018;9:12. <https://doi.org/10.1186/s13326-018-0179-8>.
9. Jonquet C, Annane A, Bouarech K, Emonet V, Melzi S. SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale JFIM'16; 2016.
10. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith NB, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*. 2009;37(web server):170–3.
11. Whetzel PL, Team N. NCBO Technology: Powering semantically aware applications. *Biomed Semant*. 2013;4S1:49.
12. Shah NH, Bhatia N, Jonquet C, Rubin DL, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*. 2009;10(9):S14.
13. Jonquet C, Shah NH, Musen MA. The Open Biomedical Annotator. In: American Medical Informatics Association Symposium on Translational Bioinformatics, AMIA-TBI'09. San Francisco: AMIA; 2009. p. 56–60.
14. Tchechmedjiev A, Abdaoui A, Emonet V, Melzi S, Jonnagadala J, Jonquet C. Enhanced Functionalities for Annotating and Indexing Clinical Text with the NCBO Annotator+. *Bioinformatics*. 2018. <https://doi.org/10.1093/bioinformatics/bty009>.
15. Névél A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In: Proceedings of the 4th Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing, BioTxtM'14. Reykjavik, Iceland: Manchester: NaCTEM; 2014. p. 24–30.
16. Goeuriot L, Kelly L, Suominen H, Névél A, Robert A, Kanoulas E, et al. CLEF 2017 eHealth Evaluation Lab Overview. In: Jones G. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017. Lecture Notes in Computer Science, vol 10456. Cham: Springer; 2017.
17. Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français. In: 4ème Symposium sur l'Ingénierie de l'Information Médicale, SIIM'17. Toulouse; 2017. p. 10. [http://www.lirmm.fr/~jonquet/publications/documents/Article\\_SIIM2017\\_FrenchContext.pdf](http://www.lirmm.fr/~jonquet/publications/documents/Article_SIIM2017_FrenchContext.pdf).
18. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007;25:1251–5.
19. Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, et al. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res*. 2016;45:D347–52.
20. Côté RG, Jones P, Apweiler R, Hermjakob H. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*. 2006;7:7.
21. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV. Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*. 2015;16:1–9.
22. Jonquet C, Emonet V, Musen MA. Roadmap for a multilingual BioPortal. In: Gracia J, McCrae JP, Vulcu G, editors. 4th workshop on the multilingual semantic web, MSW4'15. Portoroz; 2015. p. 15–26. [http://www.lirmm.fr/~jonquet/publications/documents/Article\\_MSW4\\_MultilingualBioPortal.pdf](http://www.lirmm.fr/~jonquet/publications/documents/Article_MSW4_MultilingualBioPortal.pdf).
23. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32:267–70.
24. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. *BMC Bioinformatics*. 2009;10(2):S1. <https://doi.org/10.1186/1471-2105-10-S2-S1>.
25. Graybeal J, Isenor AW, Rueda C. Semantic mediation of vocabularies for ocean observing systems. *Comput Geosci*. 2012;40(120):131.
26. Jonquet C, Toulet A, Arnaud E, Aubin S, Yeumo ED, Emonet V, et al. AgroPortal: an ontology repository for agronomy. *Comput Electron Agric*. 2018;144:126–43. <https://doi.org/10.1016/j.compag.2017.10.012>.
27. Zweigenbaum P, Baud R, Burgun A, Namer F, Jarrousse A, Grabar N, et al. Towards a unified medical lexicon for French. *Stud Health Technol Inform*. 2003;95:415–20.
28. Darmoni SJ, Jarrousse E, Zweigenbaum P, Beux PL, Namer F, Baud R, et al. VUMeF: extending the French involvement in the UMLS Metathesaurus. In: American Medical Informatics Association Annual Symposium, AMIA'03. Washington DC: AMIA; 2003. p. 884.
29. Joubert M. project consortium InterSTIS. Interopérabilité sémantique de terminologies de santé francophones. *Ingénierie Rech Biomédicale*. 2011;32:80–2.
30. S J D, Joubert M, Dahamna B, Delahousse J, Fieschi M. SMTs: a French Health Multi-terminology Server. In: American Medical Informatics Association Annual Symposium, AMIA'09. Washington DC: AMIA; 2009. p. 808.
31. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health Multi-Terminology Portal: a semantics added-value for patient safety. In: Koutkias V, Niès J, Jensen S, Maglaveras N, Beuscart R, editors. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety. Amsterdam: IOS Press; 2011. p. 129–38.
32. Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni S. Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. In: 9th International Conference on Terminology and Artificial Intelligence, TIA'11. Paris; 2011. p. 119–22.
33. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, Darmoni SJ. Health Multi-Terminology Portal: a semantics added-value for patient safety. *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, Studies in Health Technology and Informatics, Volume 166*. Amsterdam: IOS Press; 2011. p. 129–38.
34. Darmoni SJ, Thirion B, Leroy JP, Douyère M, Lacoste B, Godard C, Rigolle I, Brisou M, Videau S, Goupy E, Piot J, Quéré M, Ouazir S, Abdurab H. DocCISMEF: a search tool based on "encapsulated" MeSH thesaurus. *Studies in Health Technology and Informatics, Volume 84–1*. Amsterdam: IOS Press; 2001. p. 314–8.
35. McCool RGR, Miller E. Semantic search. In: 12th international conference on world wide web, WWW'03. Budapest, Hungary: ACM; 2003. p. 700–9.
36. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform*. 2001;84:216.
37. McCray AT. An upper-level ontology for the biomedical domain. *Comp Funct Genomics*. 2003;4:80–4.
38. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through web services: calling Whatizit. *Bioinformatics*. 2008; 24:296–8.
39. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, AMIA'01. Washington, DC: AMIA; 2001. p. 17–21.
40. Sakji S, Gicquel Q, Pereira S, Kergoulay I, Proux D, S J D, et al. Evaluation of a French Medical Multi-Terminology Indexer for the Manual Annotation of Natural Language Medical Reports of Healthcare-Associated Infections. In: et al. CS, editor. 13th World Congress on Medical Informatics, MedInfo'10. Cape Town, South Africa: IOS Press; 2010. p. 252–6.
41. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névél A, Grouin C, et al. Overview of the CLEF eHealth Evaluation Lab 2015. In: Mothe J, Savoy J, Kamps J, Pinel-Sauvagnat K, Jones G, San Juan E, et al., editors. Experimental IR meets Multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2015. p. 429–43.
42. Kelly L, Goeuriot L, Suominen H, Névél A, Palotti J, Zuccon G. Overview of the CLEF eHealth Evaluation Lab 2016. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, et al., editors. Experimental IR meets Multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2016. p. 255–66.
43. Afzal Z, Akhondi SA, van Haagen H, Van Mulligen E, Kors JA. Biomedical concept recognition in French text using automatic translation of English terms. In: Working Notes of CLEF eHealth Evaluation Lab; 2015. p. 16.
44. Soualmia LF, Cabot C, Dahamna B, Darmoni SJ. SIBM at CLEF e-health evaluation lab 2015. In: Working Notes of CLEF eHealth Evaluation Lab; 2015. p. 16.
45. van Mulligen EM, Afzal Z, Akhondi SA, Vo D, Kors JA, Erasmus MC at CLEF eHealth 2016: concept recognition and coding in French texts. In: Working Notes of CLEF eHealth Evaluation Lab; 2016. p. 16.

46. Cabot C, Soualmia LF, Dahamna B, Darmoni S. SIBM at CLEF eHealth evaluation lab 2016: extracting concepts in French medical texts with ECMT and CIMIND. In: Working Notes of CLEF eHealth Evaluation Lab; 2016. p. 16.
47. Dermouche M, Looten V, Flicoteaux R, Chevret S, Velcin J, Taright N. ECSTRA-INSERM @ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. In: Working Notes of CLEF eHealth Evaluation Lab; 2016. p. 16.
48. Zweigenbaum P, Lavergne T. LIMS ICD10 coding experiments on CépIDC death certificate statements. In: Working Notes of CLEF eHealth Evaluation Lab; 2016. p. 16.
49. Ho-Dac L-M, Fabre C, Birsik A, Boudraa I, Bourriot A, Cassier M, et al. LITL at CLEF eHealth2017: Automatic Classification of Death Reports. In: Working Notes of CLEF eHealth Evaluation Lab. Dublin; 2017. p. 16.
50. Tchechmedjiev A, Abdaoui A, Emonet V, Jonquet C. ICD10 Coding of Death Certificates with the NCBO and SIFR Annotator(s) at CLEF eHealth 2017 Task 1. In: Working Notes of CLEF eHealth Evaluation Lab. Dublin; 2017. p. 16. [http://ceur-ws.org/Vol-1866/paper\\_62.pdf](http://ceur-ws.org/Vol-1866/paper_62.pdf).
51. Atermezog GA. NoNLP: annotating medical domain by using semantic technologies. In: Working Notes of CLEF eHealth Evaluation Lab; 2017. p. 16.
52. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol*. 2013;9:1–16. <https://doi.org/10.1371/journal.pcbi.1002854>.
53. Cabot C, Soualmia LF, Darmoni SJ. SIBM at CLEF eHealth evaluation lab 2017: multilingual information extraction with CIM-IND. In: Working Notes of CLEF eHealth Evaluation Lab; 2017. p. 16.
54. Ševa J, Kittner M, Roller R, Leser U. Multi-lingual ICD-10 coding using a hybrid rule-based and supervised classification approach at CLEF eHealth 2017. In: Working Notes of CLEF eHealth Evaluation Lab; 2017. p. 16.
55. Tchechmedjiev A, Jonquet C. Enrichment of French Biomedical Ontologies with UMLS Concepts and Semantic Types for Biomedical Named Entity Recognition Through Ontological Semantic Annotation. In: Workshop on Language, Ontology, Terminology and Knowledge Structures, LOTKS'17. Montpellier: ACL; 2017. p. 8. [http://www.lirmm.fr/~jonquet/publications/documents/Article\\_LOTKS2017\\_Enrichment.pdf](http://www.lirmm.fr/~jonquet/publications/documents/Article_LOTKS2017_Enrichment.pdf).
56. Annane A, Emonet V, Azouaou F, Jonquet C. Multilingual mapping reconciliation between english-french biomedical ontologies. In: WIMS: Web Intelligence, Mining and Semantics; 2016.
57. Dai M, Shah NH, Xuan W, Musen MA, Watson SJ, Athey BD, et al. An Efficient Solution for Mapping Free Text to Ontology Terms. In: AMIA Symposium on Translational Bioinformatics, AMIA-TBI'08. San Francisco: AMIA; 2008.
58. Funk C, Baumgartner W, Garcia B, Roeder C, Bada M, Cohen KB, et al. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*. 2014;15:59.
59. Simon N, Joey T, Geiger JS. Using the NCBO web Services for Concept Recognition and Ontology Annotation of expression datasets. In: Marshall MS, Burger A, Romano P, Paschke A, Splendiani A, editors. Workshop on semantic web applications and tools for life sciences, SWAT4LS'09. Amsterdam: CEUR-WS.org; 2009.
60. Sarkar IN. Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts. In: American Medical Informatics Association Annual Symposium, AMIA'10. Washington DC: AMIA; 2010. p. 717–21.
61. Groza T, Oellrich A, Collier N. Using silver and semi-gold standard corpora to compare open named entity recognisers. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on; 2013. p. 481–5.
62. Xuan W, Dai M, Mirel B, Athey B, Watson SJ, Meng F. Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK: Linking Literature, Information and Knowledge for Biology, SIG, ISMB'08. Vienna; 2007. p. 55–8.
63. Melzi S, Jonquet C. Scoring semantic annotations returned by the NCBO Annotator. In: Paschke A, Burger A, Romano P, Marshall MS, Splendiani A, editors. 7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14. Berlin: CEUR-WS.org; 2014. p. 15.
64. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Biomed Informatics*. 2001;34:301–10. <https://doi.org/10.1006/jbin.2001.1029>.
65. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experienter, and temporal status from clinical reports. *J Biomed Inform*. 2009;42:839–51. <https://doi.org/10.1016/j.jbi.2009.05.002>.
66. Chapman WW, Hilert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. *Stud Health Technol Inform*. 2013;192:677–81. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3923890/>.
67. Navigli R. Word sense disambiguation: A survey. *ACM Comput Surv*. 2009; 41(10):1–10:69.
68. McInnes BT, Stevenson M. Determining the difficulty of word sense disambiguation. *J Biomed Inform*. 2014;47:83–90. <https://doi.org/10.1016/j.jbi.2013.09.009>.
69. Névél A, Robert N, Anderson CK, Grouin C, Lavergne T, Rey G, Robert A, et al. CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 Evaluation labs and workshop: online working notes, CEUR-WS, September, 2017; 2017.
70. Annane A, Emonet V, Azouaou F, Jonquet C. Réconciliation d'alignements multilingues dans BioPortal. In: Pernelle N, editor. 27èmes Journées Francophones d'Ingénierie des Connaissances, IC'16. Montpellier; 2016. p. 12. [http://www.lirmm.fr/~jonquet/publications/documents/Article\\_IC2016\\_Reconciliation.pdf](http://www.lirmm.fr/~jonquet/publications/documents/Article_IC2016_Reconciliation.pdf).
71. Jovanović J, Bagheri E. Semantic annotation in biomedicine: the current landscape. *J Biomed Semantics*. 2017;8:44. <https://doi.org/10.1186/s13326-017-0153-x>.
72. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS. NOBLE – flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*. 2016;17:32. <https://doi.org/10.1186/s12859-015-0871-y>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)





# Harnessing the Power of Unified Metadata in an Ontology Repository: The Case of AgroPortal

Clement Jonquet<sup>1,2</sup> · Anne Toulet<sup>1</sup> · Biswanath Dutta<sup>3</sup> · Vincent Emonet<sup>1</sup>

Received: 20 July 2017 / Revised: 20 July 2018 / Accepted: 2 August 2018  
© The Author(s) 2018

## Abstract

As any resources, ontologies, thesaurus, vocabularies and terminologies need to be described with relevant metadata to facilitate their identification, selection and reuse. For ontologies to be FAIR, there is a need for metadata authoring guidelines and for harmonization of existing metadata vocabularies—taken independently none of them can completely describe an ontology. Ontology libraries and repositories also have to play an important role. Indeed, some metadata properties are intrinsic to the ontology (name, license, description); other information, such as community feedbacks or relations to other ontologies are typically information that an ontology library shall capture, populate and consolidate to facilitate the processes of identifying and selecting the right ontology(ies) to use. We have studied ontology metadata practices by: (1) analyzing metadata annotations of 805 ontologies; (2) reviewing the most standard and relevant vocabularies (23 totals) currently available to describe metadata for ontologies (such as Dublin Core, Ontology Metadata Vocabulary, VoID, etc.); (3) comparing different metadata implementation in multiple ontology libraries or repositories. We have then built a new metadata model for our AgroPortal vocabulary and ontology repository, a platform dedicated to agronomy based on the NCBO BioPortal technology. AgroPortal now recognizes 346 properties from existing metadata vocabularies that could be used to describe different aspects of ontologies: intrinsic descriptions, people, date, relations, content, metrics, community, administration, and access. We use them to populate an internal model of 127 properties implemented in the portal and harmonized for all the ontologies. We—and AgroPortal’s users—have spent a significant amount of time to edit and curate the metadata of the ontologies to offer a better synthesized and harmonized information and enable new ontology identification features. Our goal was also to facilitate the comprehension of the agronomical ontology landscape by displaying diagrams and charts about all the ontologies on the portal. We have evaluated our work with a user appreciation survey which confirms the new features are indeed relevant and helpful to ease the processes of identification and selection of ontologies. This paper presents how to harness the potential of a complete and unified metadata model with dedicated features in an ontology repository; however, the new AgroPortal’s model is not a new vocabulary as it relies on preexisting ones. A generalization of this work is studied in a community-driven standardization effort in the context of the *RDA Vocabulary and Semantic Services Interest Group*.

**Keywords** Ontology metadata vocabulary · Semantic description · Ontology repository · Ontology selection · Ontology relation · BioPortal · AgroPortal

---

✉ Clement Jonquet  
jonquet@lirmm.fr

Anne Toulet  
toulet@lirmm.fr

Biswanath Dutta  
bisu@drtc.isibang.ac.in

Vincent Emonet  
emonet@lirmm.fr

<sup>1</sup> Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), CNRS and University of Montpellier, Montpellier, France

<sup>2</sup> Center for Biomedical Informatics Research (BMIR), Stanford University School of Medicine, Stanford, CA, USA

<sup>3</sup> Documentation Research and Training Centre (DRTC), Indian Statistical Institute, Bangalore, India

# 1 Introduction

In 2007, Swoogle's homepage [1] announced searching over 10.000 ontologies. Today, a simple Google Search for “file-type:owl” returns around 34 K results. How much ontologies are available online now? The big data deluge and the adoption of the semantic web to semantically describe and link these data [2] have made the number of ontologies grow to numbers for which machines are mandatory to index, search and select them. It has become cumbersome for domain experts to identify the ontologies to use so that automatic recommender systems have been designed to help them with this task, as for instance in the biomedical domain [3]. However, machines need metadata to facilitate the exploitation of any data, including ontologies. It is established that metadata is often too much neglected by data providers [4] even if it is now identified as a requirement to make the data FAIR [5]. But as any other data, ontologies have themselves to be *Findable, Accessible, Interoperable, and Re-usable*. Although there are multiple dimensions to make ontologies FAIR, one will agree developing open ontology repositories, and libraries is one of them. Such libraries are the best environment in which the metadata about ontologies can be described and valued. However, can we say that ontology developers describe their ontologies with relevant metadata properties that will facilitate manual or automatic search, identification and selection of ontologies? There exists a significant number of metadata vocabularies that could be used for ontologies but none of the existing ones can completely meet this need if taken independently. Therefore, how can we make ontologies more FAIR?

When someone is interested in an ontology, he/she may like to know: Who edited or contributed? When? What methodology or tool was used? Which natural language is used? Which formats are available? What is the metrics? Is it free of use or licensed? Who is using it? In addition, when someone is interested about ontologies of a domain, he/she may like to know: How ontologies can be grouped together? Which are most used? What are the relations between them? What are the common practices? Who are the key contributors of the domain? Or the most important organizations? All this information can be represented by metadata properties. Capturing that information is both a technical challenge—we need models, tools and automated population—and a data curation challenge. Indeed, the information or metadata about an ontology is often dispatched within web sites, scientific articles, documentation or sometimes not existing at all except in the brain of the original ontology developers. There is a need for metadata authoring guidelines and for harmonization of existing metadata vocabularies to simplify their use and enlarge their adoption. For instance, the recent *Minimum Information for Reporting of an Ontology* initiative (<https://github.com/owlcs/miro>) [6]

proposes the MIRO guidelines to ontology developers when reporting an ontology, e.g., in a scientific article.

In this paper, we adopt the perspective of designers of an ontology repository and report on our effort to develop a unified ontology metadata model for this repository. We measure its impact on facilitating ontology descriptions, identification and selection. In the following, we will review the current practices related to describing ontologies and using ontology metadata vocabularies. We have observed some limitations, lack of harmonization and confusions in the practices. This is not surprising when considering the efforts needed to just identify the potentially relevant vocabularies that could be used to describe ontologies.<sup>1</sup> Indeed, a few of these vocabularies are dedicated to ontologies and vocabularies (e.g., OMV, DOOR, VOF), or datasets (e.g., VOID, DCAT, SCHEMA) and others capture more general metadata (e.g., DC, DCT, PROV, DOAP).<sup>2</sup> They are often not maintained anymore, sometimes very specific or too general and of course, they are rarely aligned one another despite their significant overlaps. Furthermore, there have been several ontology repository projects that did not also take the problem seriously enough to support the description of their ontologies with standard vocabularies [7, 8]. With the exception of the Linked Open Vocabularies registry [9, 10], the MMI Ontology Registry and Repository [11], and to some extent, the NCBO BioPortal [12], the question of harmonization and standardization of ontology descriptions have not really been a central matter, although this is changing now (e.g., the OBO Foundry community metadata effort). The Linked Open Vocabularies is a good counter example; it has developed and adopted VOF as a unified model to describe metadata and relations between vocabularies. Now, even if the metadata vocabulary is limited (16 properties), the platform has more than 600 resources described with the same model.

In the rest of the paper, we will adopt a definition of metadata including anything that can be said to describe an ontology, structured data or free descriptions: how and why it is built, used, changed, accessed and how it relates to other ontologies and datasets. That will include properties going from (1) intrinsic properties, e.g., name, URI, creation date; (2) relation to other ontologies, e.g., imports,

<sup>1</sup> In this paper, we will consider the terms ontologies, terminologies, thesaurus and vocabularies as the type of knowledge organization systems [42] or knowledge artifacts [41]. Those are the subjects we are interested in describing. However, to facilitate the reading, we will use the word *ontology* to identify the subject that is described by metadata (e.g., Movie Ontology, Human Disease Ontology, MeSH thesaurus, etc.) and the word *vocabulary* to identify the semantic resources used to described ontologies (e.g., OMV, DC, DCAT, etc.).

<sup>2</sup> Please refer to column ‘prefix’ of Table 3 all along the paper for acronyms definitions of metadata vocabularies. We will consistently use upper case acronyms corresponding to the vocabulary namespace throughout the paper to refer vocabularies.

is mapped to, disagrees with; (3) community contributions, e.g., notes, project using, endorsements; (4) content-based properties, e.g., SPARQL endpoint, bulk RDF download, search endpoint. As discussed in the paper, such information when available and properly harmonized facilitates the ontology identification and selection processes, which has been assessed as crucial to enable ontology reuse [6, 13–15].<sup>3</sup> In addition, good and harmonized metadata provides information about the ontology landscape, especially when looking at a specific domain. For instance, when looking at the OBO Foundry ontologies [16], one may ask himself (1) if *OBO Edit* is actually the most used tool to develop ontologies stored in the foundry? (2) Who are the key persons in this community to talk to when starting a new ontology? (3) Which are the most involved organizations? (4) Which are the most active ontologies?

In this paper, we have made a systematic review of metadata vocabularies and their properties in order to build a list of metadata properties that can be used to describe ontologies inside our own ontology repository. The objective of this work is not to propose another “vocabulary” for ontology metadata, i.e., a SKOS or OWL resource that we would promote as a new standard to reuse in any ontology description. Indeed, our list relies completely on preexisting vocabularies (cf. discussion in Sect. 7.1). Our objective was to address the need of a common metadata model inside an ontology repository, i.e., implementing a way to compare ontologies side by side and describe the global landscape of all the ontologies in a library or repository.

The list proposed has been built following an analysis of current ontology metadata practices:

- We have reviewed the most standard and relevant vocabularies (23 totals, e.g., Dublin Core, VOID, Ontology Metadata Vocabulary, Data Catalog Vocabulary, etc.) to describe metadata for ontologies. For each of these vocabularies, we have selected the significant properties to describe objects that an ontology could be considered a certain type of, e.g., dataset, an asset, a project or a document. For instance, an ontology may be seen as a `prov:Entity` object and then the property `prov:wasGeneratedBy` may then be used to describe its provenance.
- We have reviewed the current use of metadata vocabularies by sampling 805 ontologies and measuring which vocabularies (and which properties in those vocabularies) are actually used by ontology developers.
- We have studied some of the most common ontology repositories available in the semantic web community, and especially the NCBO BioPortal (which is the reference platform to host and retrieve biomedical ontologies worldwide) to capture in our list, the properties that were actually implemented by the repositories but that would represent an information not specific to the portal. We have considered the features/properties implemented by the portal as “another vocabulary” (later called BioPortal Metadata) incorporated into our list.

As the result, we obtained a list of 346 relevant properties to describe different aspects of ontologies that we have categorized for better understanding. Someone developing an ontology will of course not have to fill them all but can consider them as a list of candidate properties to use. We then grouped those properties into a unified and simplified model of 127 properties that includes the 46 properties originally offered by the NCBO BioPortal and reuses properties of the reviewed metadata vocabularies for the rest [17]. We have implemented this new ontology metadata model within AgroPortal [18], an ontology repository, based on the NCBO technology. AgroPortal hosts ontologies and offers ontology-based services for agronomy, food, plant sciences and biodiversity domains. AgroPortal’s new metadata model supports much more metadata properties than the original NCBO one, enabling very precise description of ontologies. For instance, the model captures which kind of knowledge organization system the file uploaded to the portal is (e.g., thesaurus, ontology, taxonomy, terminology, etc.). We also have properties to capture information such as licenses, ontology editor used, syntax, etc. We can also capture how ontologies are related to other resources (web site, publication, wiki, datasets, etc.) and other ontologies. Most metadata are automatically extracted from the original ontology file, if present, when the ontology is uploaded to the portal. Or it can be in some cases automatically generated by the portal. We have completely refactored the AgroPortal ontology metadata edition page to facilitate the job to ontology developers when uploading an ontology to the portal and manually editing metadata.

With a new edition interface and a common model available for all the ontologies in the portal, we have then spent a significant amount of time to edit and curate ourselves ontology descriptions, and we have asked the ontology developers to validate our edits and complete them. This has resulted in our capability to automatically aggregate information about ontologies and vocabularies to facilitate the comprehension of the whole agronomical ontology landscape by displaying diagrams, charts and networks about all the ontologies on the portal (grouping, types of ontologies, average metrics, most frequent licenses, languages or formats, leading contributors and organizations, most active ontologies, etc.). We have

<sup>3</sup> In this paper, we define *identification and selection* of an ontology as the processes of choosing the right ontology for a given task when searching for ontologies on an ontology library or repository. It can be based on the content of the ontology, its type, community or level of adoption in a community. Sometime this process may be semi-automatized with tools such as the NCBO Recommender (also available in AgroPortal) [3].

added several new features to AgroPortal's ontology description and browsing pages and have now a specific page dedicated to visualizing the “landscape” of ontologies (<http://agroportal.lirmm.fr/landscape>) that displays synthesized information, using diagrams, charts and figures, about the ontologies developed in agronomy with the goal of facilitating ontology identification, selection and get a better comprehension of the landscape of ontologies. Of course, these new functionalities rely on the quality of the metadata extracted from the ontologies or edited on the portal. Such visualizations are also meant to motivate the ontology developers to document and describe more their ontologies. An evaluation survey conducted with AgroPortal's users shows evidence of the influence of ontology metadata on ontology identification and selection and reports on the very positive evaluation of the new functionalities by AgroPortal's users.

The rest of the paper is organized as follows: Sect. 2 presents a few motivating use cases from our work on ontology repositories; Sect. 3 discusses related work in metadata vocabularies and ontology libraries. In Sect. 4, we report on our analysis of current ontology metadata practices that have driven our methodology, described Sect. 5, to select a large list of properties and to implement a restricted and unified new ontology metadata model in AgroPortal. Section 6 presents the results obtained by implementing the new model in AgroPortal, populating the metadata and designing new interfaces to facilitate the comprehension of the ontology landscape. The section also reports about evaluating the new features with AgroPortal's user community. Sections 7 and 8, respectively, discuss the perspectives and issues in ontology metadata and concludes the paper.

## 2 Motivating Use Cases

Our work on ontology metadata is related to our research and development on ontology repositories. Indeed, LIRMM develops and maintains two ontology repositories which are based on the NCBO technology [19]. One, the SIFR BioPortal (<http://bioportal.lirmm.fr>) is developed within the context of the *Semantic Indexing of French biomedical Resources* project and focus on French biomedical ontologies and terminologies. The main goal of the SIFR project is to develop a French Annotator [20] similar to what exists within the NCBO BioPortal [21]. The second ontology repository, AgroPortal (<http://agroportal.lirmm.fr>) [18], targets the agricultural community (not restricted to any language but using English as default) and the project has for primary mission to host and describe vocabularies and ontologies. In the paper, we will only describe the use cases and implementation done within the AgroPortal project; however, it is important to note that this work is generic and has also been implemented in the SIFR BioPortal.

Data integration and semantic interoperability in agronomy—and related domains—have become a crucial scientific challenge. Recently, the research community has adopted the use of ontologies as a common and shared means to describe data make them interoperable and annotate them to build structured and formalized knowledge [22, 23]. The FAIR principles also reinforced that vision [5]. AgroPortal's main objective is to be a reference ontology repository for agronomy, plant sciences, biodiversity, and nutrition. We reused the openly available NCBO BioPortal technology (<http://bioportal.bioontology.org>) [12] to build our first ontology repository and services platform. We have now an advanced prototype, and the latest version (v1.4) was released in July 2017. It currently hosts 100 public semantic resources, with more than 2/3 of them not present in any similar ontology repository (like NCBO BioPortal) and 8 private. Today, AgroPortal offers a robust and reliable service to the community that features ontology hosting, search, versioning, visualization, comment, services for semantically annotating data with the ontologies, as well as storing and exploiting ontology alignments and data annotations.

Among the first feedbacks and requirements of new users were the ability to describe ontology metadata with additional fields that what BioPortal originally provided. For instance, the RDA Wheat Data Interoperability (WDI) working group (<http://ist.blogs.inra.fr/wdi>) recommendations [24] pointed to AgroPortal to find standard wheat-related ontologies, but they needed licensing and access rights information to be more explicit and consistent. The group also required that the endorsement of the WDI for certain ontologies shall be made explicit on AgroPortal, in order to encourage the reuse of some specific ontologies. The LovInra initiative (<http://lovinra.inra.fr>) at the French National Institute for Agricultural Research (INRA) adopted AgroPortal to publish vocabularies produced or co-produced by INRA scientists and foster their reuse beyond the original researchers. They needed to classify knowledge artifacts by types, formats, syntax, and formality.

Besides the “simple addition” of new metadata fields to the original model, the needs expressed by the early AgroPortal adopters were also related to the relations between ontologies and how would the repository help figuring out which ontologies to use. We may cite two concrete examples:

- Several ontologies are developed in parallel to capture wheat (or soy) phenotypes.<sup>4</sup> It became important for AgroPortal to capture the maximum information about the

<sup>4</sup> The Wheat Phenotype ontology [69] and IBP Wheat Trait Ontology developed within the Crop Ontology project [70]. Similarly, the Soy Ontology developed by the curator of the SoyBase database ([www.soybase.org](http://www.soybase.org)) and Soybean ontology also developed in the Crop Ontology project.

ontologies to make explicit to the community which ontology to use depending on their situation. New information such as the organization endorsing or supporting an ontology or the relation between the ontologies are useful metadata in that case.

- Ontologies are never developed isolated. Sometimes capturing the relations between the ontologies is quite cumbersome. For instance, the Planteome project [25] develops reference ontologies for plants such as the Plant Ontology and Plant Trait Ontology. The latter is connected to the specific crop trait ontologies developed within the Crop Ontology project [26]. In addition, they all use Gene Ontology [27] and Phenotype And Trait Ontology [28] to annotate gene products and qualify their phenotypes.

We will show throughout the paper how our new ontology metadata model and realization within AgroPortal help to answer these needs.

### 3 Related Work in Ontology Metadata Description

Metadata is generally described as the data about the data. The topic of ontology or vocabulary metadata is a subset of metadata research in general [4, 29]. In Sect. 4.1, we list metadata vocabularies reviewed from the literature; in the following, we only focus on general papers and references on the subject.

According to Obrst et al. [30], a metadata vocabulary must include a wider range of metadata features. For instance, metadata from a development perspective consists of information such as competency questions, ontological commitments, and design decisions; metadata from an implementation perspective consists of information for reasoning support, languages, rules, conformance to external standards and so forth. Properly defined ontology metadata has been a motivation of several applications of ontologies such as design of ontology repositories and libraries [12, 16, 31–33], ontology selection [34] automatic production of documentation [35], ontology sharing [36].

Capturing the metadata about “electronic objects” has been the original motivation of the DCMI [37] and multiple standardization bodies.<sup>5</sup> The *Dublin Core* (DC) and *DCMI Metadata Terms* (DCT) are the results of these initiatives. Today, semantically rich metadata is identified as one of requirements to produce FAIR data [5] and it becomes the core mission of research projects such as the Center for Expanded Data Annotation and Retrieval [38] which tackles

the challenge of authoring and predicting biomedical datasets metadata.

An important effort has been made in the recent years to define vocabularies for datasets. The Semantic Web Health Care and Life Sciences (HCLS) working group of the W3C have produced a community profile which reviews many of them and proposes a set of recommendations when describing datasets [39]. The FAIRsharing.org action also builds a database of “data and metadata standards, inter-related to databases and data policies” [40] to which AgroPortal’s content is now automatically pushed. More recently, the BioSchemas initiative (<http://bioschemas.org>) has also started a community effort to extend Schema.org with metadata properties that would be relevant for life sciences data. Although we do believe ontologies can somehow be seen as “datasets”—often the closest objects in vocabularies—they have some particularities that require more specific metadata vocabularies as we will see Sect. 5.2.

Ontologies are some kind of knowledge artifacts [41] or knowledge organization systems [42]. Efforts have been made to develop metadata vocabularies or application profiles adapted to such systems, for example, the Networked Knowledge Organization Systems (NKOS) working group [43] or the Ontology Metadata Vocabulary working group [44] which results will be further commented later. The Open Ontology Repository Initiative [32] was a collaborative effort to develop a federated infrastructure of ontology repositories and was also interested in the subject. In 2016, a survey was made to the wide ontology developer community with the goal to capture the *Minimum Information for Reporting of an Ontology* and lead to guidelines, recently published [6], on what should be reported about an ontology and its development, in the context of ontology description papers. Although, the intention is slightly different from our work, we believe most information that can be expressed in a scientific article presenting an ontology—including narrative sections such as motivation, knowledge acquisition or change management—can also be captured as appropriate metadata in the ontology itself; we have included in our ontology metadata model some properties to do so. Recently, a new task group (partially lead by the authors) on “ontology-metadata” has been attached to the Research Data Alliance *Vocabulary and Semantic Services Interest Group*.

Finally, the work on ontology metadata is closely related to the one on ontology libraries and repositories. Indeed, with the growing number of ontologies, ontology libraries and repositories have been of interest in the semantic web community. Ding and Fensel [45] presented in 2001 a review of ontology libraries that introduced the notion of “library.” Then Hartmann et al. [46] introduced the concept of ontology repository, with advanced features such as search, metadata management, visualization, personalization, and mappings. Most ontology libraries are always capturing some metadata

<sup>5</sup> ISO: <http://www.iso.org/schemas/iso25964/> or ISO/IEC: <http://metadata-standards.org/11179/#A3> or ISO/IEC 19763-3:2010.

as described Sect. 4.3. D'Aquin and Noy [47] provided the latest review of ontology libraries in 2012. Naskar and Dutta [8] reviews how some ontology libraries use ontology meta-data vocabularies.

## 4 Analysis of Current Ontology Metadata Practices

This analysis was made following three approaches: (1) we have reviewed the most standard and relevant metadata vocabularies available (23 totals) to select properties to describe ontologies; (2) we have reviewed how are these vocabularies used within 805 selected ontologies from known ontology libraries; (3) we have studied some of the most common ontology repositories available in the semantic web community to capture how they are dealing with ontology metadata and to which extent they rely on standard vocabularies.

### 4.1 Analysis of Existing Metadata Vocabularies to Describe Ontologies or Other General Resources

In the following, we describe the vocabularies that to some extent have been proposed to describe metadata about ontologies. It includes first of all the W3C Recommendations available to describe semantic resources: *Resource Description Framework Schema* (RDFS), *Web Ontology Language* (OWL) and *Simple Knowledge Organization System* (SKOS). Then the *Ontology Metadata Vocabulary* (OMV) [44] produced in the context of several EU projects and published in 2005. OMV (v.2.4.1) consists of 15 classes, 33 object properties, and 29 data properties. Unfortunately, the initiative stopped in 2007. Under the latest OMV version (2.4.1), two physically separated modules are proposed: OMV Core (provide the relevant metadata to support the ontology reuse settings) and OMV Extensions (to allow ontology developers and users to specify task- or application-specific ontology-related information). One limitation of OMV was not to be aligned to (or reuse) standard vocabularies at that time. This limitation has been recently partially addressed by a work published end of 2015: the *Metadata for Ontology Description* (now referred as MOD1.0) [7] which is similar to OMV (without using it). It has been designed as an ontology consisting of 15 classes (mod:Ontology+10 others+4 from FOAF), 18 object properties and 33 data properties among 7 of them were not included in OMV. For naming the metadata elements, it has reused existing properties from SKOS, FOAF, DC and DCT. Despite of the seven new properties, MOD1.0 still misses numerous relevant properties as we will see later. In Sect. 7.1, we describe

our new join work on MOD1.2 [48] done consequently to the work presented here.

In 2005, the quite simple but relevant *Vocabulary for annotating vocabulary descriptions* (VANN) was made available and quite used since then. In 2009, the *Descriptive Ontology of Ontology Relations* (DOOR) [49] has been published but never really used outside of the NeON project. It was a very formal vocabulary that described precisely and in a logical manner 32 relations between ontologies organized in a formal hierarchy. DOOR did incorporate the ontologies relations offered by OWL. More recently, the *Vocabulary of a Friend* (VOAF) [50] was created to “describe vocabularies (RDFS vocabularies or OWL ontologies) used in the Linked Data Cloud. In particular, it provides properties expressing the different ways such vocabularies can rely on, extend, specify, annotate or otherwise link to each other. It relies itself on DC and VOID.” Although VOAF was developed to capture relations between ontologies, it makes no use or reference to OWL or DOOR (with which it captures similar properties). In 2014, the NKOS working group of the Dublin Core proposed the *NKOS Application Profile* (<http://nkos.sli.s.kent.edu/nkos-ap.html>) which introduces 6 new properties and reused 22 properties from other vocabularies. [51] published a study made a few years ago to identify the relevant terminology metadata models that could form the foundation for a standard ontology profile for use by the NCI (National Cancer Institute), NCBO (National Center for Biomedical Ontology), and NCRI (National Cancer Research Institute, UK) community. This community effort on identifying the useless or ambiguous element from OMV proposed a few small changes but went no further.<sup>6</sup>

Ontologies share some characteristics with web datasets or data catalogs. Indeed, in the semantic web vision, ontologies are themselves sets of RDF triplets. We thus argue that some properties that have been defined to describe web datasets are relevant to ontologies also. Among the recent work to describe “datasets,” there are: the *Vocabulary of Interlinked Datasets* (VOID) [52], a W3C Note proposed in 2011 which can be used “to express general metadata based on DC, access metadata, structural metadata, and links between datasets.” VOID allows to describe two main objects void:Dataset and void:Linkset which are sets of links between datasets. The vocabulary also includes URIs for license or serialization formats. *Identifiers.org* (IDOT) [53] is a small vocabulary intended to “referencing of data for the scientific community, with a current focus on the Life Sciences domain.” It was developed by the European Bioinformatics Institute to specify, among other things, URI

<sup>6</sup> Some elements are removed (e.g., omv:hasPriorVersion), some element are renamed (e.g., name to fullName, acronym to shortName), some class definitions are modified and two new elements namely, certifiedBy, mandatedBy, are added into the revised set.

patterns. The *Data Catalog Vocabulary* (DCAT), which is the most recent W3C Recommendation for metadata (and uses DCT) and its profile, *Asset Description Metadata Schema* (ADMS), used to describe semantic assets (data models, code lists, taxonomies, dictionaries, vocabularies) created by the EU's Interoperability Solutions for European Public Administrations (ISA). Finally, *Schema.org* has been proposed and adopted in 2011 by Google, Bing and Yahoo! and do include a dataset class.

To describe other kinds of resources, one will find the following vocabularies: *Friend of a Friend Vocabulary* (FOAF) or *Description of a Project* (DOAP) to describe documents and projects. The *Creative Commons Rights Expression Language* (CC) for licensed work. *SPARQL 1.1 Service Description* (SD) for describing SPARQL endpoints. And the *Provenance Ontology* (PROV) and *Provenance, Authoring and Versioning* (PAV) for describing provenance (PAV specializes terms from PROV and DCT). Finally, the *OboInOwl specification* [54] converts OBO ontology header properties to OWL. This is not a standard but some of these properties are handled by the OBO Edit ontology editor and therefore often used.

Other vocabularies recently published or under development, from which we have not selected any properties in our ontology repository metadata model include *Extension to the VOID* [55], which is an extension of VOID mainly for partitions and statistical descriptions. *Citation Typing Ontology* (CiTO) describes citations between entities (one property only is actually relevant for us). The *Protocol for Web Description Resources* (POWDER) provides a mechanism to describe and discover web resources. The *DDI-RDF Discovery Vocabulary* (DISCO) which is a vocabulary to describe studies. The *Information Artifact Ontology* (IAO) [56], which was defined for representation of types of information content entities such as documents, databases, and digital images. The *Semanticscience Integrated Ontology* (SIO) [57] which describes many different types of informational entities and relations between them. [58] have proposed a metadata vocabulary for the Lemon model [59] called *Linguistic Metadata* (LIME) for describing linguistic resources and linguistically enriched datasets. Finally, we must also mention the document ISO/IEC 19763-3 (Metamodel framework for interoperability (MFI)—Part 3: metamodel for ontology registration) which latest version is from 2010 and is not public.

Table 1 summarizes and compares these vocabularies. This review of existing metadata vocabularies (and our work presented in Sect. 5.2) clearly shows no existing vocabularies really cover enough aspects of ontologies to be used solely and despite a few exceptions, metadata vocabularies do not rely on one another. Plus, there is a strong overlap in all the vocabularies studied which redefine things that have already been described several times before (such as dates for

which 25 properties are available). When dealing with harmonized metadata in the context of, for instance, an ontology repository, there exists an obvious technical and semantics challenge: being able to process ontologies that could have been described with one or several of those metadata vocabularies. Plus, many of the vocabularies do not support dereferenceability making impossible for the machine to automatically access the semantic description of the properties (e.g., domain, range) defined within the vocabulary. The fact of having multiple vocabularies for describing ontologies (or any other thing) should not be an issue: redundancies on one side enables specificity on the other side. However, in the semantic web vision, we would expect vocabularies to match and rely on one another more. To address our need of properly defining ontologies in an ontology repository, this review gave us a list of candidate metadata properties. In Sect. 5, we will present how we have built a list of properties for AgroPortal's new metadata model based on the studied vocabularies. In Sect. 7.1, we will discuss the need for metadata authoring guidelines and for harmonization of existing metadata vocabularies beyond the AgroPortal project.

## 4.2 Analysis of Current Use of Ontology Metadata Vocabularies

To get a sense of the quantity and origin of existing metadata vocabularies actually used by ontology developers, we downloaded and semi-automatically analyzed 1107 OWL ontologies taken from different sources: 594 from NCBO BioPortal, 53 from AgroPortal, 260 from MMI Ontology Registry and Repository, 97 from the OBO Foundry, 82 from DERI Vocabularies, and 21 from ProtégéWiki.<sup>7</sup> Once ontology duplicates removed—by matching name or base URIs—we obtained a corpus of 805 ontologies. Because of the sources of the ontologies, this corpus is slightly influenced by certain domains (biomedicine, biology, agronomy, environment); although it might bias the results, we are still confident they are quite representative, especially in these domains. We provide here the result of the analyzed ontologies.

We found 128 ontologies (16%) without any description or annotation. For rest of the 677 ontologies (84%), the number of properties used in describing the ontologies is ranging from 1 to 32. For instance, out of the 53 ontologies retrieved from AgroPortal, there are two ontologies having only one metadata. Overall, there are 354 ontologies (44%) for which ten or more properties (and maximum 32) are observed. For rest of the 323 ontologies (40%), the number of metadata per ontology is below 10.

<sup>7</sup> It is important to understand that we have looked at the metadata in the original ontology file, not the metadata captured by BioPortal or AgroPortal in their internal model.

**Table 1** Comparison of reviewed metadata vocabularies

Prefix	Name	Year (version)	Rely on other vocabularies	D	R	Comments
adms	Asset Description Metadata Schema	2013	dc, dcat, foaf, schema + vCard	Y	N	Profile of DCAT. Created by EU's ISA body to help standards publishers
cc	Creative Commons Rights Expression Language	2008		Y		Used to describe copyright licenses in RDF
dc	Dublin Core Elements	2012	–	Y	R	The “original” Dublin Core set of 15 classic metadata terms
dcat	Data Catalog Vocabulary	2014	dc, foaf, vcard	Y	R	W3C Recommendation for data catalog
dct	DCMI Metadata Terms	2012	–	Y	R	An up-to-date specification of all metadata terms maintained by the DCMI
doap	Description of a Project	2012	foaf	Y		Vocabulary to describe software projects
door	Descriptive Ontology of Ontology Relations	2009	–	N		Very formal ontology relation ontology
foaf	Friend of a Friend Vocabulary	2014 (v0.99)	–	Y	N	Linking people and information on the Web. Used as a reference by multiple vocabularies
idot	Identifiers.org	2018	–	Y		Provides stable and perennial identifiers for data records used in the Life Sciences
mod	Metadata for Ontology Description & Publication 1.0	2017 (v1.2)	owl, rdfs, dct, foaf, skos, omv, vann, pav, prov, sd, doap	N		Ontology designed specially to describe ontologies, extension of OMV mainly, but relies on many other metadata vocabularies. Work inspired by our work on AgroPortal
nkos	Networked Knowledge Organization Systems Application Profile	2015 (v0.2)	dc, adms, dcat, prov + frbrer, frsrd, wdrs	Y		NKOS is a Dublin Core Application Profile for describing knowledge organization systems
oboInOwl	OboInOwl Mappings	2011 (v1.2)	–	N		A namespace created when transforming OBO ontologies to OWL
omv	Ontology Metadata Vocabulary	2009 (v2.4.1)	–	N		Ontology especially created to describe ontologies. Partially adopted by ontology libraries
owl	OWL 2 Web Ontology Language	2012 (v2)	–	Y	R	W3C Recommendation to create ontologies. Offer a few properties to describe them also
pav	Provenance, Authoring and Versioning	2015 (v 2.3.1)	dc, prov	Y		Lightweight ontology specializing prov to describe provenance
prov	Provenance Ontology	2013	–	Y	R	W3C Recommendation for describing provenance metadata
rdfs	RDF Schema	2014 (v1.1)	–	Y	R	W3C Recommendation for describing any RDF resource
schema	Schema.org	2017 (v3.3)	–	Y		Google, Yahoo!, Bing agreed metadata standard for Web objects
sd	SPARQL 1.1 Service Description	2013	–	Y	R	W3C Recommendation for describing SPARQL services
skos	Simple Knowledge Organization System	2009	–	Y	R	W3C Recommendation for describing thesauri, terminologies, vocabularies
vann	Vocabulary for annotating vocabulary descriptions	2005	–	Y		Lightweight vocabulary for annotating descriptions of vocabularies
voaf	Vocabulary of a Friend	2013 (v2.3)	dc, void	Y		Vocabulary to describe vocabularies and their relations
void	Vocabulary of Interlinked Datasets	2011	dc, foaf	Y	N	Widely adopted vocabulary to describe datasets and their relations

D column states if property URIs are dereferenceable (Y or N); R column states if it is a W3C or Dublin Core Recommendation (R), note (N), or none of the two (blank)

We have also observed in total 30 metadata vocabularies that are being used to describe the ontologies. The 19 most frequently used ones are exemplified in Table 2. Notice that among these, around 1/3 of them are W3C or Dublin Core recommended vocabularies. The rest of vocabularies forms the long tail of the curve of the used metadata vocabularies with a couple of uses or mostly only one. They include recommended standards (e.g., Schema.org), community standards (e.g., CITO, ADMS, DOAP) or very specific vocabularies (e.g., PRISM, EFO, IRON). Some other findings of this study are:

- Most of all these 30 vocabularies are general in purpose. Some metadata vocabularies, which were specially proposed with the purpose of annotating/describing ontologies (e.g., VOID, VOA, DOOR), are mostly absent or barely used, with the exception of OMV which is not surprisingly among the most used vocabulary.
- However, the presence of OMV—and omvmmi complement to OMV—is mostly explained by the important number of ontologies taken from the MMI Ontology Registry and Repository that has adopted and enforced OMV in the ontologies hosted on their repository. In a previous similar study on 222 ontologies [48], which does not include MMI ontologies but included 61 ontologies randomly selected via Google, OMV was completely absent. This clearly illustrates the impact of harmonized community practices (or repository enforcement) on ontology metadata.
- Two vocabularies among the most used (oboInOwl and protege) are present because they are automatically included in ontologies by ontology development software.<sup>8</sup> Similarly, from Table 2 we can see that rdfs:comment, owl:versionInfo and owl:imports are among the most frequently used metadata elements. We think the reason for their frequent use is because of their ready availability in the ontology editors. For instance, a selected set of metadata elements from rdfs and owl are made readily available in Protégé annotation tab. We may assume most ontology developers find it handy when annotation properties are readily available in the ontology editor's annotation tab, rather than referring a vocabulary available on the Web but not in the editor. The case of owl:imports is slightly different. It is required for functional reasons to import ontologies.

<sup>8</sup> The *oboInOwl* namespace is used by the OBO2OWL converter when converting Open Biomedical Ontology format to OWL. The high frequency of this vocabulary is explained because half of our ontologies were selected from the NCBO BioPortal that contains many ontologies originally developed in OBO (often with the OBOEdit software). The *protege* name space was used in previous (~v3) of Protégé mostly to customize the user interface when displaying the ontology. It was not to describe the ontology.

- Multiple properties express the same information. For instance, in providing the name of the ontology, some have used dc:title while some other have used dct:title. Similarly, some people have used dct:license to provide the licensing information, while some others have used cc:license.
- There is a confusion between the use of DC and DCT as the latter includes and refines the 15 primary properties from the former. Some developers prefer to refer DC and some prefer DC Terms for the similar element. The reason could be the unavailability of a precise guideline on how and when to use the DC core and DCT elements. In the context of semantic web applications, although using DC is not incorrect, DCMI recommends using DCT that provides domain and range information for properties.<sup>9</sup>
- Some metadata elements are used in an improper way. For instance, skos:definition shall only be used to supply a complete explanation of the *intended meaning of a (SKOS) concept* as the other SKOS “documentation properties” and is not supposed to be used to described ontologies (unless an ontology is considered a concept).
- Generic properties such as rdfs:comment or dc:date are used instead of more specific ones such as respectively dc:description or dc:created/modified.
- The study also revealed 12 custom properties used to describe metadata (not reported in Table 2) declared in the main namespace of the ontology, e.g., primary\_author\_and\_curator, wasRevisionOf, contributing\_author. This may illustrate a not so good practice which consists in creating a new local property when in need.

We previously conducted a similar smaller study [48] and came to similar outcomes. Another one was conducted by Tejo-Alonso et al. [35]: Their study consisted of total 23 RDFS/OWL metadata vocabularies (the “most popular from prefix.cc”): They were especially interested in how much the metadata vocabularies are themselves described with proper metadata properties. The authors arrived at similar conclusions than us with our larger study: (1) rdfs/owl popularity; (2) dc/dct confusion; (3) frequency of auto-generated properties; (4) generic property over specific ones; (5) different properties for similar information.

Concerning the description of knowledge resources with metadata, we also like to mention an exceptional example found in the context of the AgroPortal project: Agrovoc, which is the reference multilingual thesaurus in agriculture developed by FAO, is explicitly and extensively defined by a so-called “VOID profile”<sup>10</sup> which lives aside from the main

<sup>9</sup> [http://wiki.dublincore.org/index.php/FAQ/DC\\_and\\_DCTERMS\\_Namespaces](http://wiki.dublincore.org/index.php/FAQ/DC_and_DCTERMS_Namespaces).

<sup>10</sup> <http://aims.fao.org/aos/agrovoc/void.ttl>.

**Table 2** Most frequent used vocabularies over a corpus of 805 ontologies

Prefix	Number	Properties used (number of times)
omv	2169	acronym (251), creationDate (251), description (251), hasCreator (251), name (251), uri (251), usedontologyengineeringtool (157), version (148), keywords (126), hasContributor (109), documentation (74), reference (49)
omvmmi	1697	creditRequired (251), origMaintainerCode (251), hasContentCreator (193), hasResourceType (186), shortNameuri (151), temporarymmirole (108), origvocManager (107), contactRole (106), contact (99), origvocuri (60), origvocDocumentationuri (40), creditCitation (38), origvocDescriptiveName (36), origvocSyntaxFormat (30), origvocKeywords (23), origvocVersionid (16), origvocLastModified (1)
dc	1599	creator (456), description (309), date (307), contributor (183), source (77), title (102), subject (47), format (31), license (28), publisher (21), rights (17), language (8), identifier (6), modified (3), coverage (2), issued (1), type (1)
dct	652	modified (86), title (85), created (84), partOf (81), status (81), type (81), description (62), publisher (60), creator (9), license (6), issued (3), subject (2), contributor (3), isreferencedby (3), identifier (1), isrequiredby (1), language (1), date (1), source (1), format (1)
owl	498	versionInfo (183), imports (210), versionIRI (74), priorVersion (22), ontology (4), incompatibleWith (3), backwardCompatibleWith (1), deprecated (1)
oboInOwl	283	default-namespace (54), hasOboFormatVersion (53), savedBy (49), date (47), auto-generated-by (40), namespaceIdRule (7), treat-xrefs-as-equivalent (5), hassubset (4), remark (4), treat-xrefs-as-is_a (4), treat-xrefs-as-genus-differentia (3), format-version (2), pairwise-disjoint (2), treat-xrefs-as-has-subclass (2), treat-xrefs-as-reverse-genus-differentia (2), comment (1), data-version (1), default-relationship-id-prefix (1), next-id (1), property-value (1)
rdfs	265	comment (174), label (68), seeAlso (16), isDefinedBy (7)
vann	166	preferredNamespacePrefix (83), preferredNamespaceUri (83)
foaf	102	homepage (91), mbox (6), page (4), isPrimaryTopicOf (2)
obo	33	iao_0000116 (10), idspace (4), date (3), default-relationship-id-prefix (3), format-version (2), remark (2), comment (1), iao_0000117 (1), iao_0000412 (1), definition (1), editorialNote (1), historyNote (1), imports (1), is_metadata_tag (1), license (1)
skos	19	altLabel (6), prefLabel (6), definition (5), changeNote (1)
protégé	19	defaultLanguage (19)
nemo_annot	10	created_date (2), curator (2), modified_date (2), pref_label (2), synonym (2)
vaem	9	dateCreated (1), hasAspectsCope (1), hasCatalogEntry (1), hasDisciplineScope (1), hasDomainScope (1), hasRole (1), lastUpdated (1), revisionNumber (1), usesNonImportedResource (1)
cc	4	license (4)
dcat	3	landingPage (2), downloadURL (1)
asthma	2	creator (1), defaultLanguage (1)
pav	2	version (2)
void	2	dataBrowse (1), dataDump (1)

For namespaces either see Table 1 or in some cases on <https://prefix.cc>

thesaurus file and uses 7 metadata vocabularies to describe Agrovoc with RDF statements.

This review helped us to decide which vocabulary and/or property shall be “prioritized” when selecting properties for our unique model in an ontology repository. The final step was then to look at how other ontology libraries were dealing with metadata.

### 4.3 Analysis of Metadata Representation Within Ontology Libraries

We have studied some of the most common ontology libraries and repositories available in the semantic web community, and especially the NCBO BioPortal, to analyze: (1) how they are dealing with ontology metadata; (2) to which extent they rely on previously analyzed metadata vocabularies. We have only been interested in the metadata that are “nonspecific” to the repository, i.e., specific fields required for implementation purposes were ignored.

We consider under the term libraries any kind of web tool (repository, registry or portal) that somehow focus on ontologies and/or vocabularies [45]. In particular, we have explicitly reviewed:

1. Repository or portals including the NCBO BioPortal [12], Ontobee [60], EBI Ontology Lookup Service [10], MMI Ontology Registry and Repository [11], the ESIP portal (based on NCBO technology), and AberOWL [61];
2. Registries or catalogs including the OKFN Linked Open Vocabularies [9], OBO Foundry [16], WebProtégé (<http://webprotege.stanford.edu>), Agrisemantics Map of Data Standards (<http://vest.agrisemantics.org>) [62], FAIR-Sharing (<https://fairsharing.org>) [40];
3. Web indexes such as Watson [63], Swoogle [1] (or Sindice.com, not reviewed because not accessible anymore).

We have reviewed the metadata properties used by all these libraries and considered them for our listing to be implemented in our portal. As later explained, we have used BioPortal as baseline. Each of the reviewed libraries uses, to some extent, some metadata fields but do not always use standard metadata vocabularies:

- NCBO BioPortal repository [12] uses 66 metadata properties that serves as the basis for our listing.<sup>11</sup> These properties are defined in an in-house vocabulary (here called BioPortal Metadata and identified with the namespace bpm) that is not formally described outside of

BioPortal but because of the portal adoption of JSON-LD, can be formally used.<sup>12</sup> For 10 properties, BioPortal reuses OMV names but redefines them in its own namespace (e.g., bpm:omvacronym). Other than the 10 OMV property names, BioPortal does not use any other metadata vocabulary. Over the 66 properties used by BioPortal, we have classified 46 (36 locally defined +10 from OMV) as nonspecific to the portal. BioPortal user interface (and web services) allows to edit most of the properties and some of them are automatically generated (e.g., metrics). Because they originally use the same source code, the situation is the same for ESIP portal and AgroPortal before our work.

- MMI Open Ontology Repository, which was originally also based on BioPortal code, did later embrace OMV more and added a few other metadata properties (omvmmi extension). The repository administrators do edit the ontology metadata of the files hosted on the portal to harmonize them.
- Linked Open Vocabulary registry [9] explicitly uses VOID and VOAf; the latter was actually created for this purpose. The LOV is a very good example of good use of harmonized metadata that has inspired us a lot. More than 600 vocabularies (as of May 2017) are described with common metadata fields facilitating manual and automatic search. In addition, LOV is not limited to VOAf and recommends the use of other standard vocabularies.<sup>13</sup> It is important to note that the metadata is either entered by the developer submitting the vocabulary then curated by the registry administrators. Some are also automatically generated and, in both cases, LOV always relies on standard vocabularies to store the information.
- OBO Foundry [16] refers metadata from around 20 vocabularies including DC, FOAF, IDOT, VOID, DOAP, DISCO, etc.<sup>14</sup> The OBO Foundry community effort is important, and they encourage the ontology developers to edit the metadata, aside from the main ontology file, in a specific document (in MD or YAML format) hosted on GitHub aside of the ontology files and parsed by the OBO Foundry application to display ontology descriptions.<sup>15</sup> OBO Foundry administrators manually curate/edit ontology metadata in complement of ontology developers.
- Ontobee [60] offers a few (6–7) common metadata (e.g., IRI, home, contact) and then display any other metadata

<sup>11</sup> <http://data.bioontology.org/documentation#OntologySubmission> and #Ontology.

<sup>12</sup> Originally, the NCBO developed the BioPortal Metadata Ontology (<http://purl.bioontology.org/ontology/BP-METADATA>) which imports OMV. But the current implementation is not completely in sync with this vocabulary anymore.

<sup>13</sup> [http://lov.okfn.org/Recommendations\\_Vocabulary\\_Design.pdf](http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf).

<sup>14</sup> <http://obofoundry.github.io/registry/context.jsonld>.

<sup>15</sup> For instance: <https://github.com/OBOFoundry/OBOFoundry.github.io/blob/master/ontology/envo.md>.

properties originally included in the ontology as “annotation properties.” The portal also counts a few metrics.

- Similarly, AberOWL [61] and OLS [10], have a few common properties and then display the rest (included in the ontology file) as annotation properties. By comparison to OBO Foundry, the common properties are not described with standard vocabularies.

For a recent review of ontology libraries and their metadata, the reader might refer to [8], briefly summarized in [7]. In these papers, authors showed that ontology metadata vocabularies are rarely used by ontology libraries: 4<sup>16</sup> ontology libraries over the 13 studied have partially used the OMV.

## 5 Building a List of Properties to Describe Ontologies

### 5.1 Method to Select Properties from Existing Vocabularies

Enlightened by the analysis presented in the previous section, we have accomplished a systematic review (as methodologically described by [64]) of the vocabularies previously identified with the following research question in mind: *Which existing properties could be used to describe ontologies?* The previously listed vocabularies have been identified from: (1) the semantic web literature; (2) investigating ontology libraries; (3) related similar studies such as the one for dataset by the HCLS working group. Vocabularies were selected based on their degree of standardization, relevance for ontologies and current usage by ontology developers. The final list of the 23 reviewed vocabularies and the numbers of property reused are available in Table 3, plus the NCBO BioPortal metadata model that we used as baseline and listed as a vocabulary with the prefix “bpm.”

We now describe selection criteria for properties to be used by our ontology portal. The goal of this list was to delimit the set of properties that our ontology repository will “parse,” i.e., the ones that will be automatically recognized and used to populate the unified ontology metadata model. Indeed, our motivation was to improve metadata management within AgroPortal, a portal based on the NCBO technology. For other important reasons in the AgroPortal project (maintenance, collaboration, support, interoperability), keeping our ontology repository backward compatible with NCBO was mandatory. Therefore, each time a property was already captured by the BioPortal model, we would add it to the list and not change it to another property that the analysis Sect. 4 would have shown more relevant. The criteria for inclusion were the following, considered by order of importance:

1. Relevance for describing an ontology—the property may have a sense if used to describe an ontology.
2. Being not “specific” to a library—even if the ontology library helps to populate or predict the property, the property would capture an information that belongs to the ontology. For instance, properties such as credentials on the portal or maintenance information, or local parsing status are considered “specific.”
3. Semantic consistency—there must not be any conflict (e.g., disjoint classes) if someone would describe an ontology with all the listed properties. For instance, an ontology may be an instance of `omv:Ontology`, `void:Dataset` and `cc:Work` at the same time.
4. Being a W3C or Dublin Core Recommendations.
5. The frequency of use in the study presented in Sect. 4.2.
6. Priority to vocabularies specific for ontologies rather than to the ones specialized for more general object (`cc:Work`, `dcat:DataSet`, `sd:Service`, etc.).

Although we agree dereferenceability is an important criterion for a vocabulary, we have not excluded properties that are not dereferenceable, even it means a machine would hardly understand the semantics of the property. We will mention this as a requirement for a future ontology metadata vocabulary in Sect. 7.1.

### 5.2 Properties Selected from Existing Metadata Vocabularies

For each of these vocabularies, we have selected the significant properties to describe objects that an ontology could be considered a certain type of, e.g., a dataset, an asset, a project or a document. For instance, an ontology may be seen as a `prov:Entity` object and then the property `prov:wasGeneratedBy` may then be used to describe its provenance. We illustrate with examples as often as possible.

The first things to look at are the properties available in the W3C standard vocabularies, such as RDFS, OWL, and SKOS. Indeed, they include some annotation properties that we can use to describe ontologies if we consider them instances of `rdfs:Resource`, `owl:Ontology` or `skos:conceptScheme`.

---

`rdfs:label`, `rdfs:seeAlso`, `rdfs:comment`, `owl:versionInfo`,  
`owl:versionIRI`, `owl:imports`, `owl:priorVersion`,  
`owl:backwardCompatibleWith`, `owl:incompatibleWith`,  
`owl:deprecated`, `skos:prefLabel`, `skos:altLabel`,  
`skos:hiddenLabel`, `skos:hasTopConcept`, `skos:notation`

---

SKOS label properties can be used to denote the alternative or non-conventional names of an ontology. For instance, the Phenotype And Trait Ontology is also known as “PATO,”

<sup>16</sup> The study reported 3 only, but the case of MMI was a mistake.

**Table 3** Vocabularies studied in this review + BioPortal

Prefix	Namespace	Resource	#T	#S	#U
adms	<a href="http://www.w3.org/ns/adms#">http://www.w3.org/ns/adms#</a>	adms:Asset	13	11	0
cc	<a href="http://creativecommons.org/ns#">http://creativecommons.org/ns#</a>	cc:Work	5	5	2
dc	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>	NA	15	15	0
dcat	<a href="http://www.w3.org/ns/dcat#">http://www.w3.org/ns/dcat#</a>	dcat:Dataset	5	4	0
dct	<a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a>	dcmi:Dataset, dcmi:Collection	55	38	13
doap	<a href="http://usefulinc.com/ns/doap#">http://usefulinc.com/ns/doap#</a>	doap:Project	25	18	3
door	<a href="http://kannel.open.ac.uk/ontology#">http://kannel.open.ac.uk/ontology#</a>	owl:Ontology	32	11	6
foaf	<a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>	foaf:Document	11	10	4
idot	<a href="http://identifiers.org/idot/">http://identifiers.org/idot/</a>	dct:Dataset	9	6	1
mod	<a href="http://www.isibang.ac.in/ns/mod#">http://www.isibang.ac.in/ns/mod#</a>	mod:Ontology	27	26	1
nkos	<a href="http://w3id.org/nkos#">http://w3id.org/nkos#</a>	rdfs:Resource	6	4	0
oboInOwl	<a href="http://www.geneontology.org/formats/oboInOwl#">http://www.geneontology.org/formats/oboInOwl#</a>	owl:Ontology	13	9	0
omv	<a href="http://omv.ontoware.org/2005/05/ontology#">http://omv.ontoware.org/2005/05/ontology#</a>	omv:Ontology	37	37	35
owl	<a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a>	owl:Ontology	11	7	2
pav	<a href="http://purl.org/pav/">http://purl.org/pav/</a>	prov:Entity	30	16	2
prov	<a href="http://www.w3.org/ns/prov#">http://www.w3.org/ns/prov#</a>	prov:Entity	22	10	2
rdfs	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>	rdfs:Resource	7	3	0
schema	<a href="http://schema.org/">http://schema.org/</a>	schema:Dataset	90	41	7
sd	<a href="http://www.w3.org/ns/sparql-service-description#">http://www.w3.org/ns/sparql-service-description#</a>	sd:Service	13	1	1
skos	<a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a>	skos:conceptScheme	14	5	1
vann	<a href="http://purl.org/vocab/vann/">http://purl.org/vocab/vann/</a>	rdfs:Resource	6	5	3
voaf	<a href="http://purl.org/vocommons/voaf#">http://purl.org/vocommons/voaf#</a>	voaf:Vocabulary	16	12	5
void	<a href="http://rdfs.org/ns/void#">http://rdfs.org/ns/void#</a>	void:Dataset	24	16	5
bpm	<a href="http://data.bioontology.org/metadata">http://data.bioontology.org/metadata</a>	bpm:Ontology bpm:OntologySubmission	36	36	34
Total			522	346	127

Column #T is the total number of properties provided by the vocabulary for column Resource type (or rdfs:Resource). Column #S is the number of properties *selected* in the list from this vocabulary (only vocabularies within the same namespace). Column #U is the number of properties *used* as default property in the implementation of the new ontology repository model. For instance, for foaf:Document, we have reviewed a total of 11 properties and considered 10 of them were relevant to describe ontologies and are now parsed by AgroPortal, but only 4 have been explicitly used as “default” property in the new model

“Phenotypic Quality Ontology,” or “Ontology of phenotypic qualities.”

Then the Dublin Core Metadata Initiative standards are available. Dublin Core does not always specify the domain of its properties. We have assumed that all of them accept rdfs:Resource as domain. We have included the 15 DC properties and 38 DCT properties that are relevant for describing ontologies (only DCT is listed hereafter):

dct:title, dct:accessRights, dct:isPartOf, dct:hasVersion, dct:bibliographicCitation, dct:language, dct:dateSubmitted, dct:description, dct:created, dct:date, dct:issued, dct:rightsHolder, dct:modified, dct:conformsTo, dct:contributor, dct:creator, dct:subject, dct:rights, dct:license, dct:format, dct:type, dct:requires, dct:isVersionOf, dct:relation, dct:coverage, dct:publisher, dct:identifier, dct:source, dct:abstract, dct:alternative, dct:hasPart, dct:isFormatOf, dct:hasFormat, dct:audience, dct:valid, dct:accrualMethod, dct:accrualPeriodicity, dct:accrualPolicy

DCT’s accrual properties can be used for instance to describe the process by which an ontology is updated and

new concepts are added or removed. This has been established as an important aspect by the *Minimum Information for Reporting of an Ontology* guidelines.

Among the vocabularies available for ontologies we have taken all the properties from OMV and MOD<sup>17</sup> considering an ontology an instance of `omv:Ontology` and `mod:Ontology`. We only list the ones in OMV namespace (when they are named the same in MOD):

---

`omv:acronym`, `omv:name`, `omv:hasOntologyLanguage`,  
`omv:reference`, `omv:URI`, `omv:naturalLanguage`, `omv:documentation`,  
`omv:version`, `omv:creationDate`, `omv:description`, `omv:status`,  
`omv:resourceLocator`, `omv:numberOfClasses`,  
`omv:numberOfIndividuals`, `omv:numberOfProperties`,  
`omv:modificationDate`, `omv:numberOfAxioms`, `omv:keyClasses`,  
`omv:keywords`, `omv:knownUsage`,  
`omv:conformsToKnowledgeRepresentationParadigm`,  
`omv:hasContributor`, `omv:hasCreator`, `omv:designedForOntologyTask`,  
`omv:endorsedBy`, `omv:hasDomain`, `omv:hasFormalityLevel`,  
`omv:hasLicense`, `omv:hasOntologySyntax`, `omv:isOfType`,  
`omv:usedOntologyEngineeringMethodology`, `omv:notes`,  
`omv:usedOntologyEngineeringTool`, `omv:useImports`,  
`omv:hasPriorVersion`, `omv:isBackwardCompatibleWith`,  
`omv:isIncompatibleWith`, `mod:accessibility`, `mod:module`,  
`mod:ontologyInUse`, `mod:sponsoredBy`, `mod:competencyQuestion`,  
`mod:vocabularyUsed`, `mod:homepage`

---

OMV properties (and individuals) are particularly relevant as they have been explicitly created to describe ontologies. They are the only ones in our study enabling to capture information such as the methodology applied to create the ontology or the task/role for which an ontology has been designed. For instance, the Medical Subject Headings (MeSH) terminology has been designed for indexing scientific medical publications (`omv:IndexingTask`), which is different from the Gene Ontology that has been developed to annotate gene products (`omv:AnnotationTask`). Among the new properties from MOD, `mod:competencyQuestion` corresponds to properties suggested for instance by [65]

There exist two specific vocabularies for representing relations. From DOOR, that is very detailed and formal, we have selected 11 of the most significant, in addition to the 4 from OWL. We had to draw the line, and we considered 15 formal relations from these two vocabularies were enough in most cases to describe ontology relations. VOAF properties (applied to a `voaf:Vocabulary`) were almost completely included, except 4 statistical properties (that are relevant only for a specific repository):

---

`door:semanticallyIncludedIn`, `door:imports`, `door:priorVersion`,  
`door:backwardCompatibleWith`, `door:owlIncompatibleWith`,  
`door:ontologyRelatedTo`, `door:similarTo`,  
`door:comesFromTheSameDomain`, `door:isAlignedTo`,  
`door:explanationEvolution`, `door:hasDisparateModelling`,  
`voaf:classNumber`, `voaf:propertyNumber`, `voaf:extends`, `voaf:reliesOn`,  
`voaf:similar`, `voaf:hasEquivalencesWith`, `voaf:specializes`,  
`voaf:usedBy`, `voaf:metadataVoc`, `voaf:generalizes`,  
`voaf:hasDisjunctionsWith`, `voaf:toDoList`

---

The property `door:explanationEvolution` or `voaf:specializes` can be used to say that an ontology is a latter version that is semantically equivalent to another ontology and specializes it. For instance, International Classification of Diseases, 10th revision (ICD-10) has for prior version ICD-9 and for specialization ICD-10-CM (Clinical Modification made by US National Center for Health Statistics).

From NKOS Application Profile, we have selected 4 properties among the 6 new ones defined in the namespace and have in that case considered the properties would be applied to `rdfs:Resource`. Two have been excluded because we already have more precise properties in other vocabularies (`nkos:serviceOffered` and `nkos:sizeNote`).

---

`nkos:alignedWith`, `nkos:basedOn`, `nkos:updateFrequency`, `nkos:usedBy`

---

Among the metadata vocabularies to describe datasets, we have reviewed VOID, a W3C Note proposed in 2011 to describe RDF datasets. It allows describing two main objects `void:Dataset` and `void:Linkset` which are set of links between datasets. The vocabulary also includes URIs for license or serialization formats. `void:Dataset` can be described with 24 properties including a few metrics plus some from DCT. From VOID, we picked-up 16 relevant properties.

---

`void:subset`, `void:classPartition`, `void:propertyPartition`,  
`void:rootResource`, `void:classes`, `void:properties`, `void:triples`,  
`void:entities`, `void:exampleResource`, `void:vocabulary`,  
`void:sparqlEndpoint`, `void:dataDump`, `void:openSearchDescription`,  
`void:uriLookupEndpoint`, `void:uriRegexPattern`, `void:uriSpace`

---

For instance, `void:uriRegexPattern` may be used to explain the pattern that some ontologies use when building their URIs and concept identifiers, e.g., (ICD-10)'s codes respect a structure that keeps track of the chapter, and hierarchy (K70.3 code for "Alcoholic cirrhosis of liver" is the 3rd of "Alcoholic liver disease" (K70) which are all in the "Diseases of the digestive system" Chapter (K)).

<sup>17</sup> Except `mod:size` that was new in MOD and ambiguous ("the size of an ontology").

A few of the properties from Indentifiers.org (IDOT) (6) shall be relevant to describe ontologies also:

---

idot:state, idot:obsolete, idot:alternatePrefix, idot:identifierPattern, idot:preferredPrefix, idot:exampleIdentifier

---

VANN is a small vocabulary created to describe vocabularies, which includes:

---

vann:preferredNamespacePrefix, vann:preferredNamespaceUri, vann:usageNote, vann:example, vann:changes

---

The property idot:preferredPrefix or vann:preferredNamespacePrefix can be used to store the preferred prefix when using the ontologies. See for example, <http://prefix.cc> for all possible prefix values.

DCAT is the W3C Recommendation since January 2014 to describe data catalogs; it offers a dcat:Dataset class relevant for ontologies. DCAT uses DCT and also offers properties with domain dcat:Distribution, but we have not taken those ones to restrict our selection to the dcat:Dataset class (among the 4 missed properties, 3 finds equivalent in other vocabularies). Then from ADMS, which is a profile of DCAT used to describe semantic assets (data models, code lists, taxonomies, dictionaries, vocabularies), we took 19 properties for class adms:Asset (or no domain) but only 11 specifically defined in the adms namespace, because ADMS used several other vocabularies treated in this study:

---

dcat:landingPage, dcat:contactPoint, dcat:keyword, dcat:theme, adms:sample, adms:status, adms:versionNotes, adms:representationTechnique, adms:prev, adms:last, adms:next, adms:includedAsset, adms:identifier, adms:supportedSchema, adms:translation

---

In the SIFR BioPortal project [20], we are interested to formally represent that some ontologies are the translated version of other ones (usually stored in the NCBO BioPortal). For instance, the French Medical Dictionary for Regulatory Activities Terminology is translated from the English version. The adms:translation can be used for this.

Schema.org (SCHEMA) can describe multiple types of resources. We have identified the schema:Dataset type as the closest one to describe ontologies. Schema.org is very rich to describe schema:Dataset (including properties inherited of schema:CreativeWork and schema:Thing), we have identified 41 relevant properties:

---

schema:distribution, schema:includedInDataCatalog, schema:spatial, schema:about, schema:alternativeHeadline, schema:associatedMedia, schema:audience, schema:author, schema:award, schema:comments, schema:contributor, schema:copyrightHolder, schema:creator, schema:dateCreated, schema:dateModified, schema:datePublished, schema:workExample, schema:fileFormat, schema:hasPart, schema:isPartOf, schema:inLanguage, schema:isBasedOn, schema:keywords, schema:license, schema:mainEntity, schema:publisher, schema:publishingPrinciples, schema:review, schema:schemaVersion, schema:sourceOrganization, schema:translator, schema:version, schema:alternateName, schema:description, schema:image, schema:mainEntityOfPage, schema:citation, schema:name, schema:url, schema:translationOfWork, schema:translation

---

For instance, the property schema:includedInDataCatalog may be used to store the fact that an ontology is hosted in different ontology libraries. This is, for instance, the cases for the OBO Foundry ontologies that are, in addition of the foundry being uploaded in NCBO BioPortal, Ontobee, OLS and AberOWL. With such a property properly populated, everyone will always know in which library to find an ontology.

If we consider an ontology as different kinds of objects, additional relevant vocabularies may be used. Thus, FOAF can be used to describe an ontology as an instance of foaf:Document, DOAP if an ontology is viewed as development project (doap:Project) and CC to see it as a cc:Work<sup>18</sup>:

---

foaf:name, foaf:homepage, foaf:isPrimaryTopicOf, foaf:page, foaf:primaryTopic, foaf:maker, foaf:topic, foaf:depiction foaf:logo, foaf:fundedBy, doap:name, doap:blog, doap:language, doap:wiki, doap:release, doap:description, doap:created, doap:download-page, doap:helper, doap:maintainer, doap:translator, doap:audience, doap:download-mirror, doap:service-endpoint, doap:screenshots, doap:repository, doap:bug-database, doap:mailing-list, cc:attributionName, cc:attributionURL, cc:license, cc:morePermissions, cc:useGuidelines

---

More and more ontology developers have turned to GitHub to store and release their ontologies, for example, the Environment Ontology (<https://github.com/EnvironmentOntology>). The DOAP properties are thus very relevant to capture the metadata about the ontology development project.

Two vocabularies for representing provenance information are included: PROV and PAV. PAV specializes terms from PROV and DCT. It contains 40 properties (including 30 specific ones) with no constraint on range or domain. When incorporating PROV and PAV, we had to focus on the main properties offered to describe prov:Entity (but potentially more maybe used):

<sup>18</sup> We have here an inconsistency as doap:Project are themselves subclasses of foaf:Project and because foaf:Project and foaf:Document are disjoint. We let to ontology developers the choice.

---

prov:generalizationOf, prov:generatedAtTime, prov:wasAttributedTo, prov:wasInfluencedBy, prov:wasDerivedFrom, prov:wasRevisionOf, prov:specializationOf, prov:invalidatedAtTime, prov:wasGeneratedBy, prov:wasInvalidatedBy, pav:hasCurrentVersion, pav:hasVersion, pav:version, pav:createdOn, pav:authoredOn, pav:contributedOn, pav:lastUpdateOn, pav:contributedBy, pav:authoredBy, pav:createdBy, pav:createdWith, pav:previousVersion, pav:hasEarlierVersion, pav:derivedFrom, pav:curatedBy, pav:curatedOn

---

From the OboInOwl specification, we took 9 of the 13 properties (and the alternative names, not listed, e.g., savedBy):

---

oboInOwl:format-version, oboInOwl:data-version, oboInOwl:date, oboInOwl:saved-by, oboInOwl:auto-generated-by, oboInOwl:import, oboInOwl:synonymtypedef, oboInOwl:default-namespaces, oboInOwl:remark

---

Finally, we have selected sd:endpoint from SPARQL 1.1 Service Description.

### 5.3 Existing Properties in Ontology Repositories

In order to manage versioning, access rights and metadata, BioPortal model stores ontologies with two objects: one Ontology which is actually the shell for multiple Submissions that contains the real content of an ontology. The Ontology object contains the most usual metadata (name, acronym, administrators, viewing restriction, group and categories) that will remain over versions, whereas the Submission objects contain the detailed metadata (description, metrics, contact, etc.) and links to the actual content of that specific version. For example, the following REST service calls will return, respectively, the Ontology object and the latest Submission for the NCI Thesaurus:

<http://data.bioontology.org/ontologies/NCIT?display=all>  
[http://data.bioontology.org/ontologies/NCIT/latest\\_submission?display=all](http://data.bioontology.org/ontologies/NCIT/latest_submission?display=all)

We have reviewed the complete list of properties offered by those two objects (including direct properties and links returned by the API): 25 for Ontology and 41 for Submission. From them, we picked-up the ones (46) that are not specific to BioPortal. For instance, the administrator (different from contact) of an ontology in BioPortal is an information that has sense only within BioPortal and therefore does not belong to the original ontology.

For homogeneity, we use the namespace bpm in the following list, even if those properties do not actually belong to a formal vocabulary (we do not include hereafter the 10 OMV properties originally used by BioPortal):

---

bpm:group, bpm:viewOf, bpm:submissions, bpm:reviews, bpm:notes, bpm:projects, bpm:views, bpm:analytics, bpm:ui, bpm:properties, bpm:classes, bpm:roots, bpm:prefLabelProperty, bpm:definitionProperty, bpm:synonymProperty, bpm:authorProperty, bpm:hierarchyProperty, bpm:obsoleteProperty, bpm:obsoleteParent, bpm:homepage, bpm:publication, bpm:released, bpm:diffFilePath, bpm:pullLocation, bpm:contact, bpm:metrics.classes, bpm:metrics.individuals, bpm:metrics.properties, bpm:metrics.maxDepth, bpm:metrics.maxChildCount, bpm:metrics.averageChildCount, bpm:metrics.classesWithOneChild, bpm:metrics.classesWithMoreThan25Children, bpm:metrics.classesWithNoDefinition, bpm:downloadRdf, bpm:downloadCsv

---

Once a primary version of the list was created from BioPortal plus the standard metadata vocabularies, we also analyzed the other ontology repositories. We did not find other properties that were not already covered by our review so far. From the OBO Foundry, the only exceptions were the properties inside the obofmd namespace (non-dereferenceable), that seems to be the ones the OBO Foundry developers did not find in any vocabulary. Although we have matches for 4 over 5 of these properties, we did not integrate those by the lack of information about them (plus this namespace was not identified in Sect. 4.2). AberOWL contains also a property species that we did not pick up as this is specific to the biomedical domain and unsatisfiable classes which are an interesting information for the ontology evaluation, but not for ontology description. OLS contains also two properties that we do not already had (reasonerType and oboSlims) but were not included by the lack of information. Even if we have an interest in biological and agronomical ontologies, we did not include in this list, properties that are domain specific. All the properties can be used to describe ontologies from any domain.

### 5.4 Results: A Complete List of Properties to Describe Ontologies and a Unified Model for AgroPortal

After the two steps described in the previous section, we end up with a complete list of 346 properties that could be used to describe ontologies. These properties will, therefore, be parsed by AgroPortal when an ontology is uploaded in order to populate the values of unified model implemented for all the ontologies on the portal. With the 346 properties of this list, we cover most of the properties identified in Table 2 except the ones in namespaces that are not relevant for ontologies (e.g., nemo\_annot, vaem and asthma), portal specific (e.g., omvmmi), format specific or not defined as a vocabulary (e.g., obo), or software specific (e.g., protege) or within the oboInOwl namespace but not in the OBO in OWL specification [54]. Among the 31 properties from Tejo-Alonso et al.'s study [35], we cover 25 properties. The six proper-

ties not included are 4 SKOS “documentation properties” (e.g., `skos:changeNote`, `skos:definition`), that according to the SKOS specification are intended to provide information relating to concepts although there is no domain restriction for these properties. The two others are `rdfs:isDefinedBy`<sup>19</sup> and `vs:terms_status` excluded for an equivalent reason. We, therefore, believe our complete list of properties that will be parsed by our ontology repository include most of the properties actually used by ontology developers.

Among those properties of the complete list, there was obvious overlap. Indeed, some properties define exactly the same thing, e.g., the version information of an ontology can be described by `omv:version`, `owl:versionInfo`, `mod:version`, `doap:release`, `pav:version` and `schema:version`. And some properties define very similar things such as for instance the homepage of an ontology project: `bpm:homepage`, `foaf:homepage`, `cc:attributionURL`, `mod:homepage`, `doap:blog`, and `schema:mainEntityOfPage`. With the purpose of simplifying our list, and implement a restricted unified model within our ontology repository, we have grouped properties of exact or similar meaning by selecting a “default” property that we would use in our ontology metadata model. The role of these equivalences (we voluntarily do not use the word mapping or alignment) is not to build a unique vocabulary for describing ontologies (although this question will be discussed in Sect. 7), but to implement an unified model for describing ontologies in an ontology repository that would help us address the challenges explained in Sects. 1 and 2. When selecting the “default” property, we applied the following rules that are specific to our context:

1. Do not change the properties that were already in BioPortal. As previously explained, we had to keep AgroPortal backward compatible with BioPortal (we will further discuss this in Sect. 7). Except for 3 metric properties that we have duplicated to enable users to reset themselves the number of classes, individuals and properties, we have reused all the 34 other properties already implemented in BioPortal;
2. Pick up the OMV property if existing (to stay consistent with BioPortal’s historical choice of using OMV);
3. If not available within OMV, choose property from any other vocabulary offering the best correspondence by giving preference when possible to W3C Recommendations or Notes. With this in mind, we prefer `dct:publisher` to `schema:publisher` and `adms:schemaAgency`. Or, `foaf:fundedBy` rather than `mod:sponsoredBy` and `schema:sourceOrganization`.

<sup>19</sup> Although the domain of `rdfs:isDefinedBy` is `rdfs:Resource`, it is defined by the RDF specification as: “may be used to indicate an RDF vocabulary in which a resource is described.”

We came up with a list of 127 properties in the restricted unified model including the 46 original ones from BioPortal (nonspecific) and 82 new ones from metadata vocabularies. For a better comprehension, we categorized the properties as illustrated in Table 4. Among them, 17 properties from BioPortal cannot be mapped to any of the studied vocabularies, which means that they are candidates for extending one of the studied vocabularies or creating a new one (cf. Sect. 7.1). For example: `bpm:group`, `bpm:downloadCsv`, or a few metrics, and properties describing the classes.

When selecting a default property for the unified model and grouping properties by equivalences, we had to make choices (that we have tried less arbitrary possible). These were guided by our context and motivation (i.e., implementing this model in AgroPortal) and shall differ from projects with other motivations. Here are a few examples of these choices:

- We kept `omv:notes` over `rdfs:comment`, or `adms:versionNotes` in order to stay consistent with BioPortal’s choice of partially adopting OMV. This choice was made in 2009 right after the OMV vocabulary was proposed and according to us, this was a good choice at that time. We would not necessarily encourage the use of `omv:notes` (or any OMV property for which a more standard vocabulary already provides something) over `rdfs:comment` anymore now. Indeed, this is a limitation of OMV that we have pointed out. Finally, our model includes 35 of the 37 relations of OMV. The two missing are `omv:reference` and `omv:resourceLocator` that we have not included because BioPortal already offered a property for them (but not the OMV one!) respectively `bpm:publication` and `bpm:pullLocation`.
- For a property that was not already captured by BioPortal or OMV, such as the fact that an ontology is deprecated, we give priority to established standards, e.g., `owl:deprecated` over `idot:obsolete` as the OWL property (which applies to any IRI) comes from a W3C Recommendation.
- We selected `dct:publisher` over `schema:publisher` as our analysis has shown that Dublin Core (and Elements) properties are widely used among ontology developers. This might of course change in the future considering the pace of adoption of Schema.org.<sup>20</sup>
- For the relation between an ontology and a view of this ontology, BioPortal defines `bpm:viewOf` and `bpm:views` that we have kept, respectively, over `dct:isPartOf` (or `schema:isPartOf` or `void:subset` or `door:sematicallyIncludedIn`) and `dct:hasPart` (or

<sup>20</sup> On that example, one can regret the fact that Schema.org has not itself adopted Dublin Core or that the two organizations do not work together. Similarly, Schema.org and DCAT are particularly rich and we shall follow closely the effort of harmonizing them in the future.

**Table 4** Restricted list of 127 properties (“default”) implemented in AgroPortal’s unified metadata model

Category	List of properties in this category
Intrinsic properties	omv:acronym, omv:name, dct:alternative, skos:hiddenLabel, omv:URI, owl:versionIRI, dct:identifier, omv:version, omv:status, owl:deprecated, omv:hasLicense, omv:hasOntologyLanguage, omv:hasFormalityLevel, omv:hasOntologySyntax, omv:naturalLanguage
Description	omv:description, bpm:publication, omv:documentation, dct:abstract, cc:morePermissions, cc:useGuidelines, schema:copyrightHolder, bpm:pullLocation, omv:notes, omv:keywords, omv:isOfType, omv:designedForOntologyTask, omv:usedOntologyEngineeringTool, omv:usedOntologyEngineeringMethodology, omv:conformsToKnowledgeRepresentationParadigm, dct:coverage, mod:competencyQuestion, foaf:depiction, foaf:logo, foaf:homepage, schema:associatedMedia, bpm:diffFilePath, vann:example, idot:exampleIdentifier, vann:preferredNamespaceUri, vann:preferredNamespacePrefix, void:uriRegexPattern, bpm:prefLabelProperty, bpm:definitionProperty, bpm:synonymProperty, bpm:authorProperty, bpm:hierarchyProperty, bpm:obsoleteProperty, bpm:obsoleteParent, schema:includedInDataCatalog
People	omv:hasCreator, omv:hasContributor, dct:publisher, pav:curatedBy, bpm:contact, schema:translator
Grouping	omv:hasDomain, bpm:group
Relation	omv:useImports, omv:hasPriorVersion, omv:isBackwardCompatibleWith, omv:isIncompatibleWith, bpm:viewOf, bpm:views, bpm:submissions, bpm:hasPart, dct:isFormatOf, dct:hasFormat, door:ontologyRelatedTo, door:similarTo, door:comesFromTheSameDomain, door:explanationEvolution, door:hasDisparateModelling, door:isAlignedTo, schema:translationOfWork, schema:workTranslation, voaf:usedBy, voaf:generalizes, voaf:hasDisjunctionsWith
Content	omv:keyClasses, bpm:ui, sd:endpoint, voaf:metadataVoc, bpm:csvDump, bpm:properties, bpm:classes, bpm:roots, void:dataDump, void:uriLookupEndpoint, void:openSearchDescription, bpm:downloadRdf, bpm:downloadCsv
Community	omv:knownUsage, omv:endorsedBy, bpm:projects, dct:audience, bpm:analytics, foaf:fundedBy, bpm:reviews, bpm:notes, voaf:toDoList, doap:repository, doap:bug-database, doap:mailing-list, schema:award
Date	omv:creationDate, bpm:released, omv:modificationDate, dct:valid, pav:curatedOn
Metrics	omv:numberOfClasses, omv:numberOfIndividuals, omv:numberOfProperties, omv:numberOfAxioms, bpm:maxDepth, bpm:maxChildCount, bpm:averageChildCount, bpm:classesWithOneChild, bpm:classesWithMoreThan25Children, bpm:classesWithNoDefinition, void:entities
Provenance	dct:source, prov:wasGeneratedBy, prov:wasInvalidatedBy, dct:accrualMethod, dct:accrualPeriodicity, dct:accrualPolicy

schema:hasPart or oboInOwl:hasSubset or adms:sample) to keep our model backward compatible.

The selection of default properties and equivalences is the more subjective part of our work. Our choices were driven by our needs and are subject to future modifications (see discussion Sect. 7.1). Somehow, they had to be made to nourish our project of demonstrating the power of harmonized metadata in an ontology repository. We shall certainly update these choices to accommodate small changes based on user feedback or experience. The latest complete list of properties and the equivalences implemented in AgroPortal are available via a web service call: [http://data.agroportal.lirmm.fr/submission\\_metadata](http://data.agroportal.lirmm.fr/submission_metadata)

## 6 Harnessing the Power of Unified Metadata in AgroPortal

Our goal was to implement a new metadata model into an ontology repository and give sense and valorize these metadata. We want to illustrate inside an ontology repository why ontology metadata are important and how they can be leveraged to provide new interesting insights to ontology developers and final users. We also believe that it is the role

of an ontology repository to capture and give sense to metadata information interlinking ontologies together (e.g., the relation between ontologies).

### 6.1 Implementation Within AgroPortal

We have used the restricted list of Table 4 to implement a unified ontology metadata model within AgroPortal. We have added the 79 new properties into the original model (of 46 properties) precisely respecting the cardinalities of the properties.<sup>21</sup> This model is used to describe the ontologies being “hosted” within the portal, not the original ontology (to which only the original developers have authority on). Technically and formally speaking, this means that the metadata properties populated within AgroPortal apply to resources created by the portal, not the original URIs of the ontologies. For example, the National Agricultural Library Thesaurus (NALT) has for URI: <http://lod.nal.usda.gov/nalt> but the metadata properties, represented in JSON-LD within AgroPortal are assigned to the following resources: <http://d>

<sup>21</sup> With the objective of keeping our implementation simple, we have decided to add every new property to the Submission object. The range is generally either an URI or a String.

[ata.agroportal.lirmm.fr/ontologies/NALT](http://ata.agroportal.lirmm.fr/ontologies/NALT) <http://data.agroportal.lirmm.fr/ontologies/NALT/submissions/3>.<sup>22</sup>

This gives us more flexibility when implementing a unified metadata model and facilitates the valorization and use of the metadata over all the ontologies, although it could create a confusion in terms of linked data being produced by the portal. For instance, an ontology creator may have used `dc:title` in the original ontology file but we will actually use the property `omv:name` for the metadata being stored on the portal.<sup>23</sup>

When an ontology is uploaded, AgroPortal extracts automatically most of the ontology metadata if they are included in the original file or populates some of them (e.g., metrics, endpoints, links, examples). Those values can manually be changed after by ontology developers or the portal administrators if they want to provide another value. We populate the 127 properties of the unified model by automatically parsing any of the 346 properties of the complete list presented in Sects. 5.2 and 5.3. When the original ontology file uses a property to capture metadata, we copy the value of this property to the default property chosen in the unified model and assign it to the resource created to represent the ontology within AgroPortal. Sometimes, the properties happen to be the same but often they are not. In the (very exceptional) case where multiple properties from the original file map to the same default property within the model, we aggregate the values or use multiple instances of the default property to keep all the original information. Then AgroPortal's REST web service will return the metadata of the hosted ontology, not the ones from the original file. Advanced users can still access the original metadata using the AgroPortal's SPARQL endpoint (<http://sparql.agroportal.lirmm.fr/test>) on which both URIs (hosted and original) are queryable. For example, if an ontology developer would use `dc:creator` for John, Alice and Tom and then `pav:createdBy` for NIH, WHO and NCBI, then AgroPortal's REST service API will return `omv:hasCreator` for John, Alice, Tom, NIH, WHO and NCBI. The SPARQL endpoint will return the original metadata.

For each ontology, available and uploaded in the portal, we collaborate with the ontology developers to extensively describe their metadata and we have spent a significant amount of time editing, curating and harmonizing the metadata. Information is generally found in other libraries (e.g.,

LovInra, VEST Registry, OBO Foundry, FAIRsharing) or identified in the publications, web sites, documentation, etc. found about the ontologies.

Now all the ontologies within AgroPortal are described with the same unified metadata model and we have invested a significant effort in editing metadata. This has resulted in three important new features for AgroPortal (Table 5):

- AgroPortal's ability to semantically capture and display a very large number of information about an ontology. The *Ontology Summary* page allows getting all the metadata information about a specific ontology. It helps users to know more about the ontologies they are using (or consider using); this will facilitate the ontology selection process and overall, make ontologies more FAIR. Plus, thanks to the portal architecture, all these data is formally described, with semantic web (standard) vocabularies and available as linked data (JSON-LD). In addition, we have entirely redesigned AgroPortal's ontology submission page to facilitate the edition of the metadata. Whenever possible, the user interface facilitates the selection of the metadata values, while in the backend those values are stored with standard URIs. For instance, the user interface will offer a pop-up menu to select the relevant license (CC, BSD, etc.) while the corresponding URI will be taken from the RDFLicense dataset (<http://rdflicense.appspot.com>). Knowledge organization systems types are taken from the NKOS Types Vocabulary of the Dublin Core initiative.<sup>24</sup> Natural languages are taken from the LEXVO vocabulary [66]. Ontology syntax values are provided by the W3C.<sup>25</sup> Some other values (the type of ontology or formality level) are taken as individuals from OMV. An example using the OntoBiotope ontology metadata page in AgroPortal is shown in Fig. 1.
- Advanced ontology search and selection thanks to AgroPortal's *Browse Ontologies* page (Fig. 2) which offers a convenient user interface with sorting, filtering, and facets that facilitate the identification of the ontology(ies) of interest. We now offer nine facets, based on the metadata, to filter ontologies including four new ones (content, natural language, formality level, type) as well as seven options to sort this list including two new ones (name, released date). These new features facilitate the process of selecting relevant ontologies.
- We have begun facilitating the comprehension of the agronomical ontology landscape by displaying diagrams, charts, and graphs about all the ontologies on the portal (average metrics, most used tools, leading contributors and organization, and more). We have created a new AgroPor-

<sup>22</sup> AgroPortal web service API requires a key to answer the `data.agroportal` calls. Users of the API will have to create an account on AgroPortal to get an APIkey (the same procedure is required with NCBO BioPortal). For the NCBO BioPortal, examples may be found here: <http://data.bioontology.org/documentation#OntologySubmission>.

<sup>23</sup> The perfect technical choice would have been the one of LOV, which only deals with a unified metadata following a specific announced vocabulary; however, we have demonstrated that only one or two standard vocabularies do not cover all the required fields for ontologies.

<sup>24</sup> [http://wiki.dublincore.org/index.php/NKOS\\_Vocabularies](http://wiki.dublincore.org/index.php/NKOS_Vocabularies) (ANSI/NISO Z39.19-2005).

<sup>25</sup> <https://www.w3.org/ns/formats/>.

**Table 5** Summary of metadata use within AgroPortal ontology repository

	Ontology Summary page	Browse Ontologies page	Landscape page
Description	Gives all the metadata information about a specific ontology	Allows to search, order and select ontologies using a faceted search approach, based on the metadata	Allows to explore the agronomical ontology landscape by automatically aggregating the metadata fields of each ontologies in explicit visualizations (charts, term cloud and graphs)
New compared to BioPortal	The whole “Additional Metadata” block which corresponds to properties from our new model. Plus the “Get my metadata back” buttons	Four additional ways to filter ontologies in the list (content, natural language, formality level, type) as well as two new options to sort this list (name, released date)	This page did not exist in the original BioPortal
Example (user interface)	<a href="http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE">http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE</a> (see also Fig. 1)	<a href="http://agroportal.lirmm.fr/ontologies">http://agroportal.lirmm.fr/ontologies</a> (see also Fig. 2)	<a href="http://agroportal.lirmm.fr/landscape">http://agroportal.lirmm.fr/landscape</a> (see also Fig. 3 to Fig. 9)
Example (API call)	<a href="http://data.agroportal.lirmm.fr/ontologies/ANAEETHES/submissions/2?display=all">http://data.agroportal.lirmm.fr/ontologies/ANAEETHES/submissions/2?display=all</a>	<a href="http://data.agroportal.lirmm.fr/ontologies">http://data.agroportal.lirmm.fr/ontologies</a>	E.g., to get omv:hasLicense property <a href="http://data.agroportal.lirmm.fr/submissions?display=hasLicense">http://data.agroportal.lirmm.fr/submissions?display=hasLicense</a>

tal *Landscape* page that displays metadata “by property” (as opposed as “by ontology” as in Fig. 1) by aggregating the metadata values (Sect. 6.2).

## 6.2 AgroPortal’s Landscape Page

We have now a specific page dedicated to visualizing the ontology landscape in AgroPortal that facilitates analysis of the repository content. The landscape page helps to figure out what are some of the main domain of interests as well as common development practices when creating a vocabulary or ontology in agronomy. Of course, this information relies on the metadata extracted from the ontologies or edited on the portal. Such visualizations are also meant to motivate the ontology developers to document and describe more their ontologies. In the following, we present some views (figures) automatically created with the content of the repository from May 2017. Whenever possible, we also explicitly mention the metadata property used to generate the view.

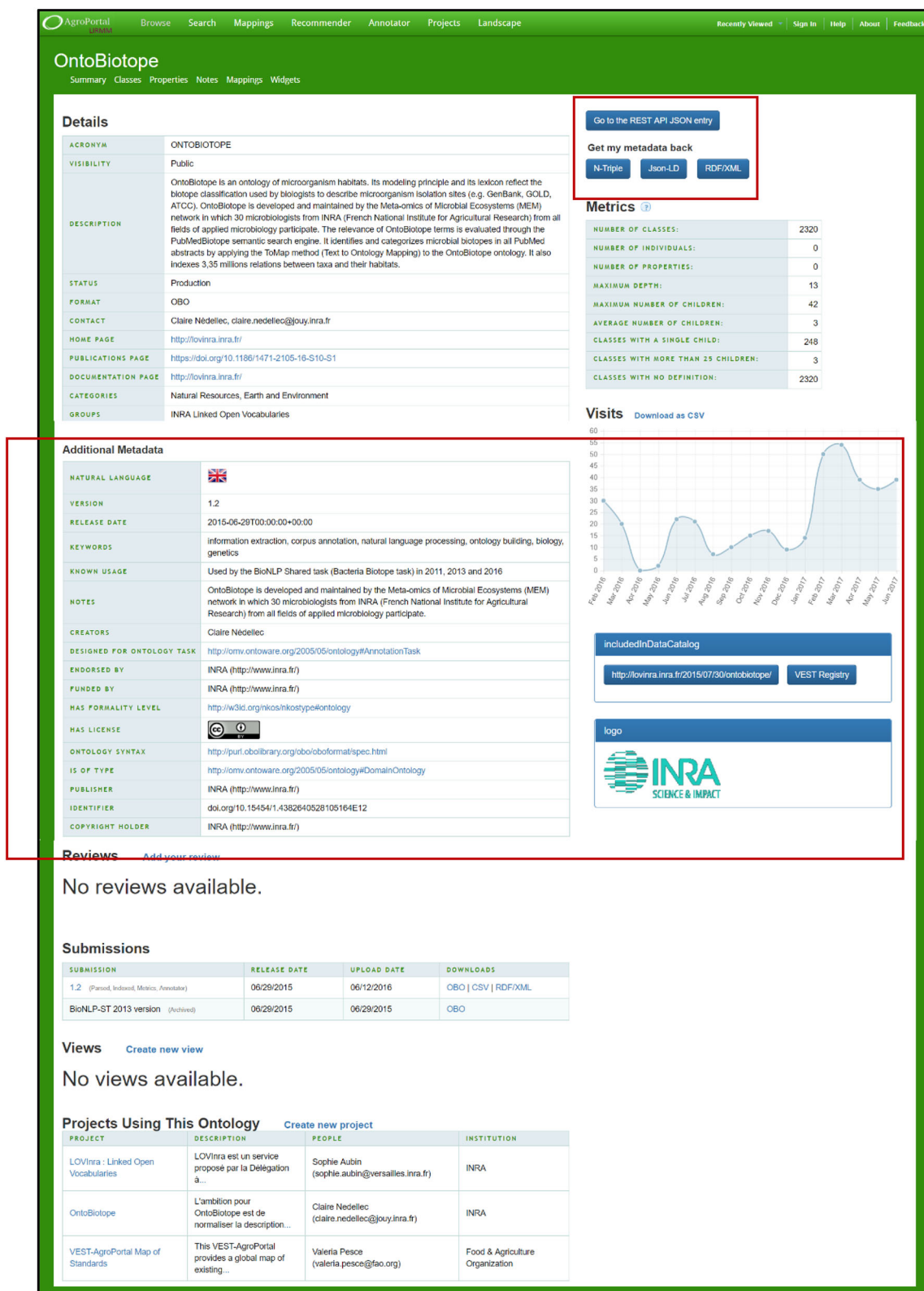
Within AgroPortal (as in the original BioPortal) we organize the ontologies in relevant group and categories (Fig. 3): each time an ontology is uploaded into the portal, it is manually assigned a group and/or category. The groups allow bringing together ontologies from the same project or organization for better identification of the provenance. The categories are another way to classify ontologies in the portal by domain. The groups and categories are customizable and will be adapted in the future to reflect the evolution of the portal’s content and community feedback. Another good aspect of the portal’s architecture is that it provides URIs for any objects in the portal including groups and categories e.g., <http://data.agroportal.lirmm.fr/categories/FARMING> identifies the category “Farms and Farming Systems.”

External applications can now use these URIs to organize ontologies or tag them.

The most commonly adopted format is OWL (Fig. 4) which confirms the agronomy community has clearly turned to the W3C Recommendation for building ontologies. In addition, we already host six vocabularies in SKOS, which shall be a format that will grow in the future. It has been adopted, for instance, by the ANAEE Thesaurus, Agrovoc, NAL and CAB Thesaurus. Figure 4 also shows that most of the ontologies are in the range between 100 and 10 K classes (or concepts), although a few big resources have been uploaded. The metrics in AgroPortal are automatically computed by the OWL-API, but they can be overridden manually. The size of the ontology is generally the number of classes (except with the SKOS format, where it is the number of individuals).

Ontology labels are mostly in English (Fig. 5) although we have seven resources that offer French labels (mostly because of our French collaborators). Multilingual resources include Agrovoc and NAL Thesaurus. Figure 5 also shows that among the 31 ontologies that have explicitly defined licensing information, all of them are openly accessible with different licenses. Note AgroPortal can also host private ontologies or restrict download for public ones.

The type and formality level of resources are described in Fig. 6. The number of upper level ontologies (not specifically dedicated to agriculture) is maintained low and not surprisingly most of the ontologies are domain or application ontologies. Acknowledging the “ambiguity” of these information, as there are no standard definitions of the type and formality level of a knowledge organization system, we do think this information is useful and may help to select the right resources for a given task [14].



**Fig. 1** Screenshot of the *Ontology Summary* page for the Ontobiotope ontology ([http://agroportal.lirmm.fr/ontologies/ONTOBIOT\\_OPE](http://agroportal.lirmm.fr/ontologies/ONTOBIOT_OPE)). The section “Additional Metadata” has been automatically extracted from the content of the original ontology file or edited by

AgroPortal admin or the ontology owner. We have not yet implemented the change at the user interface level to display nice values rather than the raw URIs. This will be done in the next future

Browse
Search
Mappings
Recommender
Annotator
Projects
Recently Viewed
Sign In
Help

# Browse

Access all ontologies that are available in IBC AgroPortal: You can filter this list by category to display ontologies relevant for a certain domain. You can also filter ontologies that belong to a certain group. [Subscribe to the IBC AgroPortal RSS feed](#) to receive alerts for the submissions of new ontologies, new versions of ontologies, new notes, and new projects. You can subscribe to feeds for a specific ontology at the individual ontology page. Add a new ontology to IBC AgroPortal using the Submit New Ontology link (you need to [sign in](#) to see this link).

Submit New Ontology

**Entry Type**  
☒ Ontology (63)  
☐ Ontology View (2)  
☐ CIMI Model (0)  
☐ NLM Value Set (0)

**Uploaded in the Last**

**Category**  
☐ Agricultural Research, Techn...  
☐ Animal Science and Animal P...  
☐ Breeding and Genetic Impro...  
☐ Farms and Farming Systems ...  
☐ Fisheries and Aquaculture ...  
☐ Food Security (1)  
☐ Food and Human Nutrition (4)  
☐ Forest Science and Forest Pro...  
☐ Geographical Locations (0)  
☐ Government, Agricultural La...  
☐ Health and Pathology (0)

**Group**  
☐ AGBIODATA (33)  
☐ AGROLD (14)  
☐ CROP (18)  
☐ LOVINRA (14)  
☐ OBO-FOUNDRY (17)  
☐ WHEAT (19)

**Format**  
☐ OBO (13)  
☐ OWL (44)  
☐ SKOS (4)  
☐ UMLS (2)

**Ontology Content**  
☐ Notes (3)  
☐ Reviews (0)  
☐ Projects (57)  
☐ Summary Only (0)

**Natural Language**  
☐ German (1)  
☐ English (58)  
☐ French (6)  
☐ Italian (1)  
☐ Portuguese (1)  
☐ Spanish (2)

**Formality Levels**  
☐ Classification scheme (1)  
☐ Dictionary (0)  
☐ Gazetteer (0)  
☐ Glossary (0)  
☐ List (0)  
☐ Name authority list (0)  
☐ Ontology (39)  
☐ Semantic network (1)  
☐ Subject heading scheme (0)  
☐ Synonym ring (0)  
☐ Taxonomy (2)

**Is of Type**  
☐ Application Ontology (15)  
☐ Core Ontology (0)  
☐ Domain Ontology (22)  
☐ Task Ontology (0)  
☐ Upper Level Ontology (5)  
☐ Vocabulary (0)

Search...

Showing 63 of 65 Sort: Popular

**AGROVOC (AGROVOC)**  
AGROVOC is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc  
Uploaded: 3/31/17  
concepts  
681,570

**AnaEE Thesaurus (ANAEETHES)**  
The AnaEE thesaurus aims to provide a controlled vocabulary for the semantic description of the study of continental ecosystems and their biodiversity  
Uploaded: 3/23/17  
projects  
1  
concepts  
3,323

**National Agricultural Library Thesaurus (NALT)**  
The Thesaurus is an online vocabulary of agricultural terms in English and Spanish and is cooperatively produced by the National Agricultural Library, USDA and the Inter-American Institute for Cooperation on Agriculture as well as other Latin American agricultural institutions belonging to the Agriculture Information and Documentation Service of the Americas (SIDALC)  
Uploaded: 4/26/17  
concepts  
67,311

**OntoBiotope (ONTOBIOTOPE)**  
OntoBiotope is an ontology of microorganism habitats  
Uploaded: 6/12/16  
projects  
3  
classes  
2,320

**Protein Ontology (PR)**  
An ontological representation of protein-related entities  
Uploaded: 6/30/15  
projects  
1  
classes  
83,656

**IBP Crop Research Ontology (CO\_715)**  
Describes experimental design, environmental conditions and methods associated with the crop study/experiment/trail and their evaluation.  
Uploaded: 6/26/15  
projects  
3  
classes  
256

**Process and Observation Ontology (PO2)**  
A core ontology for modeling transformation processes and their observations.  
Uploaded: 3/29/17  
projects  
2  
classes  
4,449

**IBP Wheat Trait Ontology (CO\_321)**  
Wheat Ontology  
Uploaded: 9/19/16  
notes  
1  
projects  
5  
classes  
1,023

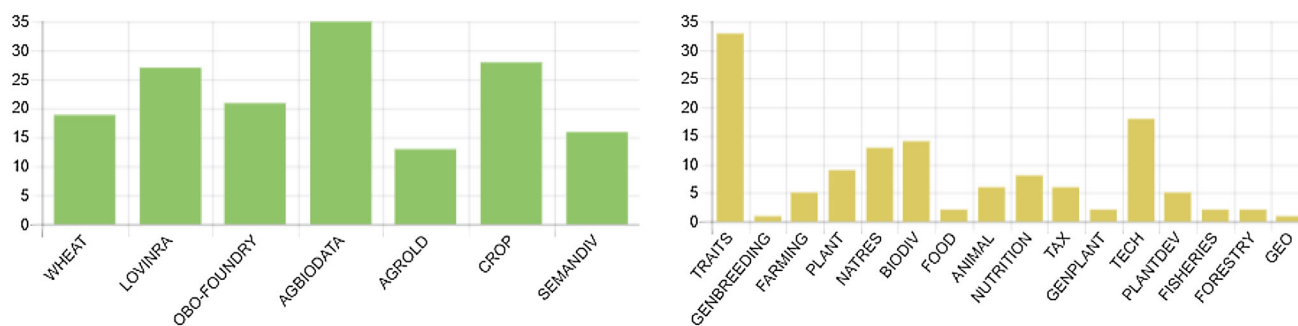
**Agricultural Experiments Ontology (AEO)**  
AEO is an ontology aimed to represent objects related to agricultural practices.  
Uploaded: 2/24/17  
projects  
1  
classes  
56

**Gene Ontology (GO)**  
The GO defines concepts/classes used to describe gene function, and relationships between these concepts  
Uploaded: 5/8/17  
projects  
4  
classes  
48,603

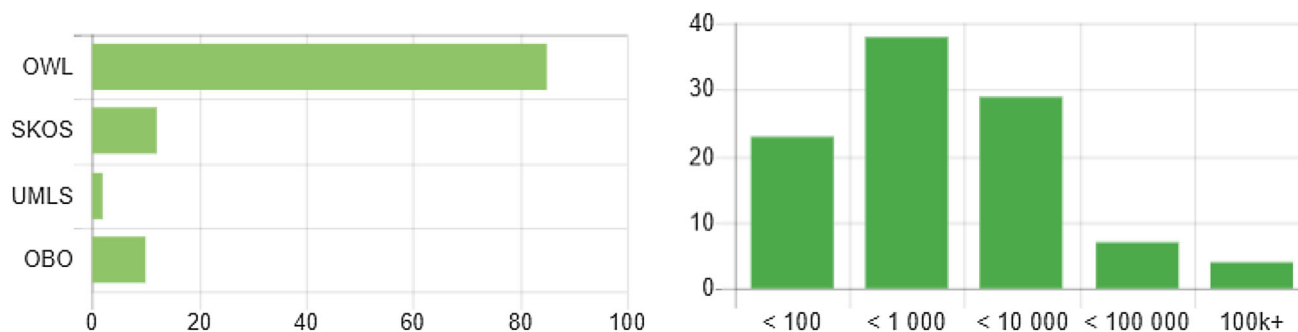
**Wheat Trait Ontology (WHEATPHENOTYPE)**  
WheatPhenotype is an ontology in Obo format that describes the traits of soft wheat (*Triticum aestivum*) and the environmental factors that affect these traits  
Uploaded: 10/9/16  
projects  
3  
classes  
466

**Animal Disease Ontology (ADO)**  
L'ontologie des maladies animales organise la base de connaissances des Maladies Animales, The Animal diseases Ontology supports the Animal diseases Knowledge Base.  
Uploaded: 3/21/17  
projects  
2  
classes  
3

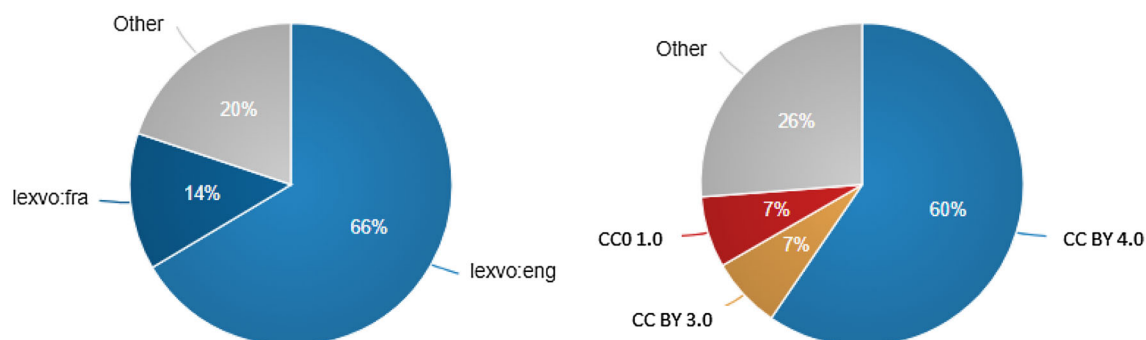
Fig. 2 Screenshot of the *Browse Ontologies* page. Facetted search (left hand side) and sorting (top right corner) offer new ways to select ontologies



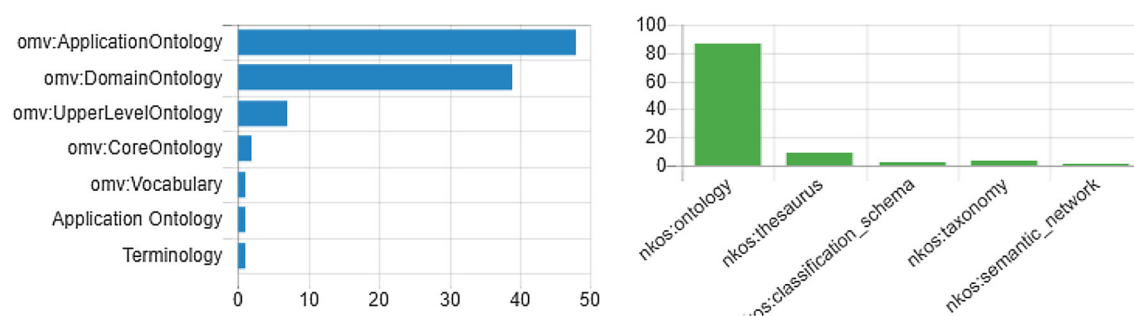
**Fig. 3** Distribution of ontologies by Group (bpm:group) and Categories (bpm:hasDomain)



**Fig. 4** Ontologies by format (omv:hasOntologyLanguage) and sizes (bpm:metrics.classes or bpm:metrics.individuals)



**Fig. 5** Natural languages (omv:naturalLanguage) used for labels and licenses (omv:hasLicense) of ontologies



**Fig. 6** Ontology types (omv:isOfType) and formality levels (omv:hasFormalityLevel)

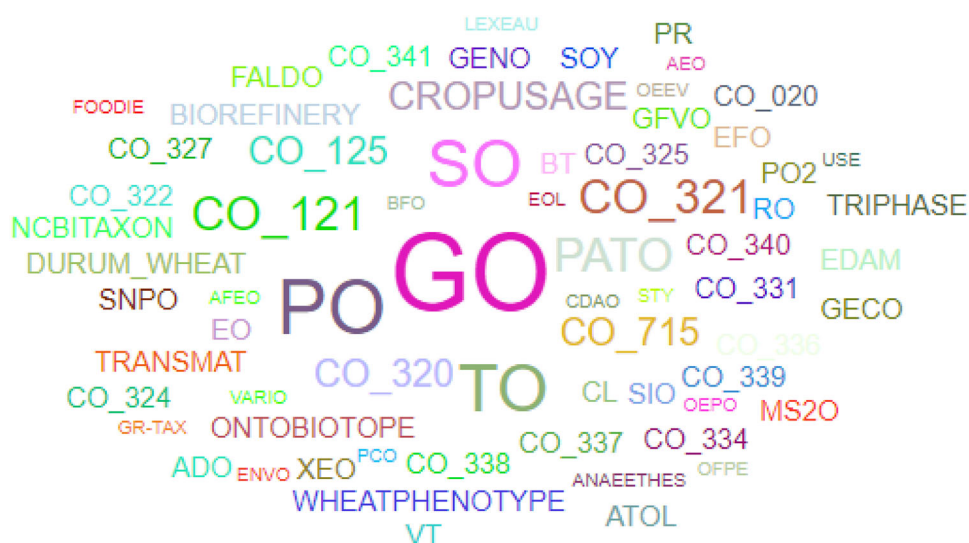
Figure 7 is an aggregation (term cloud) of several properties that relate ontologies and organizations. Such a view is interesting to identify which organizations are the most

involved in funding, adopting or endorsing ontologies. Figure 8 is a similar cloud showing which ontologies are the most actively commented, reviewed or used within research

**Fig. 7** Most mentioned organizations (aggregation as a term cloud from the properties dct:publisher, foaf:fundedBy and omv:endorsedBy)



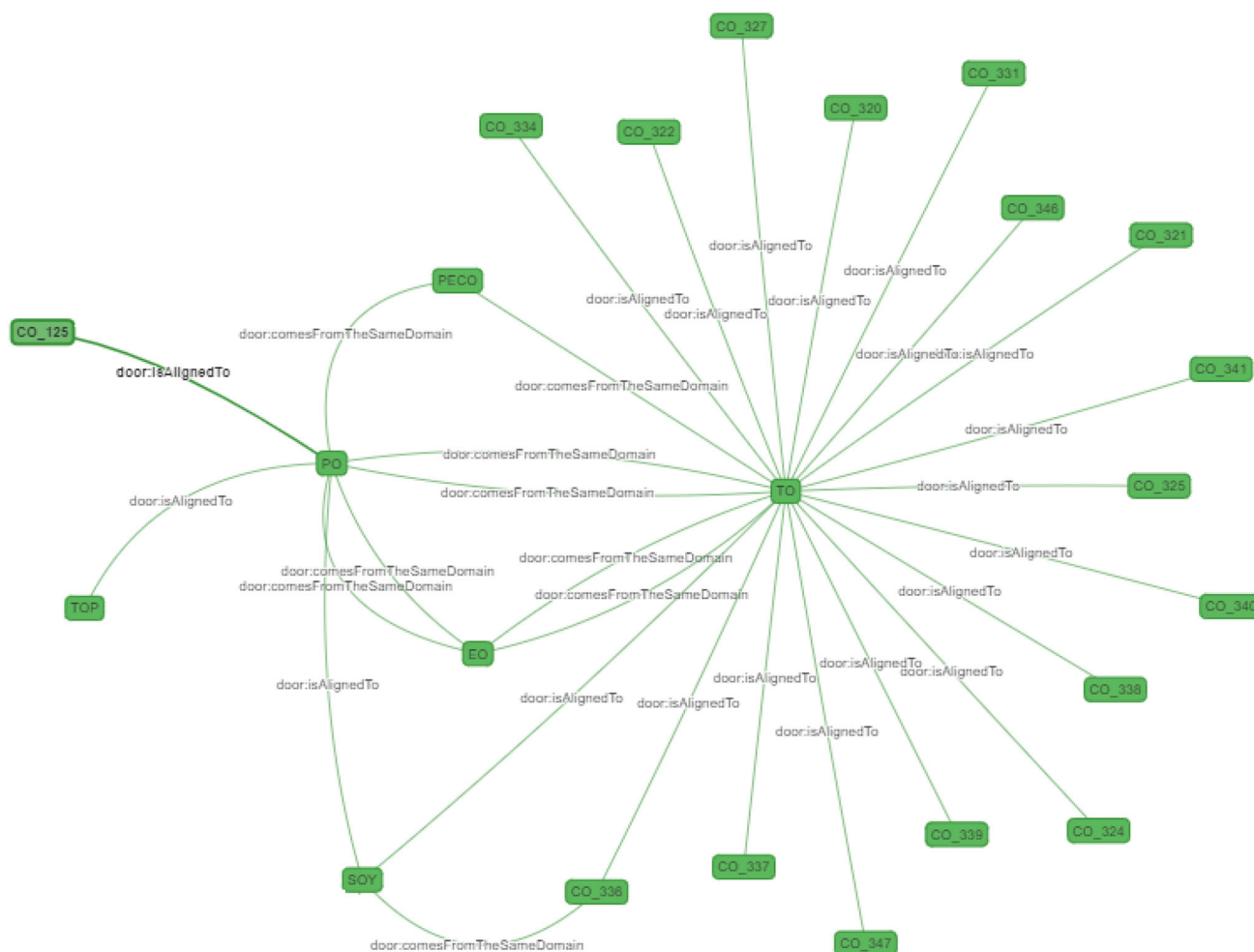
**Fig. 8** Most active ontologies (count aggregation of omv:notes, bpm:reviews and bpm:projects per ontology)



projects. Indeed, AgroPortal features a few community features [67] such as: *ontology reviews* or *notes* that can be attached in a forum-like mode to a specific ontology or class, in order to discuss the ontology (its design, use, or evolution) or allow users to propose changes to a certain class. Plus, AgroPortal provides a project list edited by its users that materialize the ontology-project relation (<http://agroportal.lirmm.fr/projects>), i.e., which project uses which ontologies.

The new metadata model allows capturing multiple relations between ontologies or between ontologies and external resources. For instance, relations to capture that an ontology is aligned to another one, represents knowledge from the same domain, is compatible or incompatible with another one, imports or uses another one, is translated from or more

generally related to another one. We have used 14 of these relations to automatically represent AgroPortal's ontologies network. Figure 9 shows the cluster of the ontologies maintained and extended within the Planteome project [25]. It captures the information that all the Crop Ontologies (CO\_\*) are aligned to the Trait Ontology, itself interconnected to the Plant Ontology and Plant Environment ontology. The Soy Ontology, developed outside of the Crop Ontology project, also appears as related to both TO and CO\_336 (the Soybean Ontology developed within the Crop Ontology project). Figure 9 is only a subset of the network. The landscape page within AgroPortal displays the whole network and filters it per ontology relations.



**Fig. 9** Subset of the ontology network showing the relations between reference plant ontologies (here properties `door:isAlignedTo` and `door:comesFromTheSameDomain`)

### 6.3 User Appreciation Survey

To evaluate the impact and appreciation of the new features enabled by our changes in AgroPortal's ontology metadata model, we conducted a survey with typical five-level Likert scale questions. Each question asked for the participant's opinion about how much the new page (or new features in the page) "helped identifying and selecting" relevant ontologies (except for the ontology submission edition page, which was concerned about editing ontology metadata). With this survey, we liked to assess AgroPortal's new metadata model ability to ease ontology identification and selection. Plus, we asked open questions about each page to get users inputs in terms of how to improve ontology metadata within AgroPortal in the future. The survey was sent only to the AgroPortal users mailing list which had 131 members then. We had 32 responses that are analyzed hereafter. 2/3 of the participants were both users and administrators of one or several ontologies in AgroPortal. The last third was only regular

AgroPortal users who usually search and find relevant ontologies and concepts. The questions and responses are presented in Table 6.

Globally, the helpfulness of the pages was clearly established by the survey with almost  $\frac{3}{4}$  of positive responses on average for all questions. Displaying more metadata on the Summary page and being able to filter out ontologies with metadata facets on the Browse page was much appreciated. The Landscape page was ranked as a bit less "useful" than the others (with 53.2% responses explicitly positive) getting still some positive feedbacks and relevant criticisms. An additional question related to the usefulness of the page to "understand about the ecosystem of ontologies in agronomy and close related domains" obtained 75% of positive responses. The absence of response in the "Not at all" and very limited responses in the "Not so" columns show that everyone agrees about the role of metadata when identifying and selecting ontologies. Still, the exploitation of metadata to facilitate this process is improvable. Among the comments

**Table 6** User appreciation survey responses (percentage)

Question/page	Extremely helpful	Very helpful	Somewhat helpful	Not so helpful	Not at all helpfull
New ontology summary page	15.6	59.4	21.9	3.1	0
New ontology browse page	31.3	56.3	9.3	3.1	0
New landscape page	12.6	40.6	37.5	9.3	0
New ontology submission edition page (optional) <sup>a</sup>	19	57.2	19	4.8	0
Average	19.63%	53.38%	21.93%	5.08%	0%

<sup>a</sup>Only 21 responses for this last question

on: (1) the Summary page, some were about improving the user interface by keeping only the relevant fields and using something else than URLs or URIs. (2) the Browse page, most were positive as facets are often appreciated to search information, although the lack of description of the facets was often reported. (3) the Landscape page, many comments were requesting a better integration with the rest of AgroPortal (e.g., links back), merging some information (e.g., Figure 6) and some were about pointing out the importance of curating the metadata to create good value in this page.

## 7 Discussions

According to us, among the main limitations of OMV that might explain why it is not much adopted today are: (1) the fact that it did not reuse any other metadata vocabulary;<sup>26</sup> (2) it was never included in a common ontology editor such as Protégé—it would have highly facilitated the adoption of the vocabulary if ontology developers would have had only to fill out a few forms directly in their preferred ontology edition software; (3) the metadata properties were never really used and valorized by ontology libraries which would have been the best way to incite to fill them up.

In the following, we come back on each of these aspects to discuss the need for a better harmonization of standard vocabularies used to described ontology metadata. Besides our work driven by the AgroPortal project, this effort may be generalized to propose recommendations and guidelines to (1) ontology developers when describing their ontologies; (2) ontology repository or library developers to harmonize their platforms.

### 7.1 Need for Metadata Authoring Guidelines and for Harmonization of Existing Metadata Vocabularies

The analysis of the existing metadata vocabularies and practices (Sect. 4) showed there is a clear need for better metadata authoring guidelines for the community of ontology developers and a need of harmonization of existing metadata vocabularies. MOD1.0 [7] was a first attempt to address OMV's limitation of not relying on any other vocabularies but was not “mapped” itself to OMV while being very similar. Plus, it still missed numerous relevant properties to capture information about ontologies. More recently, the authors joined their efforts and proposed a new version of MOD (refer as MOD 1.2) [48].<sup>27</sup> The revision carried out from multiple aspects (e.g., new labels, structural changes, and design principles) to overcome some of the limitations of MOD 1.0 and to enrich it further influenced also by our work on AgroPortal. MOD1.2 contains 88 properties taken from DCAT, DCT, DOAP, FOAF, OMV, OWL, PAV, PROV, RDFS, SD and VOA but creates only 13 new properties in the MOD namespace. Future extended versions (MOD2.0 and more) shall contain at least equivalent property for each of the 127 of AgroPortal's new metadata model. Note that because MOD development is free from any implementation constraints, we have not always selected in MOD1.2 the same default properties than in AgroPortal's unified metadata model. In [48], we also describe the application goals of MOD1.2 and illustrate our experimental results with SPARQL queries that can be run on properly defined metadata.

MOD 1.2 is a recent initiative and still a temporary proposition. It is understandable that to achieve community adoption, this work needs to engage more people, with the ultimate goal of producing a community standard endorsed by a

<sup>26</sup> Although we acknowledge that in 2005, there was not as vocabularies as today, important standards such as OWL, Dublin Core or FOAF may have been used at that time.

<sup>27</sup> <https://github.com/sifproject/MOD-Ontology>.

standardization body such as W3C. MOD 1.2 was recently introduced to the Research Data Alliance *Vocabulary and Semantic Services Interest Group* (VSSIG).<sup>28</sup> Future work will now happen in the context of the “ontology metadata” task group of the VSSIG. Among the current studied propositions is to implement MOD2.0 as a profile of DCAT. We shall also make sure we will enforce—or enable operationalization of—the recently published MIRO guidelines [6].

## 7.2 Metadata Edition

Another important aspect in metadata is that almost no one really like filling them in; therefore, how can we facilitate metadata editing for ontology developers? Within AgroPortal, we have entirely redesigned the ontology information edition page and have tried to build it in a way that will both facilitate the edition and not freak out the editors with a basic list of 127 properties to fill in. However, this page will need improvements. We do envision paths for the future:

- Metadata should be as much as possible generated or predicted automatically either by the ontology edition software or by external tools,<sup>29</sup> e.g., software used, dates, languages.
- It is the role of ontology edition software to actually support (some) metadata edition functionalities. It would highly facilitate the task (and the emergence of a standard vocabulary) if ontology editors would only need to fill out a few forms directly in their preferred ontology edition software. Indeed, as seen in Sect. 4.2 properties available for editing (or even better, automatically generated) within the ontology editor are inclined to be well used.
- It is the role of ontology libraries to facilitate the edition, generation and prediction of ontology metadata for properties that take their senses within a community-based library, e.g., relations between ontologies, reviews, related projects, etc. When relevant, the libraries should offer a mechanism to easily export the metadata edited or generated in order for ontology developers to include it in the original ontology file for other systems to use it. Within AgroPortal, we have developed such a mechanism on the *Ontology Summary* page.<sup>30</sup> In addition of an API call, the “Get my metadata back” buttons allow ontology developers, on a simple click, to download the metadata stored within the portal in RDF/XML, JSON-LD or N-triples syntax to copy/paste within the original ontology. An additional question related to the interest of this functionality

was included within the survey presented Sect. 6.3 and obtained 62.5% positive responses. Right now, this feature will return metadata following AgroPortal’s model, but when MOD2.0 will be available or any community adopted standard, it will return the metadata with respect to this standard.<sup>31</sup>

In the future, we plan to discuss with the Protégé development team the integration of some of the listed properties in the software, so that developers can edit them in the ontology development process. We are also considering results of the CEDAR project (<http://metadatacenter.org>) in terms of metadata prediction and edition [38].

## 7.3 Automatic Ontology Selection and Recommendation

An unified metadata model can also be leveraged by automatic ontology selection tools such as the Recommender also available in Agro/BioPortal [3, 68] which relies mostly on the content of ontologies to recommend them. For instance, the whole network built out of ontologies relations (Fig. 9) will help users to figure out which are the key relevant ontologies to rely on. As another example, searching “for ontologies” often rely on “searching inside” ontologies (method based on coverage) which is not very often satisfactory when instead metadata should be used. For example, searching “anatomy” in BioPortal Search will return a bunch of popular ontologies that contains the term anatomy, but the Foundational Model of Anatomy, which is the reference ontology about human anatomy will not show up in the results. To identify FMA, someone needs to browse the ontologies and filter ontologies with the word anatomy in the ontology name or description. Or better, he or she might use the “Anatomy” ontologies category that BioPortal defines. In both cases, this relies on metadata, not on the content of the ontology.

## 8 Conclusion and Future Work

In this paper, we have shown the impact of unified and harmonized metadata within an ontology repository. We have explained how it facilitates ontology description, selection and helps to capture the global landscape of ontologies from a given domain. Thanks to this new unified model served by a stable API, metadata descriptions of AgroPortal ontologies have already been automatically harvested by two external ontology libraries: the Agrisemantics Map of Data Standards (<http://vest.agrisemantics.org>) and FAIRsharing (<http://fairsaring.org>).

<sup>28</sup> The RDA Interest Group was reconfigured in 2017 (<https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html>).

<sup>29</sup> For instance, BioPortal uses the OWL-API to generate metrics. Protégé also does but does not save these metrics inside the ontology.

<sup>30</sup> This feature has been recently developed and is still in beta mode.

<sup>31</sup> Before completely changing AgroPortal’s model, we believe each library could at least import/export MOD2.0 compliant metadata.

Our motivation was first to make a review of the available vocabularies to describe ontologies or other kinds of resources (dataset, vocabulary, project, document) and pick up the properties that would be relevant for describing ontologies. Since the OMV initiative in 2005, there have been multiple propositions especially with the emergence of the web of data. Our goal was then to identify the redundancy and lacks between these vocabularies by regrouping the properties into a restricted unified model that we have implemented in the AgroPortal ontology repository. We have worked on our side and in partnership with our users to fill the metadata and in parallel developed new user interfaces. This has resulted in multiple new features within AgroPortal that we have presented and have been appreciated by our users as facilitating the ontology identification and selection processes.

We can now come back on addressing some concrete motivational use cases described in Sect. 2:

- The new ontology metadata model has been driven by and finally implemented within AgroPortal and the French SIFR BioPortal. The new model makes the description of the ontologies more complete and is unified for all the ontologies.
- The properties `omv:hasLicense`, `dct:publisher`, `cc:morePermissions` and `schema:copyrightHolder` can now be used to precisely describe the licensing information about the ontologies endorsed by the Wheat Data Interoperability working group. The endorsement itself is also captured by the property `omv:endorsedBy`.
- LovINRA ontologies can now be explicitly and unambiguously classified by syntax (`omv:hasOntologySyntax`), format (`omv:hasOntologyLanguage`), type (`omv:isOfType`) and formality level (`omv:hasFormalityLevel`).
- The alignment relations between the Crop Ontology trait ontologies and the Plant Trait Ontology are now captured by `door:isAlignedTo` and other relations such as `door:comesFromTheSameDomain` and `door:ontologyRelatedTo`.

We did not pursue the goal of integrating all the reviewed vocabularies into a new “integrated vocabulary” that could become a standard for describing ontologies (e.g., a new OMV), although the clear need for metadata authoring guidelines and for harmonization of existing metadata vocabularies has also been discussed. We are currently working in generalizing this work within a new version of MOD that would merge and harmonize existing ones. A generalization of this work is studied in a community-driven standardization effort in the context of the RDA *Vocabulary and Semantic Services Interest Group*. We are also discussing with the Stanford NCBO project how to merge back our contributions to the technology into the NCBO BioPortal.

In the future, we want to be able to describe more the usage of ontologies by defining/extending (1) generic tasks for which ontology are used (annotation, indexing, search, reasoning, etc.) and (2) small examples of usages of the ontologies. We also plan to use the same metadata analysis approach to suggest ontology development guidelines based on community practices. For instance, by looking at the most used properties to describe ontologies or their classes. In the future, by integrating more relevant ontologies and vocabularies into AgroPortal and cautiously describing them, we hope to offer a reference portal to identify and use knowledge organizations systems in agronomy, food, plant sciences and biodiversity. We will continue our metadata edition and curation effort to be sure to provide the community with the best descriptions for ontologies available.

**Acknowledgements** This work is partly achieved within the Semantic Indexing of French biomedical Resources (SIFR—[www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) project that received funding from the French National Research Agency (Grant ANR-12-JS02-01001), the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant agreement No 701771, the NUMEV Labex (Grant ANR-10-LABX-20), the Computational Biology Institute of Montpellier (Grant ANR-11-BINF-0002) as well as by University of Montpellier and the CNRS. This work has also been partially funded by the Indian Statistical Institute under the Start-Up Grant project. We also thank the National Center for Biomedical Ontologies for help and time spent with us in deploying the AgroPortal and all our users who have taken some time to author/review the metadata of their ontologies and answer our appreciation survey.

**Author Contributions** CJ conceived of the project, provided the scientific direction and wrote this manuscript. AT worked on the new metadata model and on description of the ontologies with help of the AgroPortal user community. VE implemented the new metadata model, additional features and the landscape page in AgroPortal. BD worked on vocabulary analysis as well reflections on MOD and also contributed to the survey analysis. All authors declare not conflict of interest and approved the final manuscript.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Ding L, Finin T, Joshi A, Peng Y, Cost RS, Sachs J, Pan R, Reddivari P, Doshi V (2004) Swoogle: a semantic web search and metadata engine. In: Grossman DA, Gravano L, Zhai C, Herzog O, Evans D (eds) 13th ACM conference on information and knowledge management, CIKM’04. ACM, Washington DC, USA
2. Bizer C, Heath T, Berners-Lee T (2009) Linked data—the story so far. *Sem Web Inf Syst* 5:1–22
3. Martinez-Romero M, Jonquet C, O’Connor MJ, Graybeal J, Pazos A, Musen MA (2017) NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Biomed Sem* 8:21. <https://doi.org/10.1186/s13326-017-0128-y>

4. Chao T (2015) Mapping methods metadata for research data. *Digit Curation* 10:82–94
5. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hoof R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>
6. Matentzoglou N, Malone J, Mungall C, Stevens R (2018) MIRO: guidelines for minimum information for the reporting of an ontology. *J Biomed Sem* 9:6
7. Dutta B, Nandini D, Kishore G (2015) MOD : metadata for ontology description and publication. In: International conference on Dublin core & metadata applications, DC'15, Sao Paulo, Brazil, pp 1–9
8. Naskar D, Dutta B (2016) Ontology libraries : a study from an ontologist and an ontologist perspectives. In: 19th international symposium on electronic theses and dissertations, ETD'16, Lille, France, pp 1–12
9. Vandenbussche P-Y, Atemezing GA, Poveda-Villalon M, Vatan B (2014) Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. *Sem Web* 1:1–5
10. Côté RG, Jones P, Apweiler R, Hermjakob H (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 7:97
11. Rueda C, Bermudez L, Fredericks J (2009) The MMI ontology registry and repository: a portal for marine metadata interoperability. In: MTS/IEEE Biloxi—marine technology for our future: global and local challenges, OCEANS'09, Biloxi, MS, USA, p 6
12. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith NB, Jonquet C, Rubin DL, Storey M-A, Chute CG, Musen MA (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 37:170–173
13. Sabou M, Lopez V, Motta E (2006) Ontology selection on the real semantic web: how to cover the queens birthday dinner? In: Staab S, Svátek V (eds) 15th international conference on knowledge engineering and knowledge management managing knowledge in a world of networks, EKAW'06. Springer, Pödebrady, Czech Republic, pp 96–111
14. Park J, Oh S, Ahn J (2011) Ontology selection ranking model for knowledge reuse. *Expert Syst Appl* 38:5133–5144
15. Malone J, Stevens R, Jupp S, Hancocks T, Parkinson H, Brooksbank C (2016) Ten simple rules for selecting a bio-ontology. *PLoS Comput Biol* 12:6
16. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium T.O.B.I., Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah NH, Whetzel PL, Lewis S (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25:1251–1255
17. Toulet A, Emonet V, Jonquet C (2016) Modèle de métadonnées dans un portail d'ontologies. In: Diallo G, Kazar O (eds) 6èmes Journées Francophones sur les Ontologies, JFO'16, Bordeaux, France
18. Jonquet C, Toulet A, Arnaud E, Aubin S, Dzale Yeumo E, Emonet V, Graybeal J, Laporte M-A, Musen MA, Pesce V, Larmande P (2018) AgroPortal: a vocabulary and ontology repository for agronomy. *Comput Electron Agric* 144:126–143
19. Whetzel PL, Team N (2013) NCBO technology: powering semantically aware applications. *Biomed Sem* 4S1:49
20. Jonquet C, Annane A, Bouarech K, Emonet V, Melzi S (2016) SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16, Genève, Suisse, p 16
21. Jonquet C, Shah NH, Musen MA (2009) The open biomedical annotator. In: American medical informatics association symposium on translational bioinformatics, AMIA-TBI'09, San Francisco, CA, USA, pp 56–60
22. Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, Bowers S, Buttigieg PL, Davies N, Endresen D, Gandolfo MA, Hanner R, Janning A, Krishtalka L, Matsunaga A, Midford P, Morrison N, Ó Tuama É, Schildhauer M, Smith B, Stucky BJ, Thomer A, Wiczorek J, Whitacre J, Wooley J (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PLoS ONE* 9:e89606
23. Meng X (2012) Special issue—agriculture ontology. *Integr Agric* 11:1
24. Quesneville H, Dzale Yeumo E, Alaux M, Arnaud E, Aubin S, Baumann U, Buche P, Cooper L, Ćwiek-Kupczyńska H, Davey RP, Fulss RA, Jonquet C, Laporte M-A, Larmande P, Pommier C, Protonotarios V, Reverte C, Shrestha R, Subirats I, Venkatesan A, Whan A (2017) Developing data interoperability using standards: a wheat community use case. *F1000Research* 6:1843. <https://doi.org/10.12688/f1000research.12234.2>
25. Cooper L, Meier A, Laporte M-A, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, Qu B, Preece J, Zhang E, Todorovic S, Gkoutos G, Doonan JH, Stevenson DW, Arnaud E, Jaiswal P (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res* 46:D1168–D1180
26. Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front Physiol* 3:1–10
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25:25–29
28. Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, Segerdell E, Mungall C, Westerfield M (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol* 22:345–350
29. Zeng ML (2008) Metadata. Neal-Schuman, New York
30. Obrst L, Gruninger M, Baclawski K, Bennett M, Brickley D, Berg-Cross G, Hitzler P, Janowicz K, Kapp C, Kutz O, Lange C, Levenchuk A, Quattri F, Rector A, Schneider T, Spero S, Thessen A, Vegetti M, Vizedom A, Westerinen A, West M, Yim P (2014) Semantic web and big data meets applied ontology. *Appl Ontol* 9:155–170
31. Graybeal J, Isenor AW, Rueda C (2012) Semantic mediation of vocabularies for ocean observing systems. *Comput Geosci* 40:120–131
32. Baclawski K, Schneider T (2009) The open ontology repository initiative: requirements and research challenges. In: Tudorache T, Correndo G, Noy N, Alani H, Greaves M (eds) Workshop on collaborative construction, management and linking of structured knowledge, CK'09. CEUR-WS.org, Washington, DC, USA, p 10
33. Till M, Kutz O, Codescu M (2014) Ontohub: a semantic repository for heterogeneous ontologies. In: Theory day in computer science, DACS'14, Bucharest, Romania, p 2
34. Sabou M, Lopez V, Motta E, Uren V (2006) Ontology selection: ontology evaluation on the real semantic web. In: Vrandečić D, Suarez-Figueroa MC, Gangemi A, Sure Y (eds) 4th international

- EON workshop, evaluation of ontologies for the web, EON'06. CEUR-WS.org, Edinburgh, Scotland, UK
35. Tejo-Alonso C, Berrueta D, Polo L, Fernández S (2010) Current practices and perspectives for metadata on web ontologies and rules. *Metadata Sem Ontol* 7:10
  36. Palma R, Hartmann J, Haase P (2008) Ontology metadata vocabulary for the semantic web, Report, v2.4, pp 1–85
  37. Weibel S, Kunze J, Lagoze C, Wolf M (1998) Dublin core metadata for resource discovery, RFC 2413, Internet Engineering Task Force
  38. Musen MA, Bean CA, Cheung K-H, Dumontier M, Durante KA, Gevaert O, Gonzalez-Beltran A, Khatri P, Kleinstein SH, O'Connor MJ, Pouliot Y, Rocca-Serra P, Sansone S-A, Wiser JA (2015) The CEDAR team: the center for expanded data annotation and retrieval. *Am Med Inform Assoc* 22:1148–1152
  39. Dumontier M, Gray AJG, Marshall MS, Alexiev V, Ansell P, Bader G, Baran J, Bolleman JT, Callahan A, Cruz-Toledo J, Gaudet P, Gombocz EA, Gonzalez-Beltran AN, Groth P, Haendel M, Ito M, Jupp S, Juty N, Katayama T, Kobayashi N, Krishnaswami K, Laibe C, Le Novère N, Lin S, Malone J, Miller M, Mungall CJ, Rietveld L, Wimalaratne SM, Yamaguchi A (2016) The health care and life sciences community profile for dataset descriptions. *PeerJ* 4:e2331
  40. McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, Sansone S-A (2016) BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database. <https://doi.org/10.1093/database/baw075>
  41. McGuinness DL (2003) In: Fensel D, Hendler J, Lieberman H, Wahlster W (eds) *Spinning the semantic web: bringing the World Wide Web to its full potential*, Chapter 6. MIT Press, Cambridge, MA, pp 171–194
  42. Zeng ML (2008) Knowledge organization systems (KOS). *Knowl Organ* 35:160–182
  43. Zen ML, Zeng ML (2008) Metadata for terminology/KOS resources. In: 8th networked knowledge organization systems workshop, Washington, DC, USA
  44. Hartmann J, Haase P (2005) Ontology metadata vocabulary and applications, pp 906–915
  45. Ding Y, Fensel D (2001) Ontology library systems: the key to successful ontology re-use. In: 1st semantic web working symposium, SWS'01. CEUR-WS.org, Stanford, CA, USA, pp 93–112
  46. Hartmann J, Palma R, Gómez-Pérez A (2009) Ontology repositories. In: Staab S, Studer R (eds) *Handbook on ontologies*. Springer, Berlin, Heidelberg, pp 551–571
  47. D'Aquin M, Noy NF (2012) Where to publish and find ontologies? A survey of ontology libraries. *Web Sem* 11:96–111
  48. Dutta B, Toulet A, Emonet V, Jonquet C (2017) New generation metadata vocabulary for ontology description and publication. In: Garoufallou E, Virkus S, Alemu G (eds) 11th metadata and semantics research conference, MTSR'17, Tallinn, Estonia
  49. Allocca C, d'Aquin M, Motta E (2009) DOOR—towards a formalization of ontology relations. In: International conference on knowledge engineering and ontology development, KEOD'09, Madera, Portugal, pp. 13–20
  50. Vandenbussche P-Y, Vatan B (2012) Metadata recommendations for linked open data vocabularies. Report, v1.1. [https://lov.linkeddata.es/Recommendations\\_Vocabulary\\_Design.pdf](https://lov.linkeddata.es/Recommendations_Vocabulary_Design.pdf)
  51. Min H, Turner S, de Coronado S, Davis B, Whetzel PL, Freimuth RR, Solbrig HR, Kiefer R, Riben M, Stafford GA, Wright L, Ohira R (2016) Towards a standard ontology metadata model. In: Jaiswal P, Hoehndorf R (eds) 7th international conference on biomedical ontologies, ICBO'16, Poster Session, Corvallis, Oregon, USA, p 6
  52. Hausenblas KAM (2009) Describing linked datasets—on the design and usage of void, the vocabulary of interlinked datasets. In: *Linked data on the web workshop, LDOW'09*, Madrid, Spain
  53. Juty N, Novère N Le, Laibe C (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res* 40:580–586
  54. Tirmizi SH, Aitken S, Moreira DA, Mungall C, Sequeda J, Shah NH, Miranker DP (2011) Mapping between the OBO and OWL ontology languages. *Biomed Sem* 2:16
  55. Mäkelä E (2014) Aether—Generating and viewing extended VoID statistical descriptions of RDF datasets. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Springer, Crete, pp 429–433
  56. Ceusters W (2012) An information artifact ontology perspective on data collections and associated representational artifacts. In: Mantas J et al (ed) 24th international conference of the European federation for medical informatics, MIE'12. IOS Press, Pisa, pp 68–72
  57. Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR, Duck G, Furlong LI, Keath N, Klassen D, McCusker JP, Queralt-Rosinach N, Samwald M, Villanueva-Rosales N, Wilkinson MD, Hoehndorf R (2014) The Semantic-science Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J Biomed Sem* 5:14
  58. Fiorelli M, Stellato A, McCrae JP, Cimiano P, Pazienza MT (2015) LIME: the metadata module for OntoLex. In: Gandon F, Sabou M, Sack H, d'Amato C, Cudré-Mauroux P, Zimmermann A (eds) 12th European semantic web conference, ESWC'15. Springer, Portoroz, pp 321–336
  59. McCrae J, Spohr D, Cimiano P (2011) Linking lexical resources and ontologies on the semantic web with lemon. In: Antoniou G, Grobelnik M, Simperl E, Parsia B, Plexousakis D, DeLeenheer P, Pan JZ (eds) 8th extended semantic web conference, ESWC'11. Springer, Heraklion, Crete, pp 245–259
  60. Ong E, Xiang Z, Zhao B, Liu Y, Lin Y, Zheng J, Mungall C, Courtot M, Ruttenberg A, He Y (2016) Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res* 45:D347–D352
  61. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV (2015) AberOWL: a framework for ontology-based data access in biology. *BMC Bioinformatics* 16:1–9
  62. Pesce V, Tennison J, Mey L, Jonquet C, Toulet A, Aubin S, Zervas P, Pesce V, Tennison J, Mey L, Jonquet C, Toulet A, Aubin S, Zervas P (2018) A map of agri-food data standards. Technical Report, F1000Research, 7–177
  63. D'Aquin M, Baldassarre C, Gridinoc L, Angeletou S, Sabou M, Motta E (2007) Watson: a gateway for next generation semantic web applications. In: 6th international semantic web conference, ISWC'07, Poster & Demo Session, Busan, Korea, p 3
  64. Kitchenham B (2004) Procedures for performing systematic reviews. Technical report, TR/SE-0401, Keele University
  65. Gangemi A (2005) Ontology design patterns for semantic web content. In: Gil Y, Motta E, Benjamins VR, Musen MA (eds) 4th international semantic web conference, ISWC 2005. Springer, Galway, pp 262–276

66. de Melo G (2015) Lexvo.org: language-related information for the linguistic linked data cloud. *Sem Web* 6:8
67. Noy NF, Dorf M, Griffith NB, Nyulas C, Musen MA (2009) Harnessing the power of the community in a library of biomedical ontologies. In: Clark T, Luciano JS, Marshall MS, Prud'hommeaux E, Stephens S (eds) Workshop on semantic web applications in scientific discourse, SWASD'09, Washington DC, USA, p 11
68. Jonquet C, Musen MA, Shah NH (2010) Building a biomedical ontology recommender web service. *Biomed Sem* 1:S1
69. Nédellec C, Bossy R, Valsamou D, Ranoux M, Golik W, Sourdille P (2014) Information extraction from bibliography for marker-assisted selection in wheat. In: International conference on metadata and semantics research, MTSR'14. Springer, Karlsruhe, pp 301–313
70. Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front Physiol* 3:326. <https://doi.org/10.3389/fphys.2012.00326>



Data and text mining

# Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+

Andon Tchechmedjiev<sup>1,\*</sup>, Amine Abdaoui<sup>1</sup>, Vincent Emonet<sup>1</sup>, Soumia Melzi<sup>1</sup>, Jitendra Jonnagaddala<sup>2</sup> and Clement Jonquet<sup>1,3</sup>

<sup>1</sup>Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier & CNRS, Montpellier 34090, France, <sup>2</sup>Faculty of Medicine, University of New South Wales, Sydney, New South Wales 2052, Australia and <sup>3</sup>Center for Biomedical Informatics Research (BMIR), Stanford University, Stanford, California 94305, USA

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 10, 2017; revised on January 3, 2018; editorial decision on January 3, 2018; accepted on January 9, 2018

## Abstract

**Summary:** Second use of clinical data commonly involves annotating biomedical text with terminologies and ontologies. The National Center for Biomedical Ontology Annotator is a frequently used annotation service, originally designed for biomedical data, but not very suitable for clinical text annotation. In order to add new functionalities to the NCBO Annotator without hosting or modifying the original Web service, we have designed a proxy architecture that enables seamless extensions by pre-processing of the input text and parameters, and post processing of the annotations. We have then implemented enhanced functionalities for annotating and indexing free text such as: scoring, detection of context (negation, experimenter, temporality), new output formats and coarse-grained concept recognition (with UMLS Semantic Groups). In this paper, we present the NCBO Annotator+, a Web service which incorporates these new functionalities as well as a small set of evaluation results for concept recognition and clinical context detection on two standard evaluation tasks (Clef eHealth 2017, SemEval 2014).

**Availability and implementation:** The Annotator+ has been successfully integrated into the SIFR BioPortal platform—an implementation of NCBO BioPortal for French biomedical terminologies and ontologies—to annotate English text. A Web user interface is available for testing and ontology selection ([http://bioportal.lirmm.fr/ncbo\\_annotatorplus](http://bioportal.lirmm.fr/ncbo_annotatorplus)); however the Annotator+ is meant to be used through the Web service application programming interface ([http://services.bioportal.lirmm.fr/ncbo\\_annotatorplus](http://services.bioportal.lirmm.fr/ncbo_annotatorplus)). The code is openly available, and we also provide a Docker packaging to enable easy local deployment to process sensitive (e.g. clinical) data in-house (<https://github.com/sifrproject>).

**Contact:** andon.tchechmedjiev@lirmm.fr

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Semantic annotation of clinical data with standard medical terminologies/ontologies facilitates second use and translational data discoveries. Electronic Health Records often include unstructured

elements (free text) that contain valuable information for medical research (Meystre *et al.*, 2008). Researchers have developed systems to automatically detect clinical conditions and extract valuable knowledge in order to facilitate decision support (Rothman *et al.*, 2012),

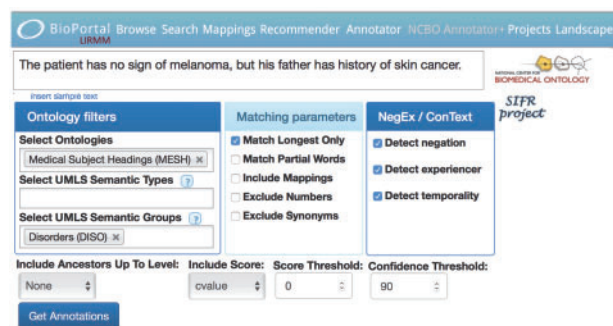


Fig. 1. User Interface of the NCBO Annotator+ Web service ([http://bioportal.lirmm.fr/ncbo\\_annotatorplus](http://bioportal.lirmm.fr/ncbo_annotatorplus)) illustrating new features. To reproduce this example with the Web service, use the URL: <https://goo.gl/BTrNzJ>

CLASS	filter	ONTOLOGY	CONTEXT	SCORE	NEGATION	EXPERIENTER	TEMPORALITY
Skin Neoplasms	MESH	...	history of <b>skin cancer</b> .	3.000	AFFIRMED	OTHER	HISTORICAL
Melanoma	MESH	...	sign of <b>melanoma</b> , but	3.322	NEGATED	PATIENT	RECENT

Format Results As:  To reproduce these results:

Fig. 2. Annotation results for the example sentence from Figure 1

the identification of patients (Liu *et al.*, 2013) and surveillance (Herasevich *et al.*, 2011). In 2009, the US National Center for Biomedical Ontologies released the NCBO Annotator (Jonquet *et al.*, 2009) within the BioPortal platform (Noy *et al.*, 2009), a publicly accessible and easily usable annotator Web service to process raw biomedical English text and identify ontology concepts. The annotation workflow is based on a highly efficient syntactic concept recognition tool [95% precision for diseases (Dai *et al.*, 2008)] that uses concept names and synonyms. The recognizer optionally allows to use names and synonyms of related concepts through semantic expansion [e.g. *is\_a* assertions and concept-to-concept mappings (Shah *et al.*, 2009)]. The NCBO Annotator has been widely adopted in the community and is one of the most actively used services from NCBO BioPortal, with a dictionary made from labels of 600+ ontologies. Yet, the Annotator lacks natural language processing capabilities (e.g. handling of morphological variants, disambiguation) required to improve the accuracy of annotations. Another limitation is the absence of scoring and of the contextualization of clinical text annotations, something it was never really designed for.

In the context of the Semantic Indexing of French biomedical Resources (SIFR) project, in which we have developed a French version of the Annotator, we have implemented some new features for French that we seamlessly ported to English through a proxy Web service called NCBO Annotator+. These new features include: annotation scoring, additional output formats (for evaluation and integration with standard clinical systems), clinical context detection (negation, experienter and temporality through the integration of the NegEx/ConText algorithm) and coarse-grained entity type annotations (with UMLS Semantic Groups, e.g. anatomy, disorders, devices). This article presents: (i) the proxy architecture and on how it enables the addition of new features, (ii) a performance evaluation of the NCBO Annotator+ on concept recognition tasks (death certificates and clinical notes) and on context detection (clinical notes only).

## 2 Materials and methods

Annotator+ is composed of a Web user interface in the SIFR BioPortal, and a proxy servlet to implement new features; it uses the

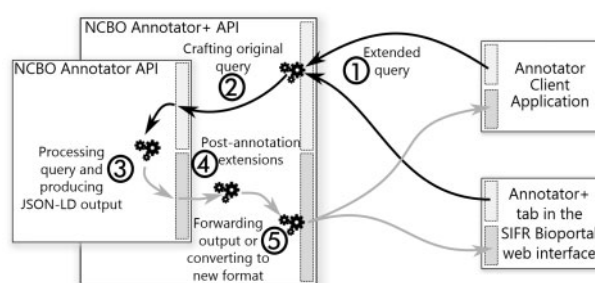


Fig. 3. NCBO Annotator+ proxy-like Web service architecture

NCBO BioPortal Annotator REST API in the backend. Figure 1 illustrates the Annotator+ interface with an example sentence (Restricted to the MESH and SNOWMED-CT vocabularies, filtered on the 'Disorder' UMLS Semantic Group, scored with a 90% relative threshold and with clinical context detection activated), while Figure 2 illustrates the resulting annotations.

### 2.1 Proxy Web service architecture

The NCBO Annotator is developed and maintained by the NCBO and does not easily support quick add-ons. To extend the NCBO Annotator without modifying the original application, we developed a proxy Web service architecture that can run independently and extend the service by pre-processing inputs and post-processing outputs. It works as follows (Fig. 3): (i) requests are sent to the proxy with extended parameters that are parsed to select/apply the additional features; (ii) a query is crafted for the original service without any extended parameters; (iii) the original NCBO Annotator processes the query and returns the results; (iv) the proxy retrieves annotations and applies post-processing/filtering (e.g. scoring); and finally, (v) the output is generated in the original format or in one of the new output formats from Annotator+. The proxy is implemented in a generic form that enables the querying of any NCBO-like annotator Web service. Indeed, we also use it for the French Annotator (Jonquet *et al.*, 2016a,b) and the AgroPortal Annotator, a similar Web service developed for agronomy (Jonquet *et al.*, 2016a).

### 2.2 New features

**Scoring.** During semantic indexing, annotations 'bring together' data elements and ontology concepts. Annotation scoring and ranking help to distinguish the most relevant annotations for a given element (e.g. a document, a clinical report) and when searching the original data. Typically, in information retrieval approaches, scoring is based on term frequency. We have implemented and evaluated a new scoring method for that purpose. By using a natural language processing term extraction measure called C-Value (Frantzi *et al.*, 2000), we were able to offer three scoring algorithms based on match frequencies that favour longer multi-word term annotations (higher scores) over shorter or single word annotation (Melzi *et al.*, 2014). We also added a mechanism to filter annotations by absolute score or in proportion (percentage) to the cumulative score distribution, to retrieve only the most relevant annotations (e.g. annotating with a threshold of 90% only retains the annotations with scores in the top 10% of the score distribution).

**New output formats.** NCBO Annotator supports XML and JSON-LD outputs. While JSON-LD is a recognized format, it is not sufficient for many annotation benchmarks and tasks, especially in the semantic Web and natural-language-processing communities.

Annotator+ adds support for standard (BRAT, RDF) and task-specific (e.g. CLEF eHealth) formats. RDF is the backbone language of the semantic Web and BRAT (<http://brat.nlplab.org>) is widely used for evaluation campaigns and for the production of annotated corpora. We also enriched the JSON-LD output with additional information (e.g. scores or clinical context).

**Clinical context.** For clinical text, the context of the annotated clinical conditions is crucial: Distinguishing between affirmed and negated occurrences (e.g. ‘no sign of metastasis’); whether a condition pertains to the patient or to others (e.g. ‘mother had breast cancer’); or temporality (i.e. if a condition is recent or historical. e.g. ‘history of poliovirus’). NegEx/ConText, is one of the best performing and fastest (open-source) algorithms for clinical context detection in English medical text (Harkema *et al.*, 2009). NegEx/ConText is based on lexical cues (trigger terms) that modify the default status of medical conditions appearing in their scope. For instance, by default the system considers a condition *affirmed*, and marks it as *negated* only if it appears under the scope of a trigger term. Each trigger term has a pre-defined scope either forward (e.g. ‘denies’) or backward (e.g. ‘is ruled out’), which ends by a colon or a termination term (e.g. ‘but’). We integrated this algorithm within the NCBO Annotator+ by post-processing the sentence in which an annotation appears. To our knowledge, this is the first implementation of a Web-based ConText-like system in a publicly accessible platform allowing non-experts in natural-language-processing to both annotate and contextualize medical conditions in clinical notes.

**Coarse-grained semantic annotation.** Recognizing broad entity types (e.g. gene, drug, disease) is a task of high interest for the BioNLP community. The 10 Semantic Groups (McCray *et al.*, 2001) are often used as coarse-grained groupings of the Unified Medical Language System (UMLS) Semantic Types (Bodenreider, 2004). Thanks to the capability of the NCBO Annotator to filter ontologies by Semantic Types, we have also added the capability to filter by Semantic Groups in Annotator+. This enables anyone to annotate free text and keep only certain broad types of annotations. For instance, a pharmacogenomics researcher doing a study may restrict the annotations to the types ‘disorders’ and ‘chemicals & drugs’ to investigate the effect of adverse drug reactions.

### 2.3 Evaluation protocol

We briefly report on the performance of the NCBO Annotator+ for: (i) annotating and contextualizing concepts in clinical text on the CLEF eHealth 2017 task 1 corpus (Névéol *et al.*, 2017), created for the automatic annotation of death certificates with ICD-10 codes; (ii) the SemEval 2015 Task 14.2 development corpus, created for the identification of biomedical concepts (i.e. names and identifiers in UMLS) and of clinical context features (we covered negation and experienter).

## 3 Results and discussion

This section provides: (i) benchmark results for concept recognition with the original NCBO Annotator and (ii) evaluation of the new features (negation & experienter detection only) of the Annotator+. The goal is both to provide additional performance evaluations to the community of the NCBO Annotator and to evaluate our own additions to the Annotator+. In 2017, we have participated to the CLEF eHealth 2017 Task 1 evaluation campaign, with the French/SIFR Annotator and the NCBO Annotator+. The campaign tackles the problem of information extraction (diagnostic coding) in written

**Table 1.** Evaluation for concept recognition (NCBO Annotator) and clinical context detection (Annotator+) expressed by Precision, Recall, F-measure, Accuracy

Task (Corpus)	P (%)	R (%)	F1 (%)	A (%)
Concept Recognition (CLEF eHealth)	69.1	51.4	58.9	
Concept Recognition (SemEval)	46.9	62.0	53.4	66.6
Negation Detection (SemEval)	87.0	88.9	88.0	89.3
Experienter Detection (SemEval)	52.9	70.4	60.4	52.7

death certificates, where the objective is to annotate each document with a set of relevant International Classification of Diseases, 10th revision (ICD-10) diagnostic codes. We have built a custom SKOS vocabulary (Simple Knowledge Organization System) from the dictionary of terms provided and uploaded it to the NCBO BioPortal (which also parses SKOS as input format). When annotating the death certificates with the NCBO Annotator, we obtained median results compared to the rest of the competitors [cf. Table 1 (Névéol *et al.*, 2017; Tchechmedjiev *et al.*, 2017)]; ahead of other knowledge-based systems but behind specifically tailored supervised learning systems. The results are encouraging considering that we have not customized the service in any way for the task. We acknowledge the better performance of supervised learning approaches, but claim that in the health domain, they are often not applicable for lack of training data.

For the evaluation of our integration of NegEx/ConText within the Annotator+, we used the SemEval 2015 corpus. For the task of concept recognition in the SemEval corpus, the NCBO Annotator obtained average scores, given that we performed no adaptation to the task (and we did not use the training data at all), the concept recognition accuracy is fair (66.6%). We did not have access to the test gold standard and thus cannot compare to other participants (we ran on the dev. corpus). For negation, Annotator+ obtained state-of-the-art performance (balanced weighted average performance) and for experienter detection, we obtain results that are not substantially lower than existing evaluations of ConText (Harkema *et al.*, 2009). These results confirm both the potential of the NCBO Annotator as a concept recognition service (never evaluated on standardized evaluation campaign tasks) and the nonreduced performance of NegEx/ConText when implemented in Annotator+.

## 4 Conclusion

We believe the NCBO Annotator+ offers a valuable framework to: (i) leverage an already performant service, which uses the biggest biomedical terms dictionary (600+ semantic resources including almost all UMLS and all the OBO Library ontologies); and (ii) improve the performance of this service on specific types of text such as in our case clinical notes. In the future, we will work on two important weaknesses of the service: disambiguation of annotations (too many polysemic terms decrease precision) and for clinical text mainly, cleaning and reformatting of the text (abbreviations, spelling mistakes, unconventional sentence structures, decrease recall). We working with the NCBO towards integrating some of this work directly into the NCBO Annotator.

## Acknowledgements

This work was achieved within the Semantic Indexing of French biomedical Resources (SIFR, [www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) and PractiKPharma project (<http://practikpharma.loria.fr>). We thank the US National Center for Biomedical

Ontology for their assistance with the NCBO Annotator. We also thank the CépiDC and CLEF eHealth 2017 organizers for their authorization to use the corpus, and the SemEval 2014 organizers.

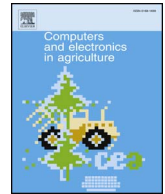
## Funding

This work is supported by the French National Research Agency within the PractiKPharma (grant ANR-15-CE23-0028) and SIFR (grant ANR-12-JS02-01001) projects as well as by the European H2020 Marie Curie actions (grant 701771), the University of Montpellier and the CNRS. We also thank the US National Center for Biomedical Ontology for their assistance with the NCBO Annotator.

*Conflict of Interest:* none declared.

## References

- Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, 267–270.
- Dai, M. et al. (2008) An Efficient Solution for Mapping Free Text to Ontology Terms. In: *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'08*. San Francisco, CA, USA.
- Frantzi, K. et al. (2000) Automatic recognition of multi-word terms: the C-value/NC-value method. *Digit. Libr.*, **3**, 115–130.
- Harkema, H. et al. (2009) ConText: an algorithm for determining negation, experimenter, and temporal status from clinical reports. *J. Biomed. Inf.*, **42**, 839–851.
- Herasevich, V. et al. (2011) Limiting ventilator-induced lung injury through individual electronic medical record surveillance. *Crit. Care Med.*, **39**, 34–39.
- Jonquet, C. et al. (2016a) Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. In: Jaiswal, P. and Hoehndorf, R. (eds) *7th International Conference on Biomedical Ontologies, ICBO'16, Demo Session, CEUR Workshop Proceedings*. Corvallis, Oregon, USA, p. 3.
- Jonquet, C. et al. (2016b) SIFR BioPortal: Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: *16th Journées Francophones d'Informatique Médicale JFIM'16*.
- Jonquet, C. et al. (2009) The open biomedical annotator. In: *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*. San Francisco, CA, USA, pp. 56–60.
- Liu, H. et al. (2013) An information extraction framework for cohort identification using electronic health records. In: *Proceedings of the American Medical Informatics Association Summits on Translational Science*, pp. 149–153.
- McCray, A.T. et al. (2001) Aggregating UMLS semantic types for reducing conceptual complexity. *Stud. Health Technol. Inf.*, **84**, 216.
- Melzi, S. et al. (2014) Scoring semantic annotations returned by the NCBO annotator. In: Paschke, A. et al. (eds) *7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14, CEUR Workshop Proceedings*. CEUR-WS.org, Berlin, Germany, p. 15.
- Meystre, S.M. et al. (2008) Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inf.*, **35**, 44.
- Névél, A. et al. (2017) CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. In: *CLEF 2017 Evaluation Labs and Workshop: Online working Notes, CEUR-WS, September, 2017*.
- Noy, N.F. et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, **37**, 170–173.
- Rothman, B. et al. (2012) Future of electronic health records: implications for decision support. *Mt. Sinai J. Med. A J. Transl. Pers. Med.*, **79**, 757–768.
- Shah, N.H. et al. (2009) Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, **10**(Suppl 9), S14. <http://doi.org/10.1186/1471-2105-10-S9-S14>.
- Tchechmedjiev, A. et al. (2017) ICD-10 coding of death certificates with the NCBO and SIFR Annotators at CLEF eHealth 2017. In: *CLEF 2017 Evaluation Labs and Workshop: Online working Notes, CEUR-WS, September, 2017*.



## Original papers

## AgroPortal: A vocabulary and ontology repository for agronomy

Clément Jonquet<sup>a,b,f,\*</sup>, Anne Toulet<sup>a,b</sup>, Elizabeth Arnaud<sup>c</sup>, Sophie Aubin<sup>d</sup>, Esther Dzalé Yeumo<sup>d</sup>, Vincent Emonet<sup>a</sup>, John Graybeal<sup>f</sup>, Marie-Angélique Laporte<sup>c</sup>, Mark A. Musen<sup>f</sup>, Valeria Pesce<sup>g</sup>, Pierre Larmande<sup>b,e</sup>

<sup>a</sup> Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier & CNRS, France

<sup>b</sup> Computational Biology Institute (IBC) of Montpellier, France

<sup>c</sup> Bioversity International, Montpellier, France

<sup>d</sup> INRA Versailles, France

<sup>e</sup> UMR DIADE, IRD Montpellier, France

<sup>f</sup> Center for BioMedical Informatics Research (BMIR), Stanford University, USA

<sup>g</sup> Global Forum on Agricultural Research (GFAR), Food and Agriculture Organization (FAO) of the United Nations, Rome, Italy



## ARTICLE INFO

## Keywords:

Ontologies  
Controlled vocabularies  
Knowledge organization systems or artifacts  
Ontology repository  
Metadata  
Mapping  
Recommendation  
Semantic annotation  
Agronomy  
Food  
Plant sciences  
Biodiversity

## ABSTRACT

Many vocabularies and ontologies are produced to represent and annotate agronomic data. However, those ontologies are spread out, in different formats, of different size, with different structures and from overlapping domains. Therefore, there is need for a common platform to receive and host them, align them, and enabling their use in agro-informatics applications. By reusing the National Center for Biomedical Ontologies (NCBO) BioPortal technology, we have designed AgroPortal, an ontology repository for the agronomy domain. The AgroPortal project re-uses the biomedical domain's semantic tools and insights to serve agronomy, but also food, plant, and biodiversity sciences. We offer a portal that features ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation; stores and exploits ontology alignments; and enables interoperation with the semantic web. The AgroPortal specifically satisfies requirements of the agronomy community in terms of ontology formats (e.g., SKOS vocabularies and trait dictionaries) and supported features (offering detailed metadata and advanced annotation capabilities). In this paper, we present our platform's content and features, including the additions to the original technology, as well as preliminary outputs of five driving agronomic use cases that participated in the design and orientation of the project to anchor it in the community. By building on the experience and existing technology acquired from the biomedical domain, we can present in AgroPortal a robust and feature-rich repository of great value for the agronomic domain.

## 1. Introduction

Agronomy, food, plant sciences, and biodiversity are complementary scientific disciplines that benefit from integrating the data they generate into meaningful information and interoperable knowledge. Undeniably, data integration and semantic interoperability enable new scientific discoveries through merging diverse datasets (Goble and Stevens, 2008). A key aspect in addressing semantic interoperability is the use of ontologies as a common and shared means to describe data, make them interoperable, and annotate them to build structured and formalized knowledge. Biomedicine has always been a leading domain encouraging semantic interoperability (Rubin et al., 2008). The domain has seen success stories such as the Gene Ontology (Ashburner et al., 2000), widely used to annotate genes and their products. And other disciplines have followed, developing among

others the Plant Ontology (Cooper et al., 2012), Crop Ontology (Shrestha et al., 2010), Environment Ontology (Buttigieg et al., 2013), and more recently, the Agronomy Ontology (Devare et al., 2016), TOP Thesaurus (Garnier et al., 2017), Food Ontology (Griffiths et al., 2016), the IC-FOODS initiative's ontologies (Musker et al., 2016), and the animal traits ontology (Hughes et al., 2014). Ontologies have opened the space to various types of semantic applications (Meng, 2012; Walls et al., 2014), to data integration (Wang et al., 2015), and to decision support (Lousteau-Cazalet et al., 2016). Semantic interoperability has been identified as a key issue for agronomy, and the use of ontologies declared a way to address it (Lehmann et al., 2012).

Communities engaged in agronomic research often need to access specific sets of ontologies for data annotation and integration. For instance, plant genomics produces a large quantity of data (annotated genomes), and ontologies are used to build databases to facilitate cross-

\* Corresponding author at: 161 Rue Ada, 34090 Montpellier, France.

E-mail address: [jonquet@lirmm.fr](mailto:jonquet@lirmm.fr) (C. Jonquet).

species comparisons (Jaiswal, 2011). More recently, the focus of many scientific challenges in plant breeding has switched from genetics to phenotyping, and standard traits/phenotypes vocabularies have become necessary to facilitate breeders' data integration and comparison. In parallel with very specific crop dictionaries (Shrestha et al., 2010), important organizations have produced large reference vocabularies such as Agrovoc (Food and Agriculture Organization) (Sachit Rajbhandari, 2012), the NAL Thesaurus (National Agricultural Library), and the CAB Thesaurus (Centre for Agricultural Bioscience International).<sup>1</sup> These thesauri are primarily used to index information resources and databases. As more vocabularies and ontologies<sup>2</sup> are produced in the domain, the greater the need to discover them, evaluate them, and manage their alignments (d'Aquin and Noy, 2012).

However, while great efforts have taken place in the biomedical domain to harmonize content (e.g., the *Unified Medical Language System* (UMLS), mostly for medical terminologies) (Bodenreider, 2004) and ontology design principles (e.g., the OBO Foundry, containing mostly biological and biomedical ontologies) (Smith et al., 2007), ontologies in agriculture are spread out around the web (or even unshared), in many different formats and artifact types, and with different structures. Agronomy (and its related domains such as food, plant sciences, and biodiversity) needs an one-stop shop, allowing users to identify and select ontologies for specific tasks, as well as offering generic services to exploit them in search, annotation or other scientific data management processes. The need is also for a community-oriented platform that will enable ontology developers and users to meet and discuss their respective opinions and wishes. This need was clearly expressed by stakeholders in various roles (developers, database maintainers, and researchers) across many community meetings, such as: 1st International Workshop for Semantics for Biodiversity in 2013 (<http://semantic-biodiversity.mpl.ird.fr>) (Larmande et al., 2013); the "Improving Semantics in Agriculture" workshop in 2015 (Baker et al., 2015); or several meetings of the Agricultural Data Interest Group (IGAD) of the Research Data Alliance.

These motivations prompted us to build a vocabulary and ontology repository to address these needs. In this paper, we present the AgroPortal project, a community effort started by the Montpellier scientific community to build an ontology repository for the agronomy domain. Our goal is to facilitate the adoption of metadata and semantics to facilitate open science in agronomy. By enabling straightforward use of agronomical ontologies, we let data managers and researchers focus on their tasks, without requiring them to deal with the complex engineering work needed for ontology management. AgroPortal offers a robust and reliable service to the community that provides ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation; stores and exploits ontology alignments; and enables interoperability with the semantic web. Our vision is to facilitate the integrated use of all vocabularies and ontologies related to agriculture, regardless of their source, format, or content type.

In order to capitalize on what is already available in other communities, we have reused the openly available NCBO BioPortal technology (<http://bioportal.bioontology.org>) (Noy et al., 2009; Whetzel et al., 2011) to build our ontology repository and services platform.

BioPortal was originally dedicated to health, biology and medicine and has some content related to agriculture, but the portal does not cover few of the facets of agronomy, food, plant sciences and biodiversity, let alone environment and animal sciences. Therefore, many in the agronomy community do not see themselves as users targeted by BioPortal. For instance, the Crop Ontology is listed on the NCBO BioPortal (along with other top-level plant-related ontologies), but is not currently fully accessible and described through this portal; none of the crop specific ontologies are available. In addition to its core repository of ontology mission, the NCBO technology also offers many applicable tools, including a mapping repository, an annotator, an ontology recommender, community support features, and an index of annotated data. All these services are reused and customized within AgroPortal to benefit its target user community.<sup>3</sup> Furthermore, our vision was to adopt, as the NCBO did, an open and generic approach where users can easily participate to the platform, upload content, and comment on others' content (ontologies, concepts, mappings, and projects). As explained below, we determined that the NCBO technology (Whetzel and Team, 2013) implemented the greatest number of our required features, while recognizing the technical challenges of adopting such a various and complex software.

In the following sections, we offer extensive descriptions of AgroPortal's features. We will focus on how they address community requirements expressed within five agronomic driving use cases involving important research organizations in agriculture such as Bioversity International (CGIAR), French INRA, and United Nations FAO. The rest of the paper is organized as follows: In Section 2, we review related work in ontology repositories in relation to our domain of interest. Section 3 describes the requirements of AgroPortal's initial five driving agronomic use cases. Section 4 presents our platform by extensively describing its content, as well as its features (both inherited from the NCBO BioPortal, and added by us). Section 5 analyzes how our initial five driving use case results benefit from AgroPortal. Finally, Section 6 provides a discussion of the contributions of AgroPortal, and Section 7 presents our conclusions.

## 2. Background and related work

With the growing number of developed ontologies, ontology libraries and repositories have been of interest in the semantic web community. Ding and Fensel (2001) presented in 2001 a review of ontology libraries that introduced the notion of "library." Then Hartman et al. Baclawski and Schneider (2009) introduced the concept of ontology repository, with advanced features such as search, metadata management, visualization, personalization, and mappings. By the end of the 2000's, the Open Ontology Repository Initiative (Baclawski and Schneider, 2009) was a collaborative effort to develop a federated infrastructure of ontology repositories.<sup>4</sup> d'Aquin and Noy (2012) provided the latest review of ontology repositories in 2012.

In the biomedical or agronomic domains there are several standards or knowledge organization systems libraries (or registries) such as FAIRSharing (<http://fairsharing.org>) Sansone et al., 2012, the FAO's VEST Registry (<http://aims.fao.org/vest-registry>), and the agINFRA linked data vocabularies (vocabularies.aginfra.eu) (Pesce et al., 2013). They usually register ontologies and provide a few metadata attributes about them. However, because they are registries not focused on vocabularies and ontologies, they do not support the level of features that an ontology repository offers. In the biomedical domain, the OBO Foundry (Smith et al., 2007) is a reference community effort to help the

<sup>1</sup> <http://aims.fao.org/agrovoc>, <https://agclass.nal.usda.gov> and <http://www.cabi.org/cabthesaurus>

<sup>2</sup> In this paper, we often use the word "ontologies" or "vocabularies and ontologies" to include ontologies, vocabularies, terminologies, taxonomies and dictionaries. We acknowledge the differences (not discussed here) in all these types of Knowledge Organization Systems (KOS) or knowledge artifacts. The reader may refer to McGuinness's discussion (McGuinness, 2003). While being an "ontology repository", AgroPortal handles all these artifact types, if they are compatibly formatted. While AgroPortal thereby enables horizontal use of these artifact types with common user interface and application programming interface, it does not leverage the full power of ontologies (e.g., reasoning), instead map all the imported artifact types to a "common simplified model."

<sup>3</sup> Except the "NCBO Resource Index" component, a database of 50+ biomedical resources indexed with ontology concepts (Jonquet et al., 2011) that we have not reused in AgroPortal because we work with the AgroLD use case to fulfill the mission of interconnecting ontologies and data.

<sup>4</sup> At that time, the effort already reused the NCBO technology that was open source, but not yet packaged in an appliance as it is today.

biomedical and biological communities build their ontologies with an enforcement of design and reuse principles that have made the effort very successful. The OBO Foundry web application is not an ontology repository per se, but relies on other applications that pull their data from the foundry, such as the NCBO BioPortal (Noy et al., 2009), OntoBee (Xiang et al., 2011), the EBI Ontology Lookup Service (Côté et al., 2006) and more recently AberOWL (Hoehndorf et al., 2015). In addition, there exist other ontology libraries and repository efforts unrelated to biomedicine, such as the Linked Open Vocabularies (Vandenbussche et al., 2014), OntoHub (Till et al., 2014), and the Marine Metadata Initiative's Ontology Registry and Repository (Graybeal et al., 2012).

Some of the known ontology repositories could be candidates for hosting agronomical ontologies. However, all of these portals either are too generic, or too narrowly focused on health, biology or medicine, and despite any existing thematic overlaps, scientific lineage and partnerships, we have identified, as established in Section 1, the crucial need for a community platform where agronomy will actually be the primary focus. To avoid building a new ontology repository from scratch, we have considered which of the previous technologies are reusable. While all of them are open source, only the NCBO BioPortal<sup>5</sup> and OLS<sup>6</sup> are really meant for reuse, both in their construction, and in their provided documentation. At the start of our project in 2014, AberOWL was not yet published and OntoBee (released in 2011) had not changed between 2011 and 2014 (a new release took place thereafter (Ong et al., 2016)). Of the two candidate technologies at the time, we will show, that the NCBO technology was the one implementing highest number of requested features.<sup>7</sup>

In the biomedical domain, the NCBO BioPortal is a well-known open repository for biomedical ontologies originally spread out over the web and in different formats. There are 656 public ontologies in this collection as of Nov. 2017, including relevant ones for agronomy. By using the portal's features, users can browse, search, visualize and comment on ontologies both interactively through a user web interface, and programmatically via web services. Within BioPortal, ontologies are used to develop an annotation workflow (Jonquet et al., 2009) that indexes several biomedical text and data resources using the knowledge formalized in ontologies to provide semantic search features that enhance information retrieval experience (Jonquet et al., 2011). The NCBO BioPortal functionalities have been progressively extended in the last 12 years, and the platform has adopted semantic web technologies (e.g., ontologies, mappings, metadata, notes, and projects are stored in an RDF<sup>8</sup> triple store) (Salvadores et al., 2013).

An important aspect is that NCBO technology (Whetzel and Team, 2013) is domain-independent and open source. A BioPortal virtual appliance<sup>9</sup> is available as a server machine embedding the complete code and deployment environment, allowing anyone to set up a local ontology repository and customize it. It is important to note that the NCBO Virtual Appliance has been quite regularly reused by organizations which needed to use services like the NCBO Annotator but, for privacy reason, had to process the data in house. Via the Virtual Appliance, NCBO technology has already been adopted for different

ontology repositories in related domains and was also chosen as foundational software of the Open Ontology Repository Initiative (Baclawski and Schneider, 2009). The Marine Metadata Interoperability Ontology Registry and Repository (Rueda et al., 2009) used it as its backend storage system for over 10 years, and the Earth Sciences Information Partnership earth and environmental semantic portal (Pouchard and Huhns, 2012) was deployed several years ago. More recently, the SIFR BioPortal (Jonquet et al., 2016) prototype was created at University of Montpellier to build a French Annotator and experiment multilingual issues in BioPortal (Jonquet et al., 2015). Although we cannot know all the applications of other technologies, the visibly frequent reuse of the NCBO technology definitively confirmed it was our best candidate. There are two other major motivations for AgroPortal to reuse the products of biomedicine: (i) to avoid re-developing tools that have already been designed and extensively used and contribute to long term support of the commonly used technology; and (ii) to offer the same tools, services and formats to both communities, to facilitate the interface and interaction between their domains. This alignment will enhance both technical reuse (for example, enabling queries to either system with the same code), and semantic reuse (knowing the same semantic capabilities and practices apply to both sets of ontologies).

More specifically to the plant domain, the Crop Ontology web application ([www.cropontology.org](http://www.cropontology.org)) (Matteis et al., 2013) publishes online sets of ontologies and dictionaries required for describing crop germplasm, traits and evaluation trials. As of Nov. 2017, it contains 28 crop-specific phenotype and trait ontologies, in addition to ontologies related to the crop germplasm domain. Besides its role as a repository, the Crop Ontology web application offers community-oriented features such as an CSV template (TDv5) for trait submission, and addition and filtering of new terms. A web Application Programming Interface (API) provides all necessary services to third party users like the Global Evaluation Trials Database, currently storing 35,000 trial records. Efforts have been made to structure and formalize the crop-specific ontologies following semantic web standards (using the Web Ontology Language (OWL)), as well as offering collaborative ontology enrichment and annotation features. The current Crop Ontology web application facilitates the ontology-engineering life cycle (Noy et al., 2010), starting with collaborative construction, publishing, use and modification. However, it would require important improvements such as: versioning, community features, multilingual aspects, visualization, data annotation, and mapping services. For instance, it is important to support the alignment (or mapping) of terms within and across different ontologies both within the Crop Ontology itself (in different crop branch) and with other top level ontologies commonly used in plant biology, like the Plant Ontology, Plant Trait Ontology, Plant Environment Ontology, Plant Stress Ontology all maintained and extended within the Planteome project (Jaiswal et al., 2016).

The Planteome platform ([www.planteome.org](http://www.planteome.org)) is reusing the Gene Ontology project AmiGO technology (Carbon et al., 2009) to build a database of searchable and browsable annotations for plant traits, phenotypes, diseases, genomes, gene expression data across a wide range of plant species. The project focuses on developing reference ontologies for plant and on integrating annotated data within the platform. Their objective is slightly different than AgroPortal's objective, and the scope is not as large as the one we envision for AgroPortal.

### 3. Driving agronomic use cases requirements

The AgroPortal project was originally driven by five agronomic use cases that were the principal sources of ontologies and vocabularies. In this section, we present their requirements in terms of ontology repository functionalities – summarized in Table 1. The results for each use case will be presented in Section 5.

<sup>5</sup> The technology has always been open source, and the appliance has been made available since 2011. However, the product became concretely and easily reusable after BioPortal v4.0 end of 2013.

<sup>6</sup> The technology has always been open source but some significant changes (e.g., the parsing of OWL) facilitating the reuse of the technology for other portals were done with OLS 3.0 released in December 2015.

<sup>7</sup> It is beyond the scope of this paper to draw a complete comparison of ontology portals. The reader may refer to d'Aquin and Noy (2012).

<sup>8</sup> The Resource Description Framework (RDF) is the W3C language to described data. It is the backbone of the semantic web. SPARQL is the corresponding query language. By adopting RDF as the underlying format, AgroPortal can easily make its data available as linked open data and queryable through a public SPARQL endpoint. To illustrate this, the reader may consult the Link Open Data cloud diagram (<http://lod-cloud.net>) that since 2017 includes ontologies imported from the NCBO BioPortal (most of the Life Sciences section).

<sup>9</sup> [www.bioontology.org/wiki/index.php/Category:NCBO\\_Virtual\\_Appliance](http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance)

**Table 1**  
Summary of agronomic use case requirements for AgroPortal.

#	Requirement	Use case	Example
1	One-stop-shop to store, browse, search, visualize agronomical ontologies	LovInra VEST	Facilitate the adoption of semantic web standards by INRA's scientist, with a focus on agriculture The registry targets specifically the agriculture community and requires content-based services. The organization of ontologies by group and categories is also necessary
2	Unique ontology access point and application programming interface (API) to ontologies	AgroLD VEST LovInra, VEST	Automatically retrieve the most recent version of ontologies currently hosted either on OBO Foundry or Cropontology.org. At the beginning of the project, a SPARQL endpoint for ontologies was also needed Access point to automatically obtain metadata about all the ontologies INRA's researchers and VEST users need to upload their resources to a platform themselves
3	Directly accessible to scientists to upload their ontologies or vocabularies	AgroLD LovInra, Crop Ontology LovInra	Annotate text data from database fields to create RDF triples Identify plant phenotypes in text descriptions
4	Ontology-based annotation service	AgroLD LovInra, Crop Ontology LovInra	INRA's develop different type of knowledge organization systems include: ontologies (AFEO, Biorefinery, OntoBiotope) but also thesauri (AnAEE, GACS)
5	Handle different level of semantic description and the corresponding standard formats (SKOS and OWL)	VEST ALL	Many resources in agronomy are in SKOS format. All use cases have expressed the need to have a place to store, describe and retrieve alignments
6	Store and retrieve mappings between ontologies	AgroLD Others	Publish AgroLD mapping annotations to reference ontologies such as SIO, EDAM, PO Reference thesauri like Agrovoc have adopted linked open data practices and offer mappings to multiple semantic web resources (not necessarily ontologies)
7	Store mappings between ontologies and external resources	ALL	All use cases have expressed the need to automatically align ontologies one another
8	Automatically generate mappings between ontologies	AgroLD	Identify AgroLD data elements when browsing ontologies in AgroPortal.
9	Query and search annotated data from ontologies	WDI	Visualize and use only the 22 vocabularies identified by the WDI working group
10	Offer a unique sub-endpoint specific to a community or group	LovInra Crop Ontology	Clearly identify resources (co-)developed by INRA's researchers Handle as a collection the Crop Ontology project, which is composed of multiple crop-specific trait ontologies. Possible alternative to cropontology.org
11	Provide rich metadata description for ontologies (using semantic web standards)	WDI LovInra	Clearly describe access rights and license information for ontologies Clearly describe the type of resources (ontology, thesaurus, vocabulary, etc.) and their format and syntax
12	Get community feedback	VEST WDI Crop Ontology VEST	Facilitate an automatic interconnection with VEST, including aligning the metadata fields Inform the community about the WDI guidelines and get their feedback on the selected ontologies Offer breeders a way to suggest new trait and comment existing ones Enable a large community of "standard" developers to provide feedback and comments on the use (or non-use) of ontologies and vocabularies in AgroPortal
13	Multilingual ontology support	VEST, Others Others	Increasingly vocabularies have labels in different languages (e.g., Agrovoc, GACS, NALt). Distinguish between these labels in lexical-based services (search, annotation) IRSTEA develops vocabularies only in French
14	Dereference URIs for ontologies	LovInra, Crop Ontology	When opening in a web browser a URI created by INRA or CO, display the corresponding class or property page
15	Mechanism to identify and select the relevant ontologies for a given task	LovInra, VEST	Facilitate the identification of relevant agronomical ontologies for non-experts
16	Enable private access to ontologies during working and/or development phases	LovInra	Access and test the AnAEE Thesaurus or GCAS before they release; work on certain versions of OntoBiotope not public in OpenMinted project
17	Export ontologies in different formats, including downgrading them to CSV	Crop Ontology	Breeders may need simpler formats, as they may not be able to use advanced semantic web formats
18	Store the project/ontology relationships	VEST, AgBioData	Select and maintain a list of ontologies used by model organism databases

### 3.1. Agronomic Linked Data (AgroLD)

Agronomic research aims to effectively improve crop production through sustainable methods. To this end, there is an urgent need to integrate data at different scales (e.g., genomics, proteomics and phenomics). However, available agronomical information is highly distributed and diverse. Semantic web technology offers a remedy to the fragmentation of potentially useful information on the web by improving data integration and machine interoperability (Schmachtenberg et al., 2014). This has been often illustrated in data integration and knowledge management in the biomedical domain (Belleau et al., 2008; Jonquet et al., 2011; Jupp et al., 2014; Groth et al., 2014). To further build on this line of research in agronomy, we have developed the Agronomic Linked Data knowledge base ([www.agrold.org](http://www.agrold.org)) (Venkatesan et al., 2015). Launched in May 2015, it serves as a platform to consolidate distributed information and facilitate formulation of research hypotheses. AgroLD offers information on genes, proteins, Gene Ontology Associations, homology predictions, metabolic pathways, plant traits, and germplasm, on the following species: rice, wheat, arabidopsis, sorghum and maize. We provide integrated

agronomic data, as well as the infrastructure to aid domain experts answering relevant biological questions (for example, "identify wheat proteins that are involved in root development"). AgroLD relies on RDF and SPARQL technologies for information modelling and retrieval, and uses OpenLink Virtuoso (version 7.1) triple store. Database contents were parsed and converted into RDF using a semi-automated pipeline implemented in Python (<https://github.com/SouthGreenPlatform/AgroLD>).

The conceptual framework for knowledge in AgroLD is based on well-established ontologies in plant sciences such as Gene Ontology, Sequence Ontology, Plant Ontology, Crop Ontology and Plant Environment Ontology. AgroLD needs a dedicated application programming interface to these ontologies, as well as a means to annotate database fields (header and values) with ontology concepts. In addition, it requires a system to store mappings annotations between key entities in the AgroLD knowledge base and reference ontologies. In the long-term vision for AgroPortal and AgroLD, the former might be an entry point to the knowledge stored in AgroLD, enabling users to easily query and locate data annotated with ontologies.

### 3.2. RDA Wheat Data Interoperability (WDI) working group

Wheat is a major source of calories and protein, especially for consumers in developing countries, and thus plays an important socio-economical role. The International Wheat Initiative ([www.wheatinitiative.org](http://www.wheatinitiative.org)) has identified easy access and interoperability of all wheat related data as a top priority, to make the best possible use of genetic, genomic and phenotypic data in fundamental and applied wheat science. For example, the identification of causative genes for an important agronomic trait is key to effective marker-assisted breeding and reverse genetics. It requires integrating information from many different sources such as gene function annotations, biochemical pathways, gene expression data, as well as comparative information from related organisms, gene knock-out and the scientific literature (Hassani-Pak et al., 2013). However, the disparate nature of the formats and vocabularies used to represent and describe the data has resulted in a lack of interoperability.

The Wheat Data Interoperability working group was created in March 2014 within the frame of the Research Data Alliance (<https://rd-alliance.org>) and under the umbrella of the International Wheat Initiative, in order to provide a common framework for describing, linking and publishing wheat data with respect to existing open standards. The working group conducted a survey to identify and describe the most relevant vocabularies and ontologies for data description and annotation in the wheat domain (Dzalé-Yeumo et al., 2017). For some data types like DNA sequence variations, genome annotations, and gene expressions, the survey showed good consensus regarding data exchange formats. However, the survey did not show good consensus about data exchange formats and data description practices for phenotypes and germplasm, suggesting the need for harmonization and standardization.

Finally, this group identified 22 relevant vocabularies and ontologies for which, beyond the consensus issue, other problems were identified: (i) format and location heterogeneity: ontology formats included OBO format, OWL, and even SKOS (or SKOS-XL); (ii) heterogeneity: these ontology coverages ranged from describing generic experimental crop study (e.g., Crop Research), to narrow wheat-related topics (Wheat Trait, Wheat Anatomy and Development), to top-level concepts in biomedicine (BioTop). The need to offer a dedicated repository of linked vocabularies and ontologies relevant for wheat having been identified, the NCBO technology was seen as a likely tool to address this needs and desired features.

### 3.3. INRA Linked Open Vocabularies (LovInra)

What does a specialist in cattle developmental biology really need to easily identify, evaluate and exploit a few potential vocabularies of interest? Whether familiar with semantics technology or not, she needs a place that reflects her scientific environment and community, where those with similar concerns can share comments and content. As an example, INRA develops models to predict feed efficiency and meat quality for beef production, using experimental data collected during decades at INRA and externally. To meet the challenge of data integration, INRA developed the Animal Trait Ontology for Livestock (ATOL). In part thanks to AgroPortal, ATOL developers have identified the Animal Disease Ontology (ADO), developed by another team at INRA, as a possible resource to expand the perimeter of actionable data. This raised the question: How many complementary or competing resources to ATOL exist?

With this vision in mind, LovInra is a service offered by the French National Institute for Agricultural Research (INRA) Scientific and Technical Information department to identify and evaluate knowledge organization sources produced by INRA's scientists, so that the agricultural community and possibly a larger public can benefit from them. Many such resources developed within specific projects remain unknown to the research community despite their value. They are often

developed by subject matter experts who are not semantic experts, and who often do not have the resources (knowledge, time, or money) to share their results. Further, they span multiple semantic levels, from simple lexical descriptions, to hierarchies, to complex semantic relations. To achieve this goal, the vocabularies must be published with respect to open standards and linked to other existing resources. INRA adopted the semantic web's practices and standards (RDF, SKOS, OWL, SPARQL) to enable the methodological and technical practices needed by INRA's scientists to standardize, document and publish the vocabularies created in their projects. Examples of INRA's projects developing vocabularies or ontologies includes: (i) the AnAAE Thesaurus for the semantic description of the study of continental ecosystems developed by the AnaEE-France infrastructure;<sup>10</sup> (ii) the OntoBiotope ontology of microorganism habitats used collaboratively in multiple projects such as OpenMinted as well as for the BioNLP shared tasks; (iii) the Agri-Food Experiment Ontology (AFEO) ontology network which cover various viticultural practices, and winemaking products and operations.

Beyond its evaluation and standardization role, LovInra also serves to assign, deference, and provide programmatic access to INRA URIs (for example, <http://opendata.inra.fr/ms20/Observation>), using its triple store and web interface (<http://lovinra.inra.fr>). Although the current service, which includes description of resource metadata and direct access to source files, is necessary for internal use, it does not meet external dissemination objectives. In addition, the LovInra registry does not support any content-based features, such as searching, browsing, visualizing, mappings and annotation. We see AgroPortal as a possible solution to the entire range of INRA's unmet semantic needs above, complementing the services already provided by LovInra.

### 3.4. The Crop Ontology project

Communities engaged in germplasm evaluation trials need to access specific sets of ontologies for plant data annotation and integration. The Crop Ontology project ([www.cropontology.org](http://www.cropontology.org)) (Shrestha et al., 2010) of the Integrated Breeding Platform (IBP) is AgroPortal's fourth use case. The main goals of this project are: (i) to publish online fully documented lists of breeding traits and standard variables used for producing standard field books and (ii) to support data analysis and integration of genetic and phenotypic data through harmonized breeders' data annotation (Shrestha et al., 2012). Crop breeders, data managers, modelers, and computer scientists created a community of practice to discuss their variables, methods and scales of measurement, and field books. They seek to develop the most complete crop-specific trait ontologies according to the Crop Ontology template and guidelines.

The Crop Ontology website, released in 2010, provides 28 crop-specific trait ontologies, in addition to ontologies describing germplasm material and evaluation trials. The website publishes each crop-specific trait ontology online, making it available for download from the user interface or through an API in various formats: CSV, OBO, RDF/SKOS. Partners like the Oat Global, the US Department of Agriculture (USDA), INRA and the Polish Genomic Network have uploaded ontologies.<sup>11</sup> The project requires a specific dedicated infrastructure that deals with the adopted multi-trait ontologies approach, and supports search and versioning of ontologies. Plus, the Crop Ontology breeders need an interface to suggest new crop traits (i.e., new terms in the trait ontologies) and simple formats (such as CSV) to export the "trait dictionary" locally.

<sup>10</sup> Analysis and Experimentation on Ecosystems is European research infrastructure dedicated to the experimental manipulation of managed and unmanaged terrestrial and aquatic ecosystems ([www.anaee.com](http://www.anaee.com)).

<sup>11</sup> In addition, the Crop Ontology is used by several third-party projects like the Next Generation Breeding (Nextgen) databases, the Integrated Breeding Platform's breeding management system, and the global repository of the Agricultural Trials or EU-SOL.

### 3.5. GODAN Map of Agri-Food Data Standards

Recently, a new project under the umbrella of the GODAN<sup>12</sup> initiative called *GODAN Action* identified as one of its outputs a global map of standards used for exchanging data in the field of food and agriculture. To avoid duplicating effort, and to reuse previous community work, the project reviewed possible sources of standards that could be integrated. Two existing suitable platforms were identified: the FAO Agricultural Information Management Standards VEST Registry (<http://aims.fao.org/vest-registry> – now merged inside the new Map of Standards presented Section 5.5) and the then-new AgroPortal project.

The VEST Registry, created by FAO in 2011, was a metadata catalog of around 200 knowledge organization sources and tools. It had a broader coverage than the AgroPortal in two facets, knowledge types and domains. (i) Types of vocabularies or standards covered: the VEST Registry covered all types of knowledge artifacts, not just vocabularies or ontologies formally defined in RDFS, OWL, SKOS, or OBO. For instance, the VEST registry would cover data exchange format specification defined in XML or text description. (ii) Domain coverage: Besides standards used specifically for food and agriculture data, the directory included resources used in neighboring disciplines (like climate and environment, sciences). The VEST Registry was conceived as a metadata catalog, providing descriptions and categorization of standards and linking to the original website or download of the standard, but it did not exploit the content of the vocabularies or ontologies, only their metadata descriptions. It did not support any alignment between the sources either. To interconnect the VEST and AgroPortal, rich and unambiguous metadata would be crucial, as well as good classification of resources per categories and types.

### 3.6. Other requirements identified

In addition to these five first driving use cases, other projects or organizations have identified AgroPortal as a relevant application to host, share and serve their ontologies:

IRSTEA's projects, such as the French Crop Usage thesaurus about crops cultivated in France, and the French Agroecology Knowledge Management ontology for design innovative crop systems. These two projects produce ontologies only in French and needed a host for their work.

The Agrovoc thesaurus (Sachit Rajbhandari, 2012), which is the most worldwide used multilingual vocabulary developed by FAO. Agrovoc contains more than 32 K concepts covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. Agrovoc Linked Open Data version contains multiple mappings to other vocabularies or resources that a resource hosting Agrovoc must incorporate.

The Consortium of Agricultural Biological Databases ([www.agbiodata.org](http://www.agbiodata.org)), a group of database developers and curators maintaining model organism databases. The group wants to identify which databases use which ontologies, and recommend a list of ontologies based on that information.

## 4. A portal for agronomic related ontologies

In 2014, the *Computational Biology Institute of Montpellier* project identified the need for an ontology-based annotation service for the AgroLD and Crop Ontology use cases above. This large bioinformatics project in France had a specific plant/agronomy data work package. In parallel, we started reusing NCBO technology (Whetzel and Team,

2013) in the context of the SIFR<sup>13</sup> project, in which we develop a French version of the Annotator (Jonquet et al., 2016). We then implemented a connector to BioPortal within WebSmash (an open environment for matching complex schemas from many heterogeneous data sources (Coletta et al., 2012) enabling calls either to the NCBO Annotator web service, or any other NCBO-based Annotator (Castanier et al., 2014). Once we had a portal prototype hosting a few specific ontologies, interest in it grew when we presented it to several interlocutors (for examples, Bioversity International, INRA, IRD, CIRAD, FAO, RDA, Planteome). Driven additionally by the other use cases presented in Section 3, we extended our reuse of the NCBO technology to the full stack, and publish it under the brand AgroPortal.

We now have an advanced prototype platform (illustrated in figures on following pages) whose latest version v1.4 was released in July 2017 at <http://agroportal.lirmm.fr>.<sup>14</sup> The platform currently hosts 77 ontologies (Table 2), with more than 2/3 of them not present in any similar ontology repository (like NCBO BioPortal), and 11 private ontologies. We have identified 93 other candidate ontologies (Table 3) and we work daily to import new ones while involving/informing the original ontology developers. The platform already has more than 90 registered users. For an overview of AgroPortal ontology analytics, see Fig. 5 (Annex).

### 4.1. Ontology organization and sources

Developers generally upload their ontologies when they think the ontologies have reached a sufficient maturity and relevance to make them publicly available. Sometime, like in the AnaEE thesaurus, or OntoBiotope, developers use/used the portal as a staging location before the ontology goes public. If the initiative comes from our side, we usually always interact with the developers before importing any new resources: the original ontology developers always stay the only authority for the ontologies in the portal. Because of the features offered by AgroPortal (Sections 4.2 and 4.3), we think it is reasonable to incorporate ontologies that are already listed on other platforms (OBO Foundry, FAIRSharing, VEST registry, or LovInra). However, in those cases we follow these practices:

Developers can configure the entry in AgroPortal to automatically pull new version of ontologies. We synchronize the ontology in AgroPortal with the one at the original location via a nightly update<sup>15</sup> so the latest version is always available. For instance, all the ontologies in the OBO-FOUNDRY group are systematically updated using their PURL (e.g., for the Plant Ontology: <http://purl.obolibrary.org/obo/po.owl>).

We always inform the ontology developers of their ontology publication on AgroPortal if they did not submit their ontology directly, and offer them to claim administration role on the ontology if desired. While we often edit ontology descriptions, we ask the ontology developers to validate our edits and complete them.

We try to avoid duplicating ontologies already hosted in the NCBO BioPortal, unless required by a specific use case. Of course, overlap exists between our domain of interest and biomedicine. Our general approach is to let ontology developers decide if their ontology should be incorporated in the AgroPortal while it is already in the NCBO BioPortal. The long-term vision for AgroPortal and BioPortal is an interconnected network of “bioportals” that will enable easy access to ontologies for anyone independently from where they are hosted and that could extend to ontology repository types beyond the NCBO technology.

<sup>13</sup> Semantic Indexing of French Biomedical Data Resources (SIFR) project - <http://www.lirmm.fr/sifr>.

<sup>14</sup> <https://github.com/agroportal/documentation/wiki/Release-notes>

<sup>15</sup> Except for three ontologies (GO, BIOREFINERY & TRANSMAT) that are updated only weekly for scalability reasons.

<sup>12</sup> Global Open Data for Agriculture and Nutrition: <http://www.godan.info>.

**Table 2**

Examples of ontologies uploaded in AgroPortal. Acronyms in parenthesis are the identifier on AgroPortal e.g., <http://agroportal.lirmm.fr/ontologies/AEO> has the acronym AEO (Size = approximate number of classes or concepts).

Title	Format	Source	Group	Size
IBP rice trait ontology (CO_320)	OWL	cropontology.org	CROP, AGBIODATA, AGROLD	~2K
IBP wheat trait ontology (CO_321)	OWL	cropontology.org	CROP, AGBIODATA, AGROLD, WHEAT	~1K
IBP wheat anatomy & development ontology (CO_121)	OBO	cropontology.org	CROP, WHEAT	~80
IBP crop research (CO_715)	OBO	cropontology.org	CROP, AGBIODATA, WHEAT	~250
Multi-crop passport ontology (CO_020)	OBO	cropontology.org	CROP	~90
Biorefinery (BIOREFINERY)	OWL	Inra	LOVINRA, WHEAT, AGBIODATA	~300
Matter transfer (TRANSMAT)	OWL	Inra	LOVINRA, WHEAT, AGBIODATA	~1.1 K
Plant ontology (PO)	OWL	OBO Foundry	OBOF, AGROLD, WHEAT, AGBIODATA	~2K
Plant trait ontology (TO)	OWL	OBO Foundry	OBOF, AGROLD, WHEAT, AGBIODATA	~4.4 K
Durum wheat (DURUM_WHEAT)	OWL	Inra	LOVINRA	~130
Agricultural experiments (AEO)	OWL	Inra	LOVINRA	~60
Environment ontology (ENVO)	OWL	OBO Foundry	WHEAT, OBOF	~6.3 K
NCBI organismal classification (NCBITAXON)	RRF	UMLS	WHEAT, AGROLD	~900 K
AnaEE thesaurus (ANAEETHES)	SKOS	Inra	LOVINRA	~3.3 K
French crop usage (CROPUSAGE)	SKOS	Irstea	None	~300
Agrovoc (AGROVOC)	SKOS	FAO (UN)	WHEAT, AGBIODATA	~32 K
Food ontology (FOODON)	OWL	OBO Foundry	OBOF	~10 K
National agricultural library thesaurus (NALT)	SKOS	NAL (USDA)	WHEAT, AGBIODATA	~67 K
Global agricultural concept scheme (GACS)	SKOS	FAO-NAL-CABI	None	~580 K
Agronomy ontology	OWL	CGIAR	OBOF	~430
Biological collections ontology	OWL	OBO Foundry	OBOF	~160
Flora phenotype ontology	OWL	AberOWL	None	~28 K

**Table 3**

Selection of candidate ontologies of interest for the agronomic community, not present in the NCBO BioPortal.

Title	Organization or source
CAB thesaurus	CABI
Chinese agricultural thesaurus	CAAS
Wine ontology	INRA
Oat, Barley, Brachiaria, Potato (etc.) trait ontologies	Crop Ontology
Plant disease ontology	INRA
Agriculture activity ontology	CAVOC
Agriculture and forestry ontology	Univ. of Helsinki
IC-FOODS ontologies (~10)	UC Davis
agINFRA soil vocabulary	FAO, GFAR
Plant-pathogen interactions ontology	CBGP
Plant phenology ontology	OBO Foundry
Thesaurus of plant characteristics	CEFE
Livestock product trait ontology	Iowa State Univ.
Livestock breed ontology	Iowa State Univ.

Within AgroPortal, each time an ontology is uploaded into the portal, it is assigned a group and/or category. Groups associate ontologies from the same project or organization, for better identification of the provenance. We have created a group for each use case, except the fifth one that is not a source of ontologies, and another one for the OBO Foundry. For each group we have deployed a specific slice (a restriction of the user interface to a specific group of ontologies) as explained later. Categories indicate the topic(s) of the ontology, providing another way to classify ontologies in the portal independently from their groups or provenance. As of now we have defined 20 general categories such as Farms and Farming Systems, Plant Phenotypes and Traits, Plant Anatomy and Development, Agricultural Research, and Technology and Engineering. These categories were established in cooperation with FAO Agricultural Information Management Standards (AIMS), which has maintained the VEST Registry since 2011.

Groups and categories, along with other metadata, can be used on the “Browse” page of AgroPortal to filter out the list of ontologies (cf. Fig. 3). Of course, groups and categories are customizable, and will be adapted in the future to reflect the evolution of the portal’s content and community feedback. The portal’s architecture provides URIs for any portal objects, including groups and categories. For example, the URI <http://data.agroportal.lirmm.fr/categories/FARMING> identifies the

group “Farms and Farming Systems.” External applications can use those URIs to organize ontologies or tag them.

#### 4.2. Features from AgroPortal inherited from the NCBO BioPortal

The main features offered by the NCBO BioPortal are described in Noy et al. (2009), Whetzel et al. (2011). They include:<sup>16</sup>

**Ontology library.** The core mission of the AgroPortal is to serve as a one-stop shop for ontology descriptions and files. The portal also allows users to specify the list of ontologies that shall be displayed in their user interface when logged-in. While not replacing source code repository such as for instance GitHub, highly used by the community, the portal stores all ontology versions as they are submitted or automatically pulled, and can display their metadata and differences from one version to the next, although only the latest ontologies are referenced for queries. Ontologies can either be harvested from specified locations, or directly uploaded by users. Ontologies are semantically described (cf. metadata), and a browsing user interface allows to quickly identify, with faceted search, the ontologies of interest based on their descriptions and metadata.

**Search across all the ontologies.** AgroPortal search service indexes the ontology content (classes, properties and values) with Lucene, and offers an endpoint to search across the ontologies by keyword or identifier. For example, a keyword search on “abiotic factor”<sup>17</sup> will identify the occurrence of this term (or similar terms if none match exactly) in all the ontologies of the portal, and sort the results by relevance to the query and ontology popularity in the portal (number of views) (Noy et al., 2013). For the above search, the first three results are Abiotic factor (CO\_715\_0000078), Abiotic stress (CO\_320:Abiotic\_stress), and abiotic stress trait (TO\_0000168).

**Ontology browsing and content visualization.** The ontology ‘classes’ and ‘properties’ tab lets users visualize a class or property within its hierarchy, as well as see the related content (labels, definition, mappings, any other relations). An important point is that each

<sup>16</sup> The features of the portal inherited from the NCBO BioPortal are more extensively described in other publications that are referenced here. We provide here only a small summary as well as relevant agronomy related examples. In addition, the documentation of the portal is also available: <https://github.com/agroportal/documentation>.

<sup>17</sup> <http://agroportal.lirmm.fr/search?q=Abiotic%20factor>

AgroPortal content page can be accessed by a direct URL, that can be potentially used to dereference an ontology URI. Dereferencing (or resolving) means to obtain a concrete representation of the identified resource (e.g., a web page), for instance, [http://agroportal.lirmm.fr/ontologies/EOL/?p=classes&conceptid=http://opendata.inra.fr/EOL/EOL\\_0000014](http://agroportal.lirmm.fr/ontologies/EOL/?p=classes&conceptid=http://opendata.inra.fr/EOL/EOL_0000014) directly points to the class ‘water salinity’ in Environment Ontology for Livestock. For each ontology, a JavaScript widget allowing autocomplete with class names is also automatically generated and can be used by external web applications to facilitate the edition of data fields restricted to ontology concepts.

**Ontology versioning.** AgroPortal handles versioning through the concept of “submission.” Once an “Ontology” (an empty skeleton with minimal metadata) has been added once to the portal, “submission” objects can be attached. A new submission is created every time that ontology is re-submitted by a user, or pulled from its original location URL. Many ontologies are not necessarily maintained in a versioning system which offers a pull URL. It is up to the developer to decide when to manually uploading the new file, thereby creating a new submission (version) in AgroPortal. However, when the ontology is configured with a pull URL, the new ontology will be pulled in automatically (and versioned as a new submission) any night that it has changed. For example, the Matter Transfer Ontology for instance is developed by INRA using the @Web application (<http://pfl.grignon.inra.fr/atWeb>).<sup>18</sup> Although only the latest version is indexed and therefore available for searching, browsing and annotation, all the previous versions are downloadable, and a difference comparison can be viewed for each submission.

**Ontology mappings.** Another key role of AgroPortal is to store mappings (or alignments) between ontologies (Ghazvinian et al., 2009). Indeed, because ontologies’ contents overlap, it is crucial to maintain their interconnections—mappings—alongside the ontologies themselves. AgroPortal implements a mapping repository where each class-to-class mapping added to the portal is a first-class citizen and can be: stored, described, retrieved and deleted. The portal automatically creates some mappings when two classes share the same URI or CUI properties,<sup>19</sup> or when they share a common normalized preferred label or synonym. Although basic lexical mapping approaches can be inaccurate and should be used with caution (Faria et al., 2014; Pathak and Chute, 2009), they usually work quite well with the LOOM mapping algorithm used in AgroPortal (Ghazvinian et al., 2009). Other mappings can be explicitly uploaded from external sources, and in that case a mapping is reified as a resource described with provenance information (e.g., automatic or manual, who added it) and one or several tags to classify the mapping (e.g., owl:sameAs, skos:exactMatch, skos:broaderMatch, gold:translation). Such information helps users decide if they want to use these mappings.

**Community feedback.** While not being a state-of-the-art Web 2.0 social platform for ontologies, the AgroPortal features a few community features (Noy et al., 2009) such as: (i) *Ontology reviews*: for each ontology, a review can be written by a logged-in user from the ontology “Summary” page. It helps keep track of the quality. (ii) *Manual mapping creation*: On each ontology class, a logged-in user can create a mapping to another class (whether the class is inside the

AgroPortal, or in the NCBO BioPortal or another resource (cf. next Section)) (Noy et al., 2008). While this is illustrative, and may stimulate propositions, the real strength of the portal comes from using the API to automatically import mappings. (iii) *Notes* can be attached in a forum-like mode to a specific ontology or class, in order to discuss the ontology (its design, use, or evolution) or allow users to propose changes to a certain class (for instance, see [http://agroportal.lirmm.fr/ontologies/CO\\_321/?p=notes](http://agroportal.lirmm.fr/ontologies/CO_321/?p=notes)). Ontology developers (or any registered users) can subscribe to email notifications to be informed each time user feedback is added to their ontologies of interest.

**Ontology-based annotation.** AgroPortal features a text annotation service that will identify ontology classes inside any text (Jonquet et al., 2009) and can filter the results per ontologies and UMLS Semantic Types (McCray, 2003).<sup>20</sup> The text annotation service provides a mechanism to employ ontology-based annotation in curation, data integration, and indexing workflow; it has been used to semantically index several data resources such as in the NCBO Resource Index (Jonquet et al., 2011).<sup>21</sup> The workflow is based on a highly efficient syntactic concept recognition tool (using concept names and synonyms) (Dai et al., 2008), and on a set of semantic expansion algorithms that leverage the semantics in ontologies (e.g., is\_a relations and mappings). The Annotator is illustrated Fig. 1. It is also used to recommend ontologies for given text input, as described hereafter.

**Ontology recommendation.** The NCBO (in collaboration with LIRMM & University of Coruña) has recently released a new version of the Recommender system in BioPortal (Martinez-Romero et al., 2017), which has also been installed in AgroPortal. This service suggests relevant ontologies from the parent repository for annotating text data. The new recommendation approach evaluates the relevance of an ontology to biomedical text data according to four different criteria: (1) the extent to which the ontology covers the input data; (2) the acceptance of the ontology in the community; (3) the level of detail of the ontology classes that cover the input data; and (4) the specialization of the ontology to the domain of the input data. This new version of a service originally released in 2010 (Jonquet et al., 2010) combines the strengths of its predecessor with a range of adjustments and new features that improve its reliability and usefulness. To our knowledge, the AgroPortal Recommender is the first ontology recommendation service made for the agronomy community to identify which ontologies are relevant for (i) a given corpus of text or (ii) a list of keywords. For instance, if used with the ‘Plant height’ text example, from Fig. 1, the service will help users to identify Trait Ontology and multiple sources from the Crop Ontology as relevant for this text.

**Register ontology related projects.** The AgroPortal provides a project list edited by its users that materialize the ontology-project relation. For instance, the relation between the Planteome project and the six ontologies it uses is described at <http://agroportal.lirmm.fr/projects/Planteome>, in a format that can be used by AgroPortal to illustrate the ontologies that are most used. This information can then be employed for instance to sort ontologies by number of projects that use them.

In addition, all the previous features are available through two endpoints allowing automatic querying of the content of the portal: (i) a REST web service API (<http://data.agroportal.lirmm.fr/>)

<sup>18</sup> There are 328 submissions as of March 2017: <http://data.agroportal.lirmm.fr/ontologies/TRANSMAT/submissions>. The latest one is always available under [http://data.agroportal.lirmm.fr/ontologies/TRANSMAT/latest\\_submission](http://data.agroportal.lirmm.fr/ontologies/TRANSMAT/latest_submission)

<sup>19</sup> Uniform Resource Identifiers (URIs) are the standard way to identify resources (classes, properties, instances) on the semantic web when using RDF-based languages such as OWL or SKOS. Concept Unique Identifiers (CUIs) are identifiers used in the UMLS Metathesaurus. They are heavily used in the biomedical domain, but not very relevant within AgroPortal, where only two sources (the Semantic Network and the NCBI Taxonomy) are extracted from the UMLS.

<sup>20</sup> This feature originally developed for the NCBO Annotator (Jonquet et al., 2009) allows to filter the annotation results using the upper level 127 UMLS semantic type (<http://agroportal.lirmm.fr/ontologies/STY>) with which each concept in the UMLS are tagged. Because this was very useful on the NCBO BioPortal, we are considering an equivalent network and mechanism in the AgroPortal.

<sup>21</sup> The ‘Resource Index’ feature is not used in AgroPortal. Our vision is to accomplish this with the AgroLD partner project.

**Annotator**

The IBC AgroPortal Annotator processes text submitted by users, recognizes relevant ontology terms in the text and returns the annotations to the user. Use the interface below to submit sample text to get ontology-based annotations. Hover the mouse pointer on any button to see what it does. Click on the (?) to see a detailed help panel.

Subscribe to the [NCBO Annotator Users Google group](#) to learn more about who and how the Annotator is being used in different projects.

Plant height is a whole plant morphology trait which is the height of a whole plant. Plant height is sometime measured as height from ground level to the top of canopy at harvest.

Insert sample text

Ontology filters: Select Ontologies (PO, X, TO, K)

Matching parameters: ☒ Match Longest Only, ☐ Match Partial Words

NegEx / ConText

Select UMLS Semantic Types: Type here to select UMLS semantic types

Select UMLS Semantic Groups: Type here to select UMLS semantic groups

Include Ancestors Up To Level: None

Get Annotations

**Annotations**

total results 7 (direct 7 / ancestor 0 / mapping 0)

CLASS	filter	ONTOLOGY	filter	TYPE	filter	CONTEXT	MATCHED CLASS	filter	MATCHED ONTOLOGY	filter	SCORE
whole plant		Plant Trait Ontology		direct		... of a whole plant. Plant height is	whole plant		Plant Trait Ontology		10.000
plant height		Plant Trait Ontology		direct		Plant height is a whole ...	plant height		Plant Trait Ontology		8.644
plant height		Plant Trait Ontology		direct		... whole plant. Plant height is sometime measured ...	plant height		Plant Trait Ontology		8.644
whole plant morphology trait		Plant Trait Ontology		direct		... is a whole plant morphology trait which is the ...	whole plant morphology trait		Plant Trait Ontology		6.644
whole plant		Plant Ontology		direct		... of a whole plant. Plant height is	whole plant		Plant Ontology		6.644
height		Plant Trait Ontology		direct		... is the height of a whole ...	height		Plant Trait Ontology		4.322
height		Plant Trait Ontology		direct		... measured as height from ground level ...	height		Plant Trait Ontology		4.322

Format Results As: JSON

To reproduce these results: Corresponding REST web service call

Additional parameters are documented at [Annotator Web Service](#)

Fig. 1. AgroPortal Annotator with scored results. (web service call: [http://services.agroportal.lirmm.fr/annotator?text=Plant height is a whole plant morphology trait which is the height of a whole plant. Plant height is sometime measured as height from ground level to the top of canopy at harvest.&ontologies=PO,TO&longest\\_only=true &whole\\_word\\_only=true&score=cvalue](http://services.agroportal.lirmm.fr/annotator?text=Plant%20height%20is%20a%20whole%20plant%20morphology%20trait%20which%20is%20the%20height%20of%20a%20whole%20plant.%20Plant%20height%20is%20sometime%20measured%20as%20height%20from%20ground%20level%20to%20the%20top%20of%20canopy%20at%20harvest.&ontologies=PO,TO&longest_only=true&whole_word_only=true&score=cvalue)).

documentation) that returns XML or JSON-LD, making it easy to use AgroPortal within any web based application (Whetzel et al., 2011); and (ii) a SPARQL endpoint (<http://sparql.agroportal.lirmm.fr/test>), which is the standard mechanism to query RDF data (Salvadores et al., 2012).

We also like to point out that by adopting the NCBO technology, including its web service APIs (Whetzel and Team, 2013), an important number of external applications developed by the biomedical semantics community become available at very low cost for the agronomy community because of backward compatibility. This includes spreadsheet annotation tools such as OntoMaton (Maguire et al., 2013) Weboulous (Jupp et al., 2015), RightField (Wolstencroft et al., 2010) and WebSmatch (Coletta et al., 2012; Castanier et al., 2014); Zooma, a tool similar to the Annotator developed by the European Bioinformatics Institute ([www.ebi.ac.uk/spot/zooma](http://www.ebi.ac.uk/spot/zooma)); the UIMA wrapper to use the Annotator web service in other NLP applications (Roeder et al., 2010); the ontology wrapper OntoCAT (Adamusiak et al., 2010); the Galaxy platform tools (Miñarro-Giménez et al., 2012); the visualization tool FlexViz (Falconer et al., 2009); and finally all the different API clients (Java, Ruby, Perl, etc.) developed by the NCBO (<https://github.com/ncbo>) or other organizations (e.g. REDCap or Protégé plugins). To some extent, other ontology platforms such as the AberOWL, which features reasoning capabilities that AgroPortal does not yet offer (Slater et al., 2016), can automatically pull content from the AgroPortal.

#### 4.3. New AgroPortal features developed since the beginning of the project

While assuring community support, day-to-day maintenance and monitoring of the portal and keeping it up-to-date with the NCBO technology, we have worked on customizations and specific services. These services target the agronomic community, but that could in some cases be used for any domains. With the vision of collaborative

development of BioPortal and AgroPortal, when relevant and possible, we push new features back to the main NCBO code branch where BioPortal users or the appliance itself can benefit. The AgroPortal open source code and documentation are accessible on GitHub: <https://github.com/agroportal>.

**Multilingualism in AgroPortal.** In the context of the SIFR project and in consultation with the NCBO, we are working on making BioPortal multilingual (Jonquet et al., 2015). This is still work in progress, although we have already added relevant metadata properties to: (i) identify the natural language in which labels are available; and (ii) link monolingual ontologies to their translations. We have also changed the representation of multilingual translation mappings. For the moment, we have chosen to consider English as the main language of AgroPortal (i.e., the one used to display content as well as indexed for Search, Annotator and Recommender services). Multilingual ontologies (i.e., with labels in multiple languages) are parsed, but only the English content is explicitly used. Non-English monolingual ontologies are attached as “views” of a main ontology that is solely described with metadata (no content). For instance, the French Agroecology Knowledge Management ontology, used in a French collaborative network (<http://agroportal.lirmm.fr/ontologies/GECO>) is only described with metadata but has attached a specific view (<http://agroportal.lirmm.fr/ontologies/GECO-FR>) with the real content in French.

**Mapping related features.** In order to interconnect AgroPortal with the NCBO BioPortal or any other repositories, we have changed the model of AgroPortal mappings to store mappings to ontologies (i) in another instance of the BioPortal technology (“inter-portal”), (ii) in any ‘external’ resources. Hence, any AgroPortal class can be linked to any class in other knowledge resource (e.g., DBpedia, WordNet, AgroLD) or the NCBO BioPortal itself). Mappings are described with

provenance data and typed with a property from a standard semantic web vocabulary (e.g., OWL, SKOS, GOLD). For instance:

- o The class ‘plant organ’ in the Plant Ontology has been manually mapped to the ‘Plant organ’ entity in the DBpedia knowledge base. The mapping tag used is `skos:exactMatch` which means that the classes represent the same entity, while not supporting a logical substitution (as with `owl:sameAs`).
- o The class ‘biomass’ in the Biorefinery ontology has been manually mapped to the class ‘Biomass’ in MeSH on the NCBO BioPortal, and automatically mapped to the class ‘biomass’ in the AnaEE Thesaurus.
- o The class ‘zooplankton’ in the AnaEE Thesaurus has been mapped to ‘zooplankton’ in the Ontology for MIRNA Target ([http://purl.obolibrary.org/obo/OMIT\\_0015869](http://purl.obolibrary.org/obo/OMIT_0015869)), which is not available in AgroPortal.

**Semantic annotation with scoring.** Within the SIFR project we develop new features and natural language based enhancement that target all the Annotator deployments (the NCBO, AgroPortal or SIFR one). For instance, to facilitate the use of annotation for semantic indexing, we have implemented three scoring methods for the Annotator. They are based on term frequency and especially useful with multi-word terms. We demonstrate the results of these new scoring measures in Melzi and Jonquet (2014). For instance, when considering annotating the text:<sup>22</sup> “*Plant height is a whole plant morphology trait which is the height of a whole plant. Plant height is sometime measured as height from ground level to the top of canopy at harvest.*” with the AgroPortal Annotator, the scoring method gives more importance to the concept ‘plant height’ (score = 8.64) than to the concept ‘height’ (score = 4.32), whose lexical form is actually more frequent in the text. The user interface of the Annotator is illustrated in Fig. 1.

**Ontology formats.** We have worked on the full support of different formats such as (i) SKOS (SKOS-XL is not handled yet), which is highly used in agronomy (AnaEE Thesaurus, Agrovoc, CAB Thesaurus and NAL Thesaurus all use SKOS); and (ii) the Crop Ontology Trait Dictionary template v5, adopted for instance by the Breeding API and Crop Ontology (import/export in this format is currently done outside of AgroPortal).

**Ontology metadata.** To facilitate the ontology identification and selection process, which has been assessed as crucial to enable ontology reuse (Park et al., 2011), we implemented a new metadata model to better support descriptions of ontologies and their relations, respecting recent metadata specifications, vocabularies, and practices used in the semantic web community (Xiang et al., 2011). We reviewed the most common and relevant vocabularies (23 in total) to describe metadata for ontologies, including Dublin Core, VoID, Ontology Metadata Vocabulary, and the Data Catalog Vocabulary. We then grouped those properties into a unified and simplified model of 127 properties (distilled from an initial list of 346 properties that will be parsed by the portal)<sup>23</sup> that includes the 45 properties originally offered by the NCBO BioPortal, and describe all the new properties with standard vocabularies.<sup>24</sup> This gives us, for example, a model to describe the type of the semantic resource uploaded to the portal (for example, thesaurus, ontology, taxonomy, or terminology). Our work provided three important new features for AgroPortal (Toulet et al., 2016):

- o Once an ontology is uploaded, AgroPortal automatically extracts most of the ontology metadata if they are included in the original file, and automatically populates some of them if possible (e.g., metrics, endpoints, links, examples). Ontology developers can

manually update those extracted or calculated values if desired. In addition, we have entirely redesigned AgroPortal’s ontology submission page to facilitate editing the metadata. Whenever possible, the user interface facilitates the selection of the metadata values, while in the backend those values are stored with standard URIs. For instance, the user interface will offer a pop-up menu to select the relevant license (CC, BSD, etc.) while the corresponding URI will be taken from the RDFLicense dataset (<http://rdflicense.appspot.com>). Knowledge organization systems types are taken from the KOS Types Vocabulary from the Dublin Core initiative.<sup>25</sup> An example using the OntoBiotope ontology metadata page in AgroPortal is shown in Fig. 2.

- o AgroPortal ontology browse page (Fig. 3) offers three additional ways to filter ontologies in the list (content, natural language, formality level) as well as three new options to sort this list. We believe these new features facilitate the process of selecting relevant ontologies.
- o We have begun facilitating the comprehension of the agronomical ontology landscape by displaying diagrams and charts about all the ontologies on the portal (average metrics, most used tools, leading contributors & organization, and more). We have created a new AgroPortal ‘landscape’ page that displays metadata “by property” –as opposed as “by ontology” as in Fig. 2 (<http://agroportal.lirmm.fr/landscape>).

For each ontology available and uploaded in the portal, we collaborate with the ontology developers to extensively describe their metadata. Information is generally found either in other registries (e.g., LovInra, VEST Registry, the OBO Foundry) or identified in the publication, web site, documentation, etc. found about the ontologies. With these curated metadata, all users can confidently select and review ontologies; any submission of the ontology can include more authoritative and more complete metadata, available to any user including the original provider, and for other linked open data users and applications; and AgroPortal’s users can better understand the landscape of ontologies in the agronomy and related domains.

## 5. Driving agronomic use case results

Now that AgroPortal has been extensively presented, we focus on the results of each use case, and illustrate the value added by this portal and its semantic content.

### 5.1. Agronomic Linked Data (AgroLD)

The OWL versions of the ontologies available in AgroPortal were retrieved from that single repository. Although AgroPortal is not the main original location for these ontologies (they are accessible on the OBO Foundry and Cropontology.org) it was convenient to find them all in one place, and to use a unique and consistent API. Plus, we also used the AgroPortal Annotator web service to annotate more than 50 datasets and produced 22% additional triples, which were validated manually (Fig. 4). Building such an annotation service for all these ontologies was one of the driving needs for AgroPortal. Encoding the original data in RDF allowed us to establish an annotation for every appropriate case, using `owl:sameAs` relations, between the data element (e.g., Protein in the SouthGreen database) defined with a new URI (<http://www.southgreen.fr/agrold/resource/Protein>) and an ontology term (e.g., the term ‘polypeptide’ in the Sequence Ontology ([http://purl.obolibrary.org/obo/SO\\_0000104](http://purl.obolibrary.org/obo/SO_0000104))). Note that we have decided to use `owl:sameAs` in this case as the resources are logically equivalent and this is a common practice in linked open data to

<sup>22</sup> Two appended definitions from the Trait Ontology and from the Crop Ontology.

<sup>23</sup> <https://github.com/agroportal/documentation/tree/master/metadata>

<sup>24</sup> For instance, the call [http://data.agroportal.lirmm.fr/ontologies/PR/latest\\_submission?display=all](http://data.agroportal.lirmm.fr/ontologies/PR/latest_submission?display=all) will display the JSON-LD format of all the metadata properties (populated or not) for the Protein Ontology.

<sup>25</sup> [http://wiki.dublincore.org/index.php/NKOS\\_Vocabularies](http://wiki.dublincore.org/index.php/NKOS_Vocabularies) (ANSI/NISO Z39.19-2005).



Fig. 2. AgroPortal's Ontology metadata page for ONTOBIOTOPE (<http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE>). The red box corresponds to the new metadata fields added in AgroPortal ontology model extracted by the portal, or provided by the administrators or by the ontology developers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

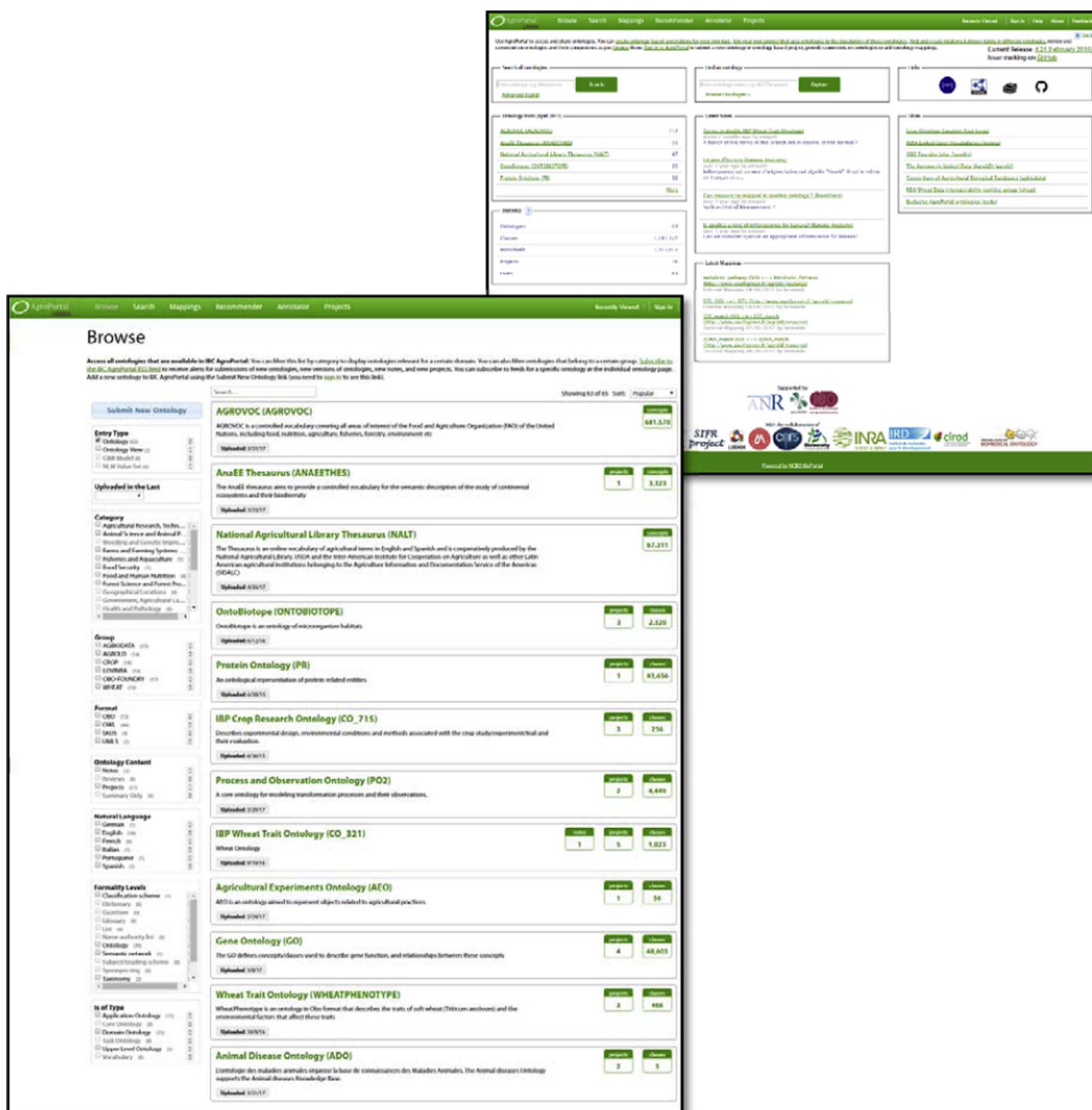


Fig. 3. Screenshots from the AgroPortal user interface (<http://agroportal.lirmm.fr>). The welcome page (back) provides a rapid overview of the content of the portal and enables a user to quickly search for and in ontologies. The browse ontology page (front) provides the list of ontologies and offer multiple sorting or faceted filtering of this list to facilitate the identification of the ontologies of interest.

interlink datasets; similar annotations have been made for properties using owl:equivalentProperty or rdfs:subPropertyOf (when an equivalent property did not exist). Now that AgroPortal handles ‘external mapping’ as described in Section 4.3, we have been able to upload all our annotations (to 23 classes and 21 properties) to fully connect the concepts from the different ontologies, and create annotations, directly within AgroPortal.<sup>26</sup>

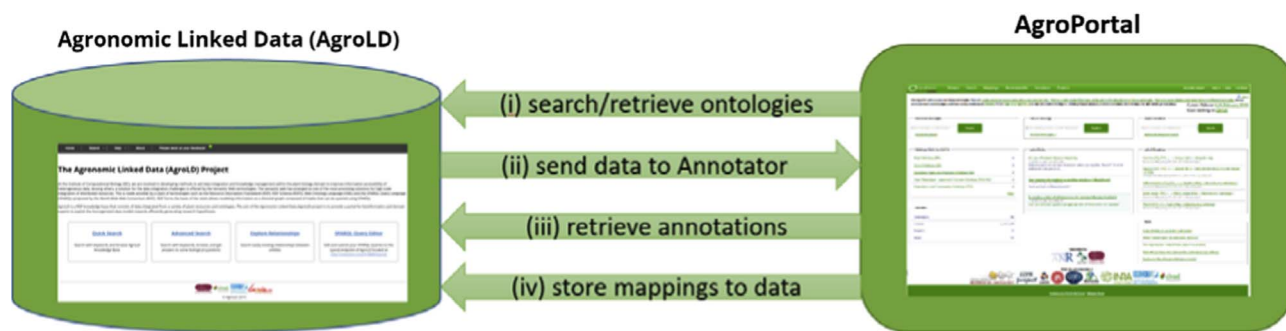
As a result, AgroLD has incorporated the data from various databases (Table 4), and produced 37 million RDF triples (Venkatesan

et al., 2015). The data source selection followed the needs and priorities of the IBC project’s work-package 5. It included important data sources such as GOA, Gramene, Oryza Tag Line, and GreenPhylDB. AgroLD can now gather genomic and phenotypic information to answer biological questions such as: “find proteins involved in plant disease resistance and high grain yield traits.” Such queries would be hard or impossible to resolve without the appropriate ontologies integrated to support the conclusion. The reader may refer to <http://agrold.org/sparqleditor.jsp> for more examples of queries in AgroLD.

## 5.2. RDA Wheat Data Interoperability (WDI) working group

We created and maintain explicit sub-parts within AgroPortal called

<sup>26</sup> The previous example ‘polypeptide’ in SO is available here in the mapping tab: [http://agroportal.lirmm.fr/ontologies/SO?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FOSO\\_0000104](http://agroportal.lirmm.fr/ontologies/SO?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FOSO_0000104)



**Fig. 4.** Interaction between AgroPortal and AgroLD. (i) AgroPortal provides a unique endpoint to retrieve heterogeneous ontologies; (ii) AgroLD's annotation pipeline sends data to the AgroPortal Annotator and (iii) retrieves annotations with ontology terms used to build AgroLD; finally (iv) AgroPortal offers a link from the ontologies to data stored in AgroLD with the 'inter portal' mapping mechanism.

**Table 4**

Plant species and data sources in AgroLD. The number of tuples gives an idea of the number of elements we have annotated from the data sources and the number of RDF triples produced. The crops and ontologies are referred as: R = rice, W = wheat, A = Arabidopsis, S = sorghum, M = maize GO = Gene Ontology, PO = Plant Ontology, TO = Plant Trait Ontology, EO = Environment Ontology, SO = Sequence Ontology, CO = Crop Ontology (specific trait ontologies).

Data sources	URL s	# tuples	Crops	Ontologies used	# triples produced
GO associations	geneontology.org	1160 K	R, W, A, M, S	GO, PO, TO, EO	2700 K
Gramene	gramene.org	1718 K	R, W, M, A, S	GO, PO, TO, EO	5172 K
UniprotKB	uniprot.org	1400 K	R, W, A, M, S	GO, PO	10000 K
OryGenesDB	orygenesdb.cirad.fr	1100 K	R, S, A,	GO, SO	2300 K
Oryza Tag Line	oryzatagline.cirad.fr	22 K	R	PO, TO, CO	300 K
TropGeneDB	tropgenedb.cirad.fr	2 k	R	PO, TO, CO	20 K
GreenPhylDB	greenphyl.org	100 K	R, A	GO, PO	700 K
SniPlay	sniplay.southgreen.fr	16 K	R	GO	16000 K
TOTAL					37000 K

slices.<sup>27</sup> The wheat slice in AgroPortal (<http://wheat.agroportal.lirmm.fr>) allows the community to share common definitions for the words they utilize to describe and annotate data, which in turn makes the data more machine-readable and interoperable. Furthermore, each slice enables ontology developers to make their ontologies more visible to targeted agronomic research communities; as of today, AgroPortal's Wheat group contains 20 of the 23 ontologies identified by the WDI.<sup>28</sup> Each ontology has been carefully described (with licenses, authority, availability, and so on), and a new metadata property (omv:endorsedBy) is used to show the ontology's endorsement by the WDI working group.

This work has been reported in the WDI's set of guidelines for wheat data description (<http://ist.blogs.inra.fr/wdi>) (Dzalé-Yeumo et al., 2017), and used since then as a reference to identify and select ontologies related to wheat. Among AgroPortal's registered users, a dozen are members of the RDA WDI working group. In the future, the slice will be maintained/managed by the WheatIS consortium to organize new wheat-related ontologies and store the alignments between them. AgroPortal's adoption by the WDI working group leveraged several advanced features of the platform as customized by the AgroPortal team. The result directly enhanced the community's processes and capabilities, provided customized access to information of particular interest to this community, and achieved wide uptake in the working group.

<sup>27</sup> Slices are a mechanism supported by the platform to allow users to interact (both via API or UI) only with a subset of ontologies in AgroPortal. If browsing the slice, all the portal features will be restricted to the chosen subset, enabling users to focus on their specific use cases. On AgroPortal, slices and groups are synchronized, so every group (described Section 4.1) has a corresponding slice displaying only the ontologies from that group.

<sup>28</sup> Among the missing ones are, CAB Thesaurus, that we are currently working on integrating; CheBI that we have decided not to upload yet; and Wheat Inra Phenotype Ontology (that is currently being merged with CO\_321).

### 5.3. INRA Linked Open Vocabularies (LovInra)

To augment the visibility of INRA's semantic resources, and achieve their mapping to resources within and external to INRA, the institute has chosen AgroPortal to publish and host INRA's resources and encourage adoption of semantic web standards. If a semantic resource is declared on the LovInra service, it is immediately uploaded and fully described on AgroPortal. Resources that are not on the LovInra service can be directly uploaded by their developers to the portal, an important consideration for such a big organization. AgroPortal assigns the new resources to the correct group and slice, and properly tags them (SKOS vocabularies, OWL/SKOS termino-ontological resources, or OBO/OWL ontologies).

The LovInra group/slice contains 16 ontologies relating to process modeling, biotopes, animal breeding, and plant phenotypes. AgroPortal has become a major element of the LovInra service and is heavily encouraged and supported by INRA. It has started to play a key resource role allowing the group's users to: (i) have a comprehensive view of the portal's ontologies (topics, types, community, etc.); (ii) quickly find a resource, and understand its content and structure by browsing it and annotating documents; (iii) discover additional vocabularies that could be used; and (iv) have access to projects linked to vocabularies, and understand how they were created or used by the projects, possibly exchanging shared experience or insights.

### 5.4. The Crop Ontology project

Currently, the AgroPortal hosts 19 crop-specific trait ontologies developed within the Crop Ontology project: Wheat, Rice, Cassava, Groundnut, Chickpea, Banana, Sweet potato, Cowpea, Soybean, Lentil, Pigeon pea, Sorghum, Pear millet, Maize, Groundnut, Castor bean, Mungbean, and Cassava. Additional ontologies will be integrated in the future with the help of the crop ontology curators. Similarly to the

LovInra or WDI use cases, these ontologies are grouped within the portal and can be browsed in a dedicated slice (<http://crop.agroportal.lirmm.fr>). Parsers for specific trait template have been developed, and in the future any of this community's formats (OBO, OWL, and CSV) shall be used to import and export trait ontologies directly within AgroPortal.<sup>29</sup>

Moreover, in the context of the Planteome project ([www.planteome.org](http://www.planteome.org)), the alignment (or mapping) of terms within and across different plant related ontologies have been created: both within the crop ontologies themselves (in different crop branch) or with other reference ontologies commonly used in plant biology (e.g., PO, TO, EO). In the future, AgroPortal will formally store the alignments between all these ontologies.<sup>30</sup>

Finally, hosting ontologies on AgroPortal offers new functionalities to the crop ontology community such as versioning, an open SPARQL endpoint, community notes, and the annotation service, while still supporting the uses of the current web site.<sup>31</sup> For instance, new traits or mappings between them can be suggested directly by breeders using AgroPortal's community features, while not directly impacting the original ontology. Each time a suggestion is made to an ontology, the breeders interested in the corresponding crop can be notified of the suggestions and comments of their peers.

### 5.5. GODAN Map of Agri-Food Data Standards

The GODAN Action project wanted to build a broadly scoped global map of standards while leveraging detailed information and content about them that could be maintained in an ontology or vocabulary. To achieve this, the new map of standards was built on top of the existing VEST Registry, but added bidirectional mechanisms linking the VEST Registry with AgroPortal. The combined system automatically imports resource descriptions from the AgroPortal into the VEST, and links records from the VEST back to the AgroPortal entries, in order to provide access to the AgroPortal content and related services. The new registry, called *Map of Agri-Food Data Standards* (<http://vest.agrisemantics.org>), was released in 2016 under two umbrellas: the GODAN Action project, and the new RDA AgriSemantics working group,<sup>32</sup> which launched at the end of 2016. The Map of Standards leverages the AgroPortal's new metadata model and application programming interface to populate the entries in the Map using a single web service call. In addition to searching by metadata, the AgroPortal's Recommender will help the agronomy community identify ontologies or vocabularies of interest.

The synchronization and interlinking of the two platforms is for the moment semi-automatic, with the content of AgroPortal being regularly imported into the global map. Users can register or edit the description of a vocabulary in the Map, and if the vocabulary is in a compatible format, they are offered, the option to add the vocabulary directly into AgroPortal. In the future, this process will be fully automatized.

## 6. Discussion

### 6.1. General reflection on research scenarios supported by AgroPortal

AgroPortal (like the NCBO BioPortal before it) adopted a vision where multiple knowledge artifacts are made available in a common

place (though not combined), and cast to a common model. While doing so, the portal arguably limits the full power of ontologies, constraining their use to features supported by the common model. We see two general scenarios of use for our portal:

The portal provides basic ontology library services for users with a “vertical need” —those who want to do very precise things (e.g., reasoning, using specific relations) using only suitable ontologies (developed by the same communities and in the same format). Such users may just use the portal to find and download ontologies, and work in their own environment.

The portal provides many semantic services (for examples, lexical analysis, search, text annotation, and use of hierarchical knowledge) to users with “horizontal needs” —those who want to work with a wide range of ontologies and vocabularies useful in their domain but developed by different communities, overlapping and in different formats. Such users greatly appreciate the unique endpoints (web application and programmatic for REST and SPARQL queries) offered by the portal under a simplified common model.

We believe there are existing resource to address the first need in agronomy (e.g., OBO Foundry, FAIRSharing, VEST registry), although without containing all the relevant ontologies and vocabularies. However, we argue the second need is unmet by any of the available platforms. If we want semantic resources like ontologies and vocabularies to achieve widespread adoption, we must facilitate their use for non-ontological experts who still want to use multiple heterogeneous semantic resources.

### 6.2. Implementation of the requirements

As presented and illustrated on examples, most of the requirements listed in Section 3 have been addressed at least partially thanks to the original BioPortal features (e.g., requirements #1–#6, #8, #10, #15, #16, #18), our new implementations (#5, #7, #11, #15), and our applying the platform to the community needs (#1, #10, #11, #17, #18). Some requirements are not yet completely achieved and/or evaluated, for instance:

(#4) The AgroPortal Annotator has been used by the AgroLD use case, but not by other ones. We have not yet evaluated the capability of the service to automatically identify entities such as plant phenotypes in text.

(#8) Automatically generating mappings is an important issue for a portal on ontologies. Although it is convenient to have some simple lexical mappings automatically generated by AgroPortal with the LOOM algorithm (Ghazvinian et al., 2009), we find that this is not enough to correctly interlink the multiple vocabularies and ontologies developed by the community. We are integrating other state-of-the-art ontology matchers such as YAM++ (Ngo and Bellahsene, 2012) as well as designing specific mapping curation interfaces. At the same time, identifying and harvesting into AgroPortal the mappings already produced by the community is a huge task, not yet begun.

(#9) We have not automatically linked databases of annotated agronomical data using ontology concepts (from within AgroPortal). While the original BioPortal has the NCBO Resource Index (Jonquet et al., 2011), we plan to rely on external annotated resources such as AgroLD (Venkatesan et al., 2015) to interlink with data. To store this information, we will build on our rich mapping model in AgroPortal as presented Section 4.3. As another example, being part of the map of standards will allow ontologies in AgroPortal to link directly to

<sup>29</sup> Most of these conversions are still achieved outside of AgroPortal. The automatically generated CSV output format is not yet compliant with the Crop Ontology trait template (v5).

<sup>30</sup> For instance, something to capture that plant height for wheat (CO\_321:0000024) is somehow linked to the general plant height trait (TO\_0000207) that is itself a morphology trait (TO:0000398). This work is ongoing, and the data is not yet publicly released.

<sup>31</sup> In the future, to offer to breeders a simple and customized interface while avoiding duplication effort, we will consider serving the Crop Ontology website use cases by directly accessing AgroPortal's backend through the REST API.

<sup>32</sup> <https://www.rd-alliance.org/groups/agrisemantics-wg.html>

datasets that use them such as the CIARD RING directory (<http://ring.ciard.net>) (Pesce et al., 2011), as that was previously indexed with some of the VEST content. The CIARD RING can be queried via SPARQL or REST API and the links between vocabularies and datasets can therefore be retrieved by any system. Such a feature, has been requested and will be among the next features of AgroPortal. In the long-term vision, AgroPortal will directly query the CIARD RING, AgroLD, or any relevant data sources like Bio2RDF or Planteome, so that a user browsing ontologies can get direct access to the data to which these ontologies link.

(#12) Although community feedback is an important aspect for working group and communities, we have not successfully engaged yet our user groups to add reviews, notes, or comments about the ontologies. A complete rethinking of this issue is a future challenge for AgroPortal.

(#13) The roadmap to make the technology fully multilingual has been identified, but not yet fully implemented.

(#15) AgroPortal can be used as a destination for dereferenced URIs. In the future, we shall discuss these strategic questions with our collaborators.

### 6.3. Future and perspectives

Considering the need for a repository of ontologies for agronomy, food, plant sciences, and biodiversity, we expect broad community adoption of the AgroPortal. The endorsement of associated partners (IRD, CIRAD, INRA, IRSTEA) illustrates the impact and interest not just in France, but also internationally (e.g., FAO, Bioversity International, IC-FOODS consortium, NCBO, Planteome, RDA working groups). More recently, two other RDA working groups (Rice Data Interoperability<sup>33</sup> and AgriSemantics<sup>34</sup>) have expressed interest in using AgroPortal as a backbone for data integration and standardization.

In the future, we will identify more potential users for the portal and support new research scenarios. For instance, within the RDA AgriSemantics WG, we are interested in using AgroPortal to host the future Global Agricultural Concept Scheme (GACS) (Baker et al., 2016), which will result from the integration and alignment of Agrovoc, NAL Thesaurus and CAB Thesaurus. The portal is considered by the GACS working group as a candidate to host the three source vocabularies (it already includes two of them), as well as the GACS itself. GACS beta version 3.1 is currently available in AgroPortal, but no specific customization has been performed. In addition, we will be offering our services to these projects:

the new IC-FOODS project (International Center for Food Ontology, Operability, Data & Semantics - [www.ic-foods.org](http://www.ic-foods.org)) that will be developing ontologies related to food, nutrition, eating behaviors (Muskier et al., 2016);

ecologists developing the Thesaurus of Plant characteristics (Garnier et al., 2017);

the French IRESTA organization, to facilitate the use of ontologies in the design of the future government-led open data repository for agriculture project (AgGate).<sup>35</sup>

To foster interest in agronomy and the semantic web and identify potential AgroPortal applications, we launched in 2016 a series of AgroHackathons ([www.agrohackathon.org](http://www.agrohackathon.org)) that focused among other things on AgroPortal and AgroLD. Finally, in the next future, we plan to achieve a community survey evaluation to capture the feedback of our community, review the requirements, and drive the future directions of the project.

## 7. Conclusion

In this paper we have presented AgroPortal, an open vocabulary and ontology repository for agronomy. We have discussed five use cases already using the portal to support their work on data interoperability, and demonstrated that beyond these use cases the portal offers services of value to the broader community. The thematic boundaries of the portal are evolving (agriculture also includes animals, and is strongly related to environmental science), and over time the community will communicate what they expect to find in such a repository.

The community outreach challenge of such a project is huge. It involves identifying already existing resources, whether already shared or not, encouraging their developers to make them available, and finally harvesting them into the single ontology repository, capable of providing many services across the heterogeneous content. We recognize that this challenge was highly facilitated by previous important efforts such as the NCBO BioPortal, OBO Foundry, Planteome, and Crop Ontology projects. In addition, we are conscious that by adopting an open library approach, knowledge “conflicts” or redundancies as well as convergences and consolidations will appear. We believe the AgroPortal will help the scientific community to fully understand these issues, and address them as appropriate.

The technological challenges of such a project are also huge; therefore, we have built upon technology previously developed in the biomedical domain. We see here an opportunity to capitalize technology and scientific outcomes of the last twelve years in a closely related domain. We illustrated in the context of five important driving agronomic use cases how AgroPortal can enable new science for the community developing and using agronomical ontologies and vocabularies worldwide. In addition, the AgroPortal platform offers a terrain for pursuing important informatics and semantic web issues, such as semantic annotation, multilingual ontologies, metadata description, ontology engineering and alignment, and ontology recommendation, and will.

Ultimately, we believe AgroPortal provides powerful services, standards, and information that will greatly facilitate the adoption of open data in agriculture and benefit the extended agronomic community, the semantic web and data science communities, and the biomedical community that in many ways laid the groundwork that AgroPortal now leverages.

## Acknowledgment

This work is partly achieved within the Semantic Indexing of French biomedical Resources (SIFR – [www.lirmm.fr/sifr](http://www.lirmm.fr/sifr)) project that received funding from the French National Research Agency (grant ANR-12-JS02-01001), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 701771, the NUmEV Labex (grant ANR-10-LABX-20), the Computational Biology Institute of Montpellier (grant ANR-11-BINF-0002), as well as by the University of Montpellier and the CNRS. We also thank the National Center for Biomedical Ontologies for their help and time spent with us in deploying the AgroPortal.

## Author contributions

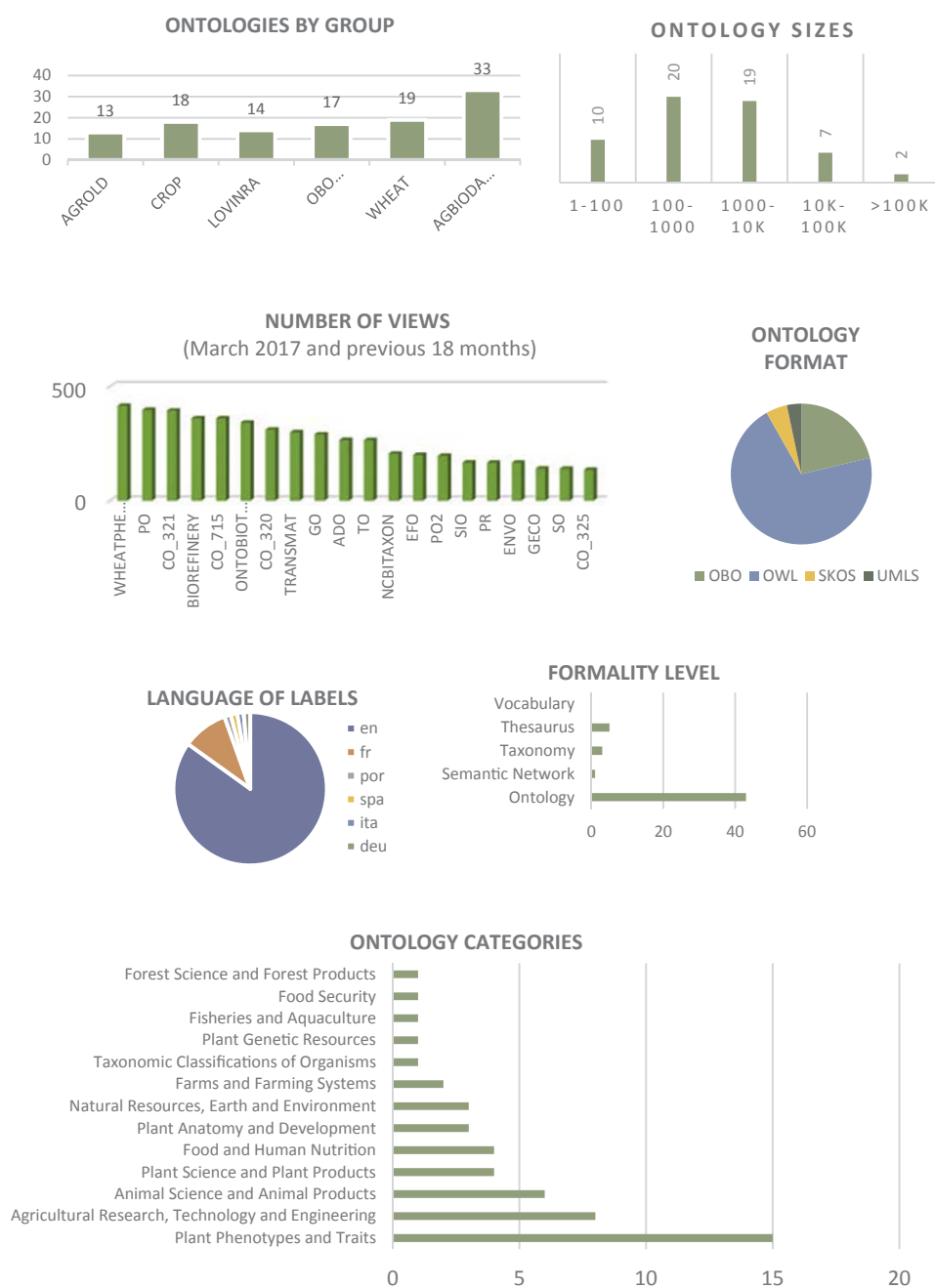
CJ conceived of the project, provided the scientific direction and led the writing of this manuscript. VE & AT respectively implemented/maintained the portal and managed the content with help of the community. JG and MAM helped and gave directions in realizing the project in collaboration with NCBO, and JG provided extensive final review and editing. Then, EA, SA, MAL, EDY, VP & PL respectively presented each of the use cases. All authors declare no conflict of interest and approved the final manuscript.

<sup>33</sup> <https://rd-alliance.org/groups/rice-data-interoperability-wg.html>

<sup>34</sup> <https://rd-alliance.org/groups/agrisemantics-wg.html>

<sup>35</sup> <https://www.economie.gouv.fr/files/files/PDF/rapport-portail-de-donnees-agricoles.pdf>

## Appendix



**Fig. 5.** AgroPortal public ontology analytics (May 2017). Updated versions of these statistics are automatically generated from AgroPortal's new metadata model, and made available on its Landscape page (<http://agroportal.lirmm.fr/landscape>).

## References

- Goble, C., Stevens, R., 2008. State of the nation in data integration for bioinformatics. *Biomed. Inf.* 41, 687–693.
- Rubin, D.L., Shah, N.H., Noy, N.F., 2008. Biomedical ontologies: a functional perspective. *Brief. Bioinform.* 9 (1), 75–90.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al., 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25 (1), 25–29.
- Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungall, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P., 2012. The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* 54, e1.
- Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G.F., Hancock, D., Morrison, N., Bruskewich, R., McLaren, G., 2010. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants*, vol. 2010, May.
- Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E., 2013. The environment ontology: contextualising biological and biomedical entities. *Biomed. Semantics* 4, 43.
- Devare, M., Aubert, C., Laporte, M.-A., Valette, L., Arnaud, E., Buttigieg, P.L., 2016. Data-driven agricultural research for development - a need for data harmonization via semantics. In: Jaiswal, P., Hoehndorf, R. (Eds.), 7th International Conference on Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 2, August.
- Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougnot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönisch, G., Bunker, D.E., Cornelissen, J.H.C., Díaz, S., Enquist, B.J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., Pérez-Harguindeguy, N., Poorter, H., Schildhauer, M., Shipley, B., Violle, C., Weiher, E., Wirth, C., Wright, I.J., Klotz, S., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. *Ecology* 105, 298–309.
- Griffiths, E., Brinkman, F., Buttigieg, P.L., Dooley, D., Hsiao, W., Hoehndorf, R., 2016. FoodON: a global farm-to-fork food ontology - the development of a universal food vocabulary. In: Jaiswal, P., Hoehndorf, R. (Eds.), 7th International Conference on

- Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 2, August.
- Musker, R., Lange, M., Hollander, A., Huber, P., Springer, N., Riggall, C., Quinn, J.F., Tomich, T.P., 2016. Towards designing an ontology encompassing the environment-agriculture-food-diet-health knowledge spectrum for food system sustainability and resilience. In: Jaiswal, P., Hoehndorf, R. (Eds.), 7th International Conference on Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 5, August.
- Hughes, L.M., Bao, J., Hu, Z.-L., Honavar, V., Reecy, J.M., 2014. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *Anim. Sci.* 86, 1485–1491.
- Meng, X.-X., 2012. Special issue – agriculture ontology. *Integrative Agriculture*, vol. 11, pp. 1, May.
- Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L., Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishnalka, L., Matsunaga, A., Midford, P., Morrison, N., Tuama, Éamonn Ó., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wiczkorek, J., Whitacre, J., Wooley, J., 2014. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PLoS One* 9, 13.
- Wang, Y., Wang, Y., Wang, J., Yuan, Y., Zhang, Z., 2015. An ontology-based approach to integration of hilly citrus production knowledge. *Comput. Electron. Agric.* 113, 24–43.
- Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., Vialle, C., 2016. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Comput. Electron. Agric.* 127, 351–367.
- Lehmanna, R.J., Reichera, R., Schiefera, G., 2012. Future internet and the agri-food sector: State-of-the-art in literature and research. *Comput. Electron. Agric.* 89, 158–174.
- Jaiswal, P., 2011. Plant Reverse Genetics: Methods and Protocols, ch. Gramene Database: A Hub for Comparative Plant Genomics. Humana Press, pp. 247–275.
- Sachit Rajbhandari, J.K., 2012. The AGROVOC concept scheme – a walkthrough. *Integrative Agriculture* 11, 694–699.
- d'Aquin, M., Noy, N.F., 2012. Where to publish and find ontologies? a survey of ontology libraries. *Web Semantics* 11, 96–111.
- Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, 267–270.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, T.O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N.H., Whetzel, P.L., Lewis, S., 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.
- Larmande, P., Arnaud, E., Mougnot, I., Jonquet, C., Libourel, T., Ruiz, M., (Eds.), 2013. Proceedings of the 1st International Workshop on Semantics for Biodiversity, Montpellier, France, May.
- Baker, T., Caracciolo, C., Jaques, Y., (Eds.), 2015. Report on the Workshop “Improving Semantics in Agriculture, (Rome, Italy), Food and Agriculture Organization of the UN, July.
- Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A., 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 37, 170–173.
- Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 39, 541–545.
- Whetzel, P.L., Team, N., 2013. NCBO technology: powering semantically aware applications. *Biomed. Semantics* 49, 451.
- Ding, Y., Fensel, D., 2001. Ontology library systems: the key to successful ontology re-use. In: 1st Semantic Web Working Symposium, SWWS'01, Stanford, CA, USA, pp. 93–112, CEUR-WS.org, August.
- Baclawski, K., Schneider, T., 2009. The open ontology repository initiative: Requirements and research challenges. In: Tudorache, T., Correndo, G., Noy, N., Alani, H., Greaves, M., (Eds.), Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, CK'09, vol. 514 of CEUR Workshop Proceedings, Washington, DC, USA, pp. 10, CEUR-WS.org, October.
- Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S.J.H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., Hide, W., 2012. Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126.
- Pesce, V., Geser, G., Protonotarios, V., Caracciolo, C., Keizer, J., 2013. Towards linked agricultural metadata: directions of the agINFRA project. In: 7th Metadata and Semantics Research Conference, AgroSem track, Thessaloniki, Greece, pp. 12, November.
- Xiang, Z., Mungall, C., Ruttenberg, A., He, Y., 2011. Ontobee: a linked data server and browser for ontology terms. In: Bodenreider, O., Martone, M.E., Ruttenberg, A., (Eds.), 2nd International Conference on Biomedical Ontology, ICBO'11, vol. 833 of CEUR Workshop Proceedings, Buffalo, NY, USA, p. 3, July.
- Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H., 2006. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinf.* 7, 7.
- Hoehndorf, R., Slater, L., Schofield, P.N., Gkoutos, G.V., 2015. AberOWL: a framework for ontology-based data access in biology. *BMC Bioinf.* 16, 1–9.
- Vandenbussche, P.-Y., Atemez, G.A., Poveda-Villalón, M., Vatan, B., 2014. Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. *Semantic Web*.
- Till, M., Kutz, O., Codescu, M., 2014. Ontohub: A semantic repository for heterogeneous ontologies. In: Theory Day in Computer Science, DACS'14, (Bucharest, Romania), p. 2, September.
- Graybeal, J., Isenor, A.W., Rueda, C., 2012. Semantic mediation of vocabularies for ocean observing systems. *Comput. Geosci.* 40, 120–131.
- Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., He, Y., 2016. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res.* 45, D347–D352.
- Jonquet, C., Shah, N.H., Musen, M.A., 2009. The open biomedical annotator. In: American Medical Informatics Association Symposium on Translational Bioinformatics, AMIA-TBI'09, San Francisco, CA, USA, pp. 56–60, March.
- Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N.F., Musen, M.A., Shah, N.H., 2011. NCBO resource index: ontology-based search and mining of biomedical resources, web semantics. In: 1st Prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC'10, Shanghai, China, vol. 9, pp. 316–324, September.
- Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F., 2013. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semantic Web* 4 (3), 277–284.
- Rueda, C., Bermudez, L., Fredericks, J., 2009. The MMI ontology registry and repository: a portal for marine metadata interoperability. In: MTS/IEEE Biloxi - Marine Technology for Our Future: Global and Local Challenges, OCEANS'09, Biloxi, MS, USA, pp. 6, October.
- D.A., Pouchard, L., Huhns, M., 2012. Lessons learned in deploying a cloud-based knowledge platform for the ESIP Federation. In: American Geo-physical Union Fall Meeting, poster session, San Francisco, USA, December.
- Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S., 2016. SIFR BioPortal: Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16, Genève, Suisse, pp. 16, July.
- Jonquet, C., Emonet, V., Musen, M.A., 2015. Roadmap for a multilingual BioPortal. In: In: Gracia, J., McCrae, J., Vulcu, G. (Eds.), 4th Workshop on the Multilingual Semantic Web, MSW4'15, vol. 1532. CEUR Workshop Proceedings, Portoroz, Slovenia, pp. 15–26.
- Matteis, L., Chibon, P., Espinosa, H., Skofic, M., Finkers, R., Bruskiewicz, R., Arnaud, E., 2015. Crop ontology: vocabulary for crop-related concepts. In: In: Larmande, P., Arnaud, E., Mougnot, I., Jonquet, C., Libourel, T., Ruiz, M. (Eds.), 1st International Workshop on Semantics for Biodiversity, vol. 1. CEUR Workshop Proceedings, Montpellier, France, pp. 37–46.
- Noy, N.F., Tudorache, T., Nyulas, C., Musen, M.A., 2010. The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies. In: AMIA Annual Symposium, Washington DC, USA, pp. 552–556, November.
- Jaiswal, P., Cooper, L., Elser, J.L., Meier, A., Laporte, M.-A., Mungall, C., Smith, B., Johnson, E.K., Seymour, M., Preece, J., Xu, X., Kitchen, R.S., Qu, B., Zhang, E., Arnaud, E., Carbon, S., Todorovic, S., Stevenson, D.W., 2016. Planteome: A resource for Common Reference Ontologies and Applications for Plant Biology. In: 24th Plant and Animal Genome Conference, PAG'16, San Diego, USA, January.
- Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25 (2), 288–289.
- Schmachtenberg, M., Bizer, C., Paulheim, H., 2014. Adoption of the linked data best practices in different topical domains. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C., (Eds.), 13th International Semantic Web Conference, ISWC'14, vol. 8796 of Lecture Notes in Computer Science, Riva del Garda, Italy, Springer, pp. 245–260, October.
- Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J., 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Biomed. Inf.* 41, 706–716.
- Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laipe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novere, N., Parkinson, H., Birney, E., Jenkinson, A.M., 2014. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 30, 1338–1339.
- Groth, P., Loizou, A., Gray, A.J., Goble, C., Harland, L., Petifer, S., 2014. API-centric linked data integration: the open PHACTS discovery platform case study. *Web Semantics* 29, 12–18.
- Venkatesan, A., Hassouni, N.E., Philippe, F., Pommier, C., Quesneville, H., Ruiz, M., Larmande, P., 2015. Exposing French agronomic resources as linked open data. In: 8th Semantic Web Applications and Tools for Life Sciences International Conference, SWAT4LS'15, vol. 546 of CEUR Workshop Proceedings, Cambridge, UK, pp. 205–207, December.
- Hassani-Pak, K., Zorc, M., Taubert, J., Rawlings, C., 2013. QTLNetMiner - candidate gene discovery in plant and animal knowledge networks. In: 21st Plant & Animal Genome Conference, poster session, San Diego, USA, pp. P0980, January.
- Dzálé-Yuemo, E., et al., 2017. Developing data interoperability using standards: A wheat community use case, F1000 Research, 6–1843, October 2017. (In preparation).
- Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., Arnaud, E., 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Frontiers Physiol.* 3.
- Coletta, R., Castanier, E., Valduriez, P., Frisch, C., Ngo, D., Bellahsene, Z., 2012. Public data integration with Webmatch. In: Raschia, G., Theobald, M., (Eds.), 1st International Workshop on Open Data, WOD'12, Nantes, France, pp. 5–12, ACM, May.
- Castanier, E., Jonquet, C., Melzi, S., Larmande, P., Ruiz, M., Valduriez, P., 2014. Semantic

- annotation workflow using bio-ontologies. In: Workshop on Crop Ontology and Phenotyping Data Interoperability, Montpellier, France, CGIAR, April.
- Noy, N.F., Alexander, P.R., Harpaz, R., Whetzel, P.L., Fergerson, R.W., Musen, M.A., 2013. Getting lucky in ontology search: a data-driven evaluation framework for ontology ranking. In: 12th International Semantic Web Conference, ISWC'13, vol. 8218 of Lecture Notes in Computer Science, Sydney, Australia, pp. 444–459, Springer, October 2013.
- Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N.H., Musen, M.A., 2009. What four million mappings can tell you about two hundred ontologies. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K., (Eds.), 8th International Semantic Web Conference, ISWC'09, vol. 5823 of Lecture Notes in Computer Science, Washington DC, USA, pp. 229–242, Springer, November 2009.
- Faria, D., Jiménez-Ruiz, E., Pesquita, C., Santos, E., Couto, F.M., 2014. Towards annotating potential incoherences in biportal mappings. In: 13th International Semantic Web Conference, ISWC'13, vol. 8797 of Lecture Notes in Computer Science, Riva del Garda, Italy, Springer, pp. 17–32.
- Pathak, J., Chute, C.G., 2009. Debugging mappings between biomedical ontologies: preliminary results from the NCBO BioPortal mapping repository. In: Smith, B., (Ed.), International Conference on Biomedical Ontology, Buffalo, NY, USA, pp. 95–98, July.
- Ghazvinian, A., Noy, N.F., Musen, M.A., 2009. Creating mappings for ontologies in biomedicine: simple methods work. In: American Medical Informatics Association Annual Symposium, AMIA'09, Washington DC, USA, pp. 198–202, November.
- Noy, N.F., Dorf, M., Griffith, N.B., Nyulas, C., Musen, M.A., 2009. Harnessing the power of the community in a library of biomedical ontologies. In: Clark, T., Luciano, J.S., Marshall, M.S., Prud'hommeaux, E., Stephens, S., (Eds.), Workshop on Semantic Web Applications in Scientific Discourse, SWASD'09, vol. 523 of CEUR Workshop Proceedings, Washington DC, USA, pp. 11, CEUR-WS.org, November.
- Noy, N.F., Griffith, N.B., Musen, M.A., 2008. Collecting community-based mappings in an ontology repository. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W., Thirunarayan, K. (Eds.), 7th International Semantic Web Conference, ISWC'08, vol. 5318 of Lecture Notes in Computer Science, Karlsruhe, Germany, October. Springer, pp. 368–371.
- McCray, A.T., 2003. An upper-level ontology for the biomedical domain. *Comp. Funct. Genomics* 4, 80–84.
- Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F., 2008. An efficient solution for mapping free text to ontology terms. In: AMIA Symposium on Translational Bioinformatics, AMIA-TBI'08, San Francisco, CA, USA, March.
- Martinez-Romero, M., Jonquet, C., O'Connor, M.J., Graybeal, J., Pazos, A., Musen, M.A., 2017. NCBO ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Biomed. Semantics* 8 (21).
- Jonquet, C., Musen, M.A., Shah, N.H., 2010. Building a biomedical ontology recommender web service, biomedical semantics, vol. 1. Selected in Pr. R. Altman's 2011 Year in Review at AMIA TBI.
- Salvadores, M., Horridge, M., Alexander, P.R., Fergerson, R.W., Musen, M.A., Noy, N.F., 2012. Using SPARQL to query biportal ontologies and metadata. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J., Hendler, J., Schreiber, G., Bernstein, A. (Eds.), 11th International Semantic Web Conference, ISWC'12, vol. 7650 of Lecture Notes in Computer Science, Boston, MA, USA, November. Springer, pp. 180–195.
- Maguire, E., González-Beltrán, A., Whetzel, P.L., Sansone, S.-A., Rocca-Serra, P., 2013. OntoMaton: a Biportal powered ontology widget for Google spreadsheets. *Bioinformatics* 29, 525–527.
- Jupp, S., Welter, D., Burdett, T., Parkinson, H., Malone, J., 2015. Collaborative ontology development using the webulous architecture and google app. In: Malone, J., Stevens, R., Forsberg, K., Splendiani, A., (Eds.), 8th International Conference on Semantic Web Applications and Tools for Life Sciences, SWAT4LS'15, vol. 1546, Cambridge, UK, pp. 120–121, December.
- Wolstencroft, K., Horridge, M., Owen, S., Mueller, W., Bacall, F., Snoep, J., Krebs, O., Goble, C., 2010. Rightfield: embedding ontology term selection into spreadsheets for the annotation of biological data. In: Polleres, A., Chen, H., (Eds.), 9th International Semantic Web Conference, Posters & Demonstrations, ISWC'10, vol. 658 of CEUR Workshop Proceedings, Shanghai, China, pp. 141–144, November.
- Roeder, C., Jonquet, C., Shah, N.H., Jr, W.A.B., Hunter, L., 2010. A UIMA wrapper for the NCBO annotator. *Bioinformatics* 26, 1800–1801.
- Adamusiak, T., Burdett, T., van der Velde, K.J., Abeygunawardena, N., Antonakaki, D., Parkinson, H., Swertz, M., 2010. OntoCAT – a simpler way to access ontology resources. In: ISMB Conference, Poster session, ISMB'10, Nature Precedings, Boston, MA, USA, July.
- Miñarro-Giménez, J.A., Mikel, J.T.F.-B., Aranguren, E., Antezana, E., 2012. NCBO-galaxy: bridging the BioPortal web services and the Galaxy platform. In: Cornet, R., Stevens, R., (Eds.), 3rd International Conference on Biomedical Ontologies, ICB'12, vol. 897 of CEUR Workshop Proceedings, Graz, Austria, pp. 2, July.
- Falconer, S.M., Callendar, C., Storey, M.-A., 2009. FLEXVIZ: visualizing biomedical ontologies on the web. In: Smith, B., (Ed.), International Conference on Biomedical Ontology, ICB'09, Buffalo, NY, USA, pp. 2, July.
- Slater, L., Gkoutos, G.V., Schofield, P.N., Hoehndorf, R., 2016. Using AberOWL for fast and scalable reasoning over BioPortal ontologies. *Biomed. Semantics* 7, 49.
- Melzi, S., Jonquet, C., 2014. Scoring semantic annotations returned by the NCBO Annotator. In: Paschke, A., Burger, A., Romano, P., Marshall, M., Splendiani, A., (Eds.), 7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14, vol. 1320 of CEUR Workshop Proceedings, Berlin, Germany, pp. 15, CEUR-WS.org, December.
- Park, J., Oh, S., Ahn, J., 2011. Ontology selection ranking model for knowledge reuse. *Expert Syst. Appl.* 38 (5), 5133–5144.
- Toulet, A., Emonet, V., Jonquet, C., 2016. Modèle de métadonnées dans un portail d'ontologies. In: Diallo, G., Kazar, O., (Eds.), 6èmes Journées Francophones sur les Ontologies, JFO'16, Bordeaux, France, October. Best paper award.
- Ngo, D., Bellahsene, Z., 2012. YAM+ + : a multi-strategy based approach for ontology matching task. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (Eds.), 18th International Conference on Knowledge Engineering and Knowledge Management, EKAW'12, vol. 7603 of Lecture Notes in Computer Science, Galway City, Ireland, October. Springer, pp. 421–425.
- Pesce, V., Maru, A., Keizer, J., 2011. The CIARD RING, an infrastructure for interoperability of agricultural research information services. *Agric. Inf. Worldwide* 4 (1), 48–53.
- Baker, T., Caracciolo, C., Arnaud, E., 2016. Global agricultural concept scheme (GACS)a hub for agricultural vocabularies. In: Jaiswal, P., Hoehndorf, R., (Eds.), 7th International Conference on Biomedical Ontologies, ICB'16, Poster Session, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 2, August.
- McGuinness, D.L., 2003. Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential, ch. Ontologies Come of Age. MIT Press, pp. 171–194.



Contents lists available at ScienceDirect

# Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: <http://www.elsevier.com/locate/websem>



## NCBO Resource Index: Ontology-based search and mining of biomedical resources

Clement Jonquet<sup>a,b,\*</sup>, Paea LePendur<sup>a</sup>, Sean Falconer<sup>a</sup>, Adrien Coulet<sup>a,c</sup>, Natalya F. Noy<sup>a</sup>, Mark A. Musen<sup>a</sup>, Nigam H. Shah<sup>a,\*</sup>

<sup>a</sup>Stanford Center for Biomedical Informatics Research, Stanford University, 251 Campus Drive, Stanford, CA 94305-5479, USA

<sup>b</sup>Laboratory of Informatics, Robotics, and Microelectronics of Montpellier (LIRMM), University of Montpellier, 161 Rue Ada, 34095 Montpellier, Cedex 5, France

<sup>c</sup>Lorraine Informatics Research and Applications Laboratory (LORIA), INRIA Nancy, Grand-Est, Campus Scientifique – BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France

### ARTICLE INFO

#### Article history:

Received 4 February 2011

Received in revised form 16 June 2011

Accepted 17 June 2011

Available online 29 June 2011

#### Keywords:

Ontology-based indexing

Semantic annotation

Information mining

Information retrieval

Biomedical data

Biomedical ontologies

### ABSTRACT

The volume of publicly available data in biomedicine is constantly increasing. However, these data are stored in different formats and on different platforms. Integrating these data will enable us to facilitate the pace of medical discoveries by providing scientists with a unified view of this diverse information. Under the auspices of the National Center for Biomedical Ontology (NCBO), we have developed the Resource Index – a growing, large-scale ontology-based index of more than twenty heterogeneous biomedical resources. The resources come from a variety of repositories maintained by organizations from around the world. We use a set of over 200 publicly available ontologies contributed by researchers in various domains to annotate the elements in these resources. We use the semantics that the ontologies encode, such as different properties of classes, the class hierarchies, and the mappings between ontologies, in order to improve the search experience for the Resource Index user. Our user interface enables scientists to search the multiple resources quickly and efficiently using domain terms, without even being aware that there is semantics “under the hood.”

© 2011 Elsevier B.V. All rights reserved.

### 1. Introduction

Researchers in biomedicine produce and publish enormous amounts of data describing everything from genomic information and pathways to drug descriptions, clinical trials, and diseases. These data are stored on many different databases accessible through Web sites, using idiosyncratic schemas and access mechanisms. Our goal is to enable a researcher to browse and analyze the information stored in these diverse resources. Then, for instance, a researcher studying allelic variations in a gene can find all the pathways that the gene affects, the drug effects that these variations modulate, any disease that could be caused by the gene, and the clinical trials that involve the drug or diseases related to that specific gene. The information that we need to answer such questions is available in public biomedical resources; the problem is finding that information.

The research community agrees that terminologies and ontologies are essential for data integration and translational discoveries to occur [1–3]. However, the metadata that describe the information in data resources are usually unstructured, often come in the form of free-text descriptions, and are rarely labelled or tagged

using terms from ontologies that are available for the domains. Users often prefer labels from ontologies because they provide a clear point of reference during their search and mining tasks [4–6]. For example, researchers and curators widely use the Gene Ontology to describe the molecular functions, cellular location, and biological processes of gene products. These annotations enable the integration of the descriptions of gene products across several model organism databases [7].

However, besides these examples, semantic annotation of biomedical resources is still minimal and is often restricted to a few resources and a few ontologies [8]. Usually, the textual content of these online resources is indexed (e.g., using Lucene) to enable querying the resources with keywords. However, there are obvious limits to keyword-based indexing, such as the use of synonyms, polysemy, lack of domain knowledge. Furthermore, having to perform keyword searches at each Web site individually makes the navigation and aggregation of the available information extremely cumbersome, if not impractical. Search engines, like Entrez ([www.ncbi.nlm.nih.gov/Entrez](http://www.ncbi.nlm.nih.gov/Entrez)), facilitate search across several resources, but they do not currently use as many of the available and relevant biomedical ontologies.

The National Center for Biomedical Ontology (NCBO) Resource Index addresses these two problems by (1) providing a unified index of and access to multiple heterogeneous biomedical resources; and (2) using ontologies and the semantic representation that they

\* Corresponding authors. Tel.: +1 650 725 6236; fax: +1 650 725 7944.

E-mail addresses: [jonquet@lirmm.fr](mailto:jonquet@lirmm.fr) (C. Jonquet), [nigam@stanford.edu](mailto:nigam@stanford.edu) (N.H. Shah).

encode to enhance the search experience for the user. The NCBO BioPortal – an open library of more than 200 ontologies in biomedicine [9] – serves as the source of ontologies for the Resource Index. We use the terms from these ontologies to annotate, or “tag,” the textual descriptions of the data that reside in biomedical resources and we collect these annotations in a searchable and scalable index (Fig. 1). The key contributions to the field are (i) to build the search system for such an important number of ontologies and resources and (ii) to use the semantics that the ontologies encode.

In the context of our research, we call data *element* any identifiable entity or record (e.g., document, article, experimentation report) which belongs to a biomedical data *resource* (e.g., database of articles, experiments, trials). Usually, an element has an identifier and can be linked by a URL. For instance, the trial NCT00924001 is an element of the ClinicalTrials.gov data resource that can be accessed with: <http://clinicaltrials.gov/ct2/show/NCT00924001>. We call *annotation* – a central component – a link from an ontology term to a data element, indicating that the data element refers to the term either explicitly or not [10,11]. We then use these annotations to “bring together” the data elements.

We currently index 22% resources, which are maintained by a variety of different institutions, with terms from more than 200 ontologies included in BioPortal (Appendix A). As of January 2011, our 1.5 Tb MySQL database, which stores the annotations in the Resource Index, contains 11 Billion annotations, 3.3 Million ontology concepts, and 3.2 Million data elements. The user interface is available at <http://bioportal.bioontology.org/resources>.

A preliminary version of the system was presented in [12]. In this paper, we illustrate use case scenarios (Section 2), describe the system implementation (Section 3) and the details of the indexing workflow (Section 3.3), and the different means to access the Resource Index (Section 3.4). We demonstrate how semantic technologies enable information retrieval and mining scenarios that were not possible otherwise (Section 4).

## 2. Use case scenarios

We will describe the functionality of the Resource Index through three use case scenarios.

**Scenario 1: Multiple-term search across resources.** The user is interested in the role of tumor protein p53 in breast cancer. He can search the Resource Index for “Tumor Protein p53” AND “Breast Carcinoma” as defined in the NCI Thesaurus (Fig. 2). The search results summarize the number of elements per resources anno-



Fig. 2. Resource Index user interface. The search for resources that contain both “Tumor Protein p53” AND “Breast Carcinoma.”

tated with both terms. The user can see there is relevant data linking p53 to breast cancer in such resources as ArrayExpress, ClinicalTrials.org, Gene Expression Omnibus (GEO), Stanford Microarray Database (SMD) and others. He can access the data elements within each resource quickly and navigate between resources.

**Scenario 2: Exploratory search across resources.** A researcher studying the causes and treatments for stroke in humans is interested in learning more about the genetic basis of the response to related conditions by searching the literature. She already knows that some related conditions such as stroke, tran-

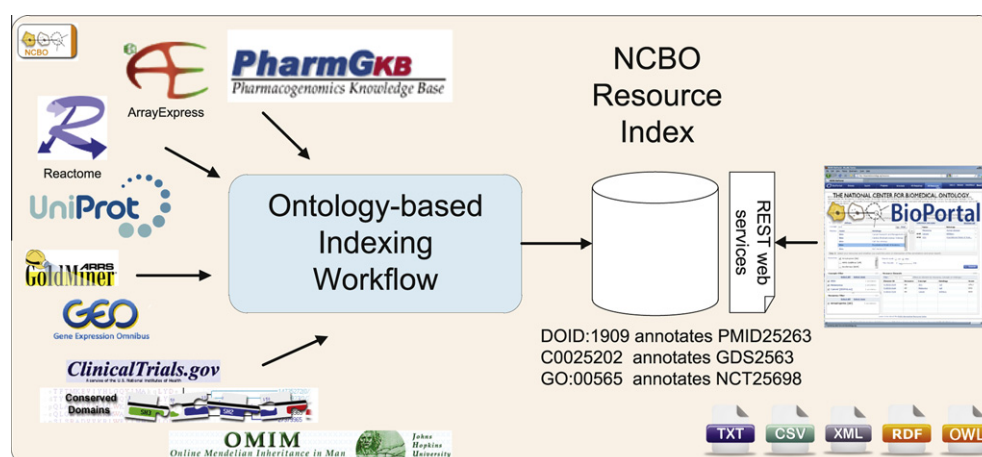


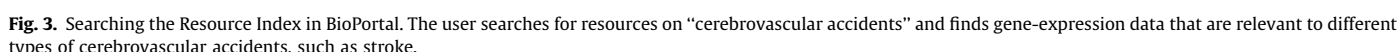
Fig. 1. NCBO Resource Index overview. We process each biomedical resource using the ontology-based indexing workflow. We store the resulting annotations in a database and make them available in several formats via REST Web services. BioPortal provides userfriendly interfaces to search and navigate the Resource Index.

**Scenario 3: Semantically enriched search across resources.** The user wants to search gene expression data about “retroperitoneal neoplasms.” A direct keyword search with “retroperitoneal neoplasm” on the GEO Web site will return no results. However, there are several datasets in GEO about “pheochromocytoma” and “renal cell cancer” both of which are retroperitoneal neoplasms and thus relevant to the previous search. When our user queries the Resource Index with “retroperitoneal neoplasm,” he will get the results that use the hierarchy represented in the BioPortal

In the next section, we describe the implementation of the Resource Index, which enables these use cases.

To create the Resource Index, we process metadata describing data elements in a variety of heterogeneous resources to create semantic annotations of these metadata. We use the publicly available biomedical ontologies in BioPortal as a source of terms, their synonyms, and the relations between terms (Section 3.1). We use resource-specific access tools to process metadata that describe data elements in different resources (Section 3.2). We use an off-the-shelf concept-recognition tool to identify terms from BioPortal ontologies within the textual metadata and annotate, or tag, the corresponding element with the recognized terms. We expand these annotations using available ontology knowledge (Section 3.3). Finally, the Web services and user interface provide users with fast and scalable access to this index and support different use cases such as information retrieval and mining (Section 3.4).

BioPortal, an open library of biomedical ontologies [9], provides uniform access to the largest collection of publicly available bio-



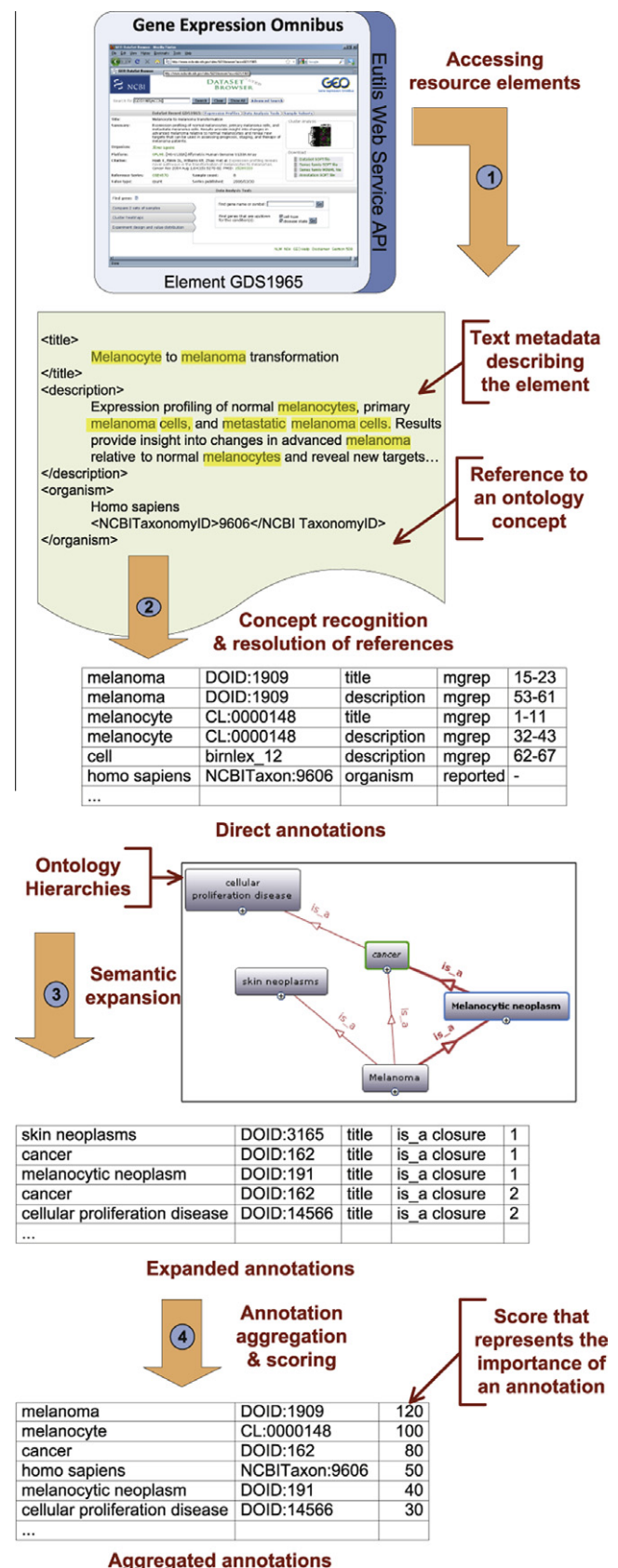
medical ontologies. At the time of this writing, there are 245 ontologies in this collection. BioPortal users can browse, search, visualize, and comment on ontologies both interactively, through a Web interface, and programmatically, via Web services. The majority of BioPortal ontologies were contributed by their developers directly to BioPortal. A number of ontologies come from Open Biomedical Ontologies (OBO) Foundry [13], a collaborative effort to develop a set of interoperable ontologies for biomedicine. BioPortal also includes publicly available terminologies from the Unified Medical Language System (UMLS), a set of terminologies which are manually integrated and distributed by the United States National Library of Medicine [14]. BioPortal includes ontologies that are developed in a variety of formats, including OWL, RDF(S), OBO (which is popular with many developers of biomedical ontologies), and RRF (which is used to distribute UMLS terminologies). BioPortal provides a uniform set of REST Web services to access basic lexical and structural information in ontologies represented in these heterogeneous formats.

We use the BioPortal REST services to traverse the ontologies and to create a *dictionary* of terms to use for direct annotations of data elements in biomedical resources. We use preferred name and synonym properties of classes for this dictionary. Some ontology formats have preferred name and synonym properties as part of the format (e.g., OBO and RRF). For OWL, ontology developers can either use the relevant SKOS properties to represent this information, or specify in the ontology metadata which are the properties that they use for preferred names (e.g., `rdfs:label`) and synonyms. Currently, our dictionary contains 6,835,997 terms, derived from the 3,349,338 concepts from 206 ontologies (the subset of BioPortal ontologies that are usable for annotation). We identify each concept by a URI defined in the original ontology or provided by NCBO.

### 3.2. Accessing biomedical resources

In addition to the ontology terms, the data elements from the biomedical resources are another major source of information for the Resource Index (Fig. 1). As of January 2011, we have indexed 22% public biomedical resources of different sizes (up to 3.2 Million elements and 1.4 Gb of data). We provide a list of sample resources in Appendix A. Data resources provide their data in idiosyncratic formats (often XML) and offer different means of access (often Web services). To access the information in the resources, we build a custom *wrapper* for each resource. The wrapper extracts the fields describing the data elements within a resource as illustrated in step 1 of Fig. 4. In developing each wrapper, we work with a subject matter expert to determine which textual metadata fields (later called contexts) we must process (e.g., title, description). We also assign each context a weight [0, 1] representing the importance of the field. We later use this weight to score annotations.<sup>1</sup> For example, we may give annotations appearing in the title a higher weight based on the expert's recommendation for that resource. In some cases, resources already tag elements with ontology terms, so the wrapper directly extracts the curated annotations and applies an appropriate weight. We call these annotations *reported annotations*. For example, the description of gene-expression data in GEO contains an *organism* field where a domain expert manually puts a term from the National Center for Biotechnology Information taxonomy, which refers to the relevant organism.

Our resource-specific wrappers access the data elements incrementally, enabling us to process only the data elements that were



**Fig. 4.** Example of annotations generated for a GEO element. *Direct annotations* are generated from textual metadata and already existing ontology references of the data element. Then, *expanded annotations* are created using the ontology *is\_a* hierarchy. Finally, all the annotations are *aggregated* and scored taking into consideration their frequency and context.

<sup>1</sup> Researchers have previously demonstrated the influence and importance of the original context in which a term appears on information retrieval [4].

added to the resource since the last time that we processed the resource.

### 3.3. Ontology-based annotation

After we access the data elements describing the resource, we perform the following steps to create annotations for the data elements in the resource: (a) direct annotation with ontology terms; (b) semantic expansion of annotations; (c) aggregation and scoring of annotations (Fig. 4).

**(a) Creation of direct annotations.** We process each textual metadata using a *concept-recognition* tool that detects the presence of concepts in text. Our workflow accepts different concept recognizers ranging from simple string matching techniques to advanced natural language processing algorithms. We currently use Mgrep [15,16] which enables fast and efficient exact matching against a very large set of input strings (however without any advanced natural language processing (e.g., stemming, permutation, morphology)). Concept recognizers usually use a *dictionary*. The dictionary (or lexicon) is a list of strings that correspond to preferred names and synonyms of ontology concepts. At this step, Mgrep uses the 6.8 Million terms dictionary built before. In the example in Fig. 4, the recognizer identifies the terms *melanoma*, *melanocyte*, and *cell* and creates a set of *direct annotations* with the corresponding concepts in the Human Disease, Cell type, and BIRN Lex ontologies. We preserve the identified term, the context in which it appears, and its character position as provenance information about the annotation.

**(b) Semantic expansion of annotations.** After direct annotations step, several semantic-expansion components leverage the knowledge in the ontologies to create *expanded annotations* from the direct annotations.

First, the *is\_a transitive closure* component traverses an ontology subclass–superclass hierarchy using a customized algorithm to create new annotations with superclasses of the classes that appear in direct annotations. We used the subclass transitive relation as defined by the original ontology e.g., *is\_a* (OBO), *rdfs:subClassOf* (OWL) and abstracted by BioPortal to compute the transitive closure on the whole ontology graph. For instance, we will expand a direct annotation of a data element with the concept *melanoma* from NCI Thesaurus, to annotations with *melanocytic neoplasm*, *cancer*, and *cellular proliferation disease* because NCI Thesaurus defines *melanoma* as a subclass of *melanocytic neoplasm*, which in turn is a subclass of *cellular proliferation disease* (Fig. 4). We preserve the shortest ancestor level (direct parent, grandparent, etc.) as provenance information to use for scoring annotations. Naturally, the farther away the ancestor term is from the term in the direct annotation, the less relevant the corresponding expanded annotation is.

Second, the *ontology-mapping* component creates new annotations based on existing mappings between ontologies. BioPortal provides point-to-point mappings between terms in different ontologies. Some of these mappings were defined manually and some were created automatically using various mapping algorithms [17].<sup>2</sup> We use the mappings that BioPortal stores and provides to expand our annotations and we do not follow them transitively. For instance, if a text is directly annotated with the concept *treatment* in Medical Subject Headings (MeSH), the mapping component will generate a new annotation with the concept *therapeutic procedure* from Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) because there is a mapping between these two terms in BioPortal. We preserve the type of mapping as

provenance information to use for scoring annotations. It allows to score those expanded annotations proportionally to the mapping confidence (e.g., *owl:sameAs*, *skos:exactMatch*, *skos:closeMatch*, manually curated or automatically generated).

**(c) Annotation aggregation and scoring.** We use the provenance information that we collect in creating direct and expanded annotations to assign each annotation a weight from 0 to 10 representing its relevance. For example, a match based on a preferred label gets a weight of 10 versus a synonym, which gets an 8; a match originating from a mapping gets a weight of 7 whereas one from an *is\_a* relationship get a diminishing weight based on ancestor level. Because several annotations with the same concept but with different provenance and context can co-exist we aggregate all those annotations of an element to a unique pair [concept–element], called *aggregated annotation*, to which a score is assigned. Those are the annotations used for searches. The scoring algorithm takes into account frequency, provenance and context of the annotation by doing the sum of the weights assigned to each annotation normalized by the weights of the original contexts.

At each step, the annotation workflow populates several relational tables and stores the *detailed* (direct & expanded) and *aggregated annotations*. Because both ontologies and resources are changing often, we need to automatically update the Resource Index tables regularly. The workflow handles (i) *resource updates* (i.e., incremental processing of new elements added to resources) using wrappers that pull only the data elements that have not been processed yet and (ii) *ontology updates* (i.e., incremental processing of new ontologies and new ontology versions) because BioPortal provides version specific identifiers for ontologies. For simplicity, when a new ontology version is added to BioPortal, the previous annotations associated with the ontology are removed from the Resource Index and new ones are added. The indexing workflow has been specifically optimized for this to occur rapidly [18]. We run these two different updates respectively weekly and monthly.

### 3.4. Accessing the NCBO Resource Index

The annotation and the scores that we described in the previous section constitute the Resource Index. The index contains 3 Billion aggregated annotations and 11 Billion detailed annotations (10% direct 90% expanded) as illustrated by Fig. 5. We provide both a Web service access to the index and a special-purpose easy-to-use graphical user interface, which enables domain experts to explore and analyze the information in the Resource Index.

The main Resource Index user interface, illustrated in Figs. 2 and 3, is a search-based interface geared towards biomedical end-users. Users do not even need to be aware that semantic technologies are driving the user interface, and can use it through a simple search-box mechanism. As the user types in terms that she is interested in, she gets a list of auto-complete suggestions for the search terms and the source ontologies for these terms. Users can search data elements using AND and OR constructs.<sup>3</sup> She is presented with a list of search results (as snippets) as well as a tag cloud of related terms (selected in the top 10 results) to help refine her search further. For each identified element, a user can see the details of the annotations highlighted in the original text and link back to the URLs of the original data elements.

Users can retrieve the content of the Resource Index programmatically by calling a Web service and specifying either *ontology concepts* or specific *data elements* that they are looking for. Specifically, we provide the following services:

<sup>2</sup> In this work we assume mappings between ontologies already exists, the creation of biomedical mappings is discussed in numerous other papers.

<sup>3</sup> The OR construct is currently available only through Web service; it is not available through the graphical user interface.

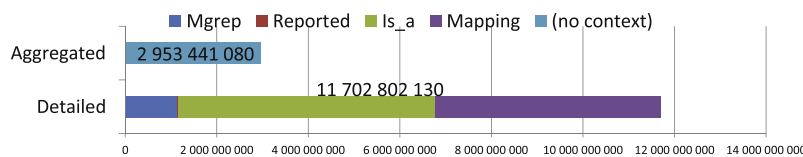


Fig. 5. Number and types of annotations in the Resource Index.

1. For a given concept, obtain the set of elements in one or several resources annotated with this concept (e.g., GEO and ArrayExpress elements annotated with concept `DOID:1909`).
2. For several concepts, obtain the union or intersection of the set of elements annotated with these concepts (e.g., GEO and ArrayExpress elements annotated with both `DOID:1909` and `CL:0000148`).
3. For a given data element, obtain the set of concepts in one or several ontologies annotating this element (e.g., NCI Thesaurus concepts annotating the GEO dataset `GDS1965`).

The first two information-retrieval services offer a unique endpoint to query several heterogeneous data resources and facilitate data integration (defined as *view integration* in Goble & Stevens [3]). The third service supports the type of exploration that the original resource may have never supported. This use case enables users to gather more information about a data element that they have already identified.

When retrieving annotations for a given element, users can filter out annotations using several mechanisms, such as limiting results to annotations with specific UMLS semantic types, using only results that match the whole word in the query, disabling the results obtained by matching synonyms, or selecting the type of mapping used for expanding annotations. Users can retrieve annotations in several formats (text, tab delimited, XML, RDF and OWL). The results are ordered by the scores assigned during the indexing phase.

#### 4. Discussion and related work

The Resource Index provides semantically-enabled uniform access to a large set of heterogeneous biomedical resources. It leverages the semantics expressed in the ontologies in several different ways:

**Preferred names and synonyms:** Many biomedical ontologies specify, as class properties, not only labels (preferred names) but also synonyms for the class names, which we use during annotation. For example, a keyword search of caNanoLab resource with “adriamycin” would normally obtain no results. However, because the ontologies that we use have defined “doxorubicin” as a synonym for “adriamycin,” the Resource Index retrieves all caNanoLab elements annotated with the term “doxorubicin.”

**Auto-complete:** As users type a term into the search box, they receive immediate feedback giving both preferred names and synonyms for matching classes from different ontologies.

**Hierarchies:** We use subclass relations to traverse ontology hierarchies to create expanded annotations, therefore improving the recall of search on general terms. For example, a search with “retroperitoneal neoplasm,” will retrieve data annotated with “pheochromocytoma” (Section 2). Notice that subclass relationships are present in all ontologies thus enable to provide the same feature for all ontologies. Specific ontology relationships are not considered, although we acknowledge there are often useful on a per-ontology approach.

**Mappings:** We use BioPortal mappings to expand the set of annotations. For example, a search with the concept “treatment” from MeSH retrieves the elements annotated with “therapeutic

procedure” in SNOMED-CT because there is a mapping between these two concepts in BioPortal.<sup>4</sup>

The use of ontologies significantly enhances recall of searches (i.e., more relevant data elements are retrieved) without affecting precision of the top results. Our aggregation and scoring addresses the issue of precision by ranking relevant results for the user e.g., the algorithm ranks the direct matches higher over the ones obtained via semantic expansion. Semantic disambiguation is not handled yet e.g., someone searching elements for “Cell” in NCI Thesaurus will obtain the elements mentioning the word ‘cell’ as the abbreviation of cell phone. However, given the characteristics of the resources indexed (biomedical databases as opposed to general Web sites) the issue has not come up in practice.

Because the goal of the Resource Index is to improve runtime information retrieval and data-mining tasks, we decided to precompute inferences with ontologies (i.e., *is\_a* and mapping expansion) rather than to implement semantic query-expansion algorithms [19] that would have computed inferences dynamically but would have required longer response time. Our technical decisions in terms of design and architecture were often driven by benchmarking analysis and metrics [18]. The indexing workflow execution times range from a couple of minutes for the small resources to more than a week for the biggest one. Because it is impossible to include in the Resource Index all possible biomedical resources, NCBO provides the ontology-based annotation workflow as a Web service [8], the NCBO *Annotator*, which allows researchers to annotate their text data automatically and get the annotations back. They can use this service to develop their own semantic-search applications. Researchers at the Medical College of Wisconsin have already created one such application for mining associations between gene expression levels and phenotypic annotations for microarray data from GEO (cf. <http://gminer.mcw.edu>).

Semantic annotation is an important research topic in the Semantic Web community [10]. Tools vary along with the types of documents that they annotate (e.g., image annotation [20]). For an overview and comparison of semantic annotation tools the reader may refer to the study by Uren and colleagues [11].

As we have mentioned earlier, our annotation workflow can be configured to use any concept-recognition tool. A number of publicly available concept recognizers identify entities from ontologies or terminologies in text. These recognizers include IndexFinder [21], SAPHIRE [22], CONANN [23], and the University of Michigan’s Mgrep [15]. The National Library of Medicine (NLM)’s MetaMap [24], which identifies UMLS Metathesaurus concepts in text, is generally used as the gold standard for evaluating tools in the biomedical domain. Many of these tools are not under active development and are restricted to a particular ontology or the UMLS.

Related tools in the biomedical domain include Terminizer [25], which is an annotation service similar to the NCBO Annotator. Terminizer recognizes concept names and synonyms and their possible permutations but only for OBO ontologies. Terminizer does not allow any automatic semantic expansion of the annotations but allows refining annotations using broader or narrower terms in the user interface. Whatizit [26], which is a set of text mining Web

<sup>4</sup> Notice there is no composition of the semantic expansion components e.g., mapping ancestors are not used for annotations.

services that can recognize several types of entities such as protein and drug names, diseases, and gene products. Reflect [27], which highlights gene, protein, and small-molecule names and can perform the recognition in HTML as well as PDF and MS Word documents. The originality of Reflect, when used in a Web browser, is that the tool links the identified terms to corresponding entries in biomedical resources e.g., UniProt, DrugBank. However, the tool is not driven by ontologies and does not execute any semantic expansion.

We have conducted a comparative evaluation of two concept recognizers used in the biomedical domain – Mgrep and MetaMap – and found that Mgrep has clear advantages in large-scale service oriented applications, specifically addressing flexibility, speed and scalability [8]. The precision of concept recognition varies depending on the text in each resource and type of entity being recognized: from 93% for recognition biological processes in descriptions of gene expression experiments to 60% in clinical trials, or from 88% for recognizing disease terms in descriptions of gene expression experiments to 23% for PubMed abstract [8]. Other studies reported similar results [28,29]. The average precision is approximately 73%, average recall is 78%.

Most of the other annotation tools do not perform any semantic expansion, which gives the Resource Index and the Annotator a significant advantage. There are however other tools in the biomedical domain that use semantics internally including MedicoPort [30], which uses UMLS semantics to expand user queries; the work of Moskovitch and colleagues [4], who use ontologies for annotation (concept based search) and demonstrate the importance of the context (context-sensitive search) when annotating structured documents. HealthCyberMap [31] uses ontologies and semantic distances for visualizing biomedical resources information. Essie [32] shows that a judicious combination of exploiting document structure, phrase searching, and concept based query expansion is useful for domain optimized information retrieval. Finally, other studies such as Khelif and colleagues [33] illustrate the annotation of a specific resource with specific ontologies (the GeneRIF resource annotated with UMLS and Galen in this case).

Currently, we create annotations based only on textual fields. However, we can extend our approach to other kinds of documents (i.e., images, sounds) by changing the tool that we use for concept recognition. We currently process only text meta-data in English. However, as BioPortal now contains ontologies in multiple languages, we can start using concept recognizers for other languages in the future.

## 5. Challenges and future plans

We are currently working on expanding the Resource Index to include more resources. Our goal is to index up to 100 public resources, including PubMed, which provides access to all research articles in biomedicine (approximately 20 Million elements). We have analyzed the metrics on ontologies in order to re-structure the database backend for the Resource Index. This restructuring has enabled us to reduce the processing time for one of our larger datasets from one week to one hour [18]. With this type of optimizations, we can now annotate extremely large datasets such as PubMed. We have already indexed the last five years of it (20%). We note that since 2010, changes in MetaMap allow it to be deployed with ontologies outside of UMLS. We are investigating the possibility of including MetaMap as an alternative concept recognizer in the annotation workflow.

One limiting factor in increasing the number of resources that we index is the need to develop custom access tools for most resources. However, most resource access tools follow the same principles, so we have built templates that enable our collaborators to build them easily and quickly to process their own datasets and to include them in the Resource Index.

Our next challenge is to evaluate the user interface and to understand what works best for domain experts. We have performed small-scale formative evaluations, but will need to work on larger scale evaluation, with different groups of users.

## 6. Conclusions

We have presented an ontology-based workflow to annotate biomedical resources automatically as well as an index constructed using this workflow. Ontology-based indexing is not new in biomedicine, however it is usually restricted to indexing a specific resource with a specific ontology (vertical approach). We adopt a horizontal approach, accessing annotations for many important resources using a large number of ontologies. This approach follows the translational bioinformatics and Semantic Web vision to discover new knowledge by recombining already existing knowledge (i.e., resources and ontologies) in a manner that the knowledge providers have not previously envisaged.

The Resource Index enables domain experts to search heterogeneous, independently developed resources. While we use ontologies and semantics “under the hood” to improve the quality of the results and to simplify the user interaction, the users are not aware of this complexity. They use a simple search-box interface

**Table 1**  
A sample of ontologies included in the Resource Index. Please refer to <http://bioportal.bioontology.org/ontologies> for a complete listing.

















Ontology	Maintained By	Format	# Classes
 <b>NCI Thesaurus (NCIt)</b>	National Cancer Institute	OWL	80 K
 <b>Medical Subject Headings (MSH)</b>	National Library of Medicine	RRF	223 K
 <b>Gene Ontology (GO)</b>	GO Consortium	OBO	33 K
 <b>Systematized Nomenclature of Medicine-Clinical Terms (SNOMEDCT)</b>	International Health Terminology Standards Development Organisation	RRF	391 K
 <b>Medical Dictionary for Regulatory Activities (MDR)</b>	Maintenance and Support Services Organization	RRF	69 K
 <b>RadLex (RID)</b>	Radiological Society of North America	PROTEGE	30 K
 <b>International Classification of Diseases (ICD10)</b>	World Health Organization	RRF	12 K
 <b>NCBI organismal classification (NCBITaxon)</b>	National Center for Biotechnology Information	OBO	513 K
 <b>Mouse adult gross anatomy (MA)</b>	The Jackson Laboratory	OBO	3 K

Table 2

A sample of resources included in the Resource Index. Please refer to [http://rest.bioontology.org/resource\\_index/resources/list/](http://rest.bioontology.org/resource_index/resources/list?urlhttp://rest.bioontology.org/resource_index/resources/list/) for a complete listing.

	Resource/contexts indexed	Maintained By	Elements
	<b>Gene Expression Omnibus (GEO)</b> gene expression and molecular abundance repository /title, summary, organism	National Center for Biotechnology Information (NCBI)	23,287 (21 Mb)
	<b>ArrayExpress (AE)</b> microarray data and geneindexed expression profiles /name, description, species, experiment_type	European Bioinformatics Institute (EMBL-EBI)	16,444 (23.4 Mb)
	<b>caNanoLab (CaNano)</b> biomedical nanotechnology research results /Composition, Association, Method, etc.	National Cancer Institute's cancer Nanotechnology Lab	890 (19.1 Mb)
	<b>Adverse Event Reporting System Data (AERS)</b> adverse events reported to FDA by doctors and other professionals /Drug_char, Drug_names, Drug_admin_route	AersData.org	1,172,881 (278.4 Mb)
	<b>Clinical Trials (CT)</b> reports on clinical research in human volunteers /title, description, condition, intervention	ClinicalTrials.gov	101,606 (187.3 Mb)
	<b>Research Crossroads (RXRD)</b> medical funding data /title	ResearchCrossroads.org	1,033,651 (89.6 Mb)
	<b>UniProt KB (UPKB)</b> protein sequence and functional information /genesymbol, goAnnotationList, proteinName	UniProt.org	18,461 (4.2 Mb)

and can drill down on the specific resources that contain their terms of interest or any other relevant terms.

## Acknowledgements

This work was supported in part by the National Center for Biomedical Ontology, under roadmap-initiative Grant U54 HG004028 from the National Institutes of Health. The NCBO Resource Index won the First prize in the Semantic Web Challenge 2010 (<http://challenge.semanticweb.org/>).

## Appendix A. Lists of ontologies and resources

see Tables 1 and 2.

## References

- [1] O. Bodenreider, R. Stevens, Bio-ontologies: current trends and future directions, *Briefing in Bioinformatics* 7 (3) (2006) 256–274.
- [2] A.J. Butte, R. Chen, Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics, in: *American Medical Informatics Association Annual Symposium, AMIA'06*, Washington DC, USA, 2006, pp. 106–110.
- [3] C. Goble, R. Stevens, State of the nation in data integration for bioinformatics, *Biomedical Informatics* 41 (5) (2008) 687–693.
- [4] R. Moskovitch, S.B. Martins, E. Behiri, A. Weiss, Y. Shahar, A comparative evaluation of full-text, concept-based, and context-sensitive search, *American Medical Informatics Association* 14 (2) (2007) 164–174.
- [5] I. Spasic, S. Ananiadou, J. McNaught, A. Kumar, Text mining and ontologies in biomedicine: making sense of raw text, *Briefing in Bioinformatics* 6 (3) (2005) 239–251.
- [6] C.A. Sneiderman, D. Demner-Fushman, M. Fiszman, N.C. Ide, T.C. Rindflesch, Knowledge-based methods to help clinicians find answers in medline, *American Medical Informatics Association* 14 (6) (2007) 772–780.
- [7] S.Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations, *Nature Reviews Genetics* 9 (2008) 509–515.
- [8] N.H. Shah, N. Bhatia, C. Jonquet, D.L. Rubin, A.P. Chiang, M.A. Musen, Comparison of concept recognizers for building the Open Biomedical Annotator, *BMC Bioinformatics* 10 (9:S14).
- [9] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N.B. Griffith, C. Jonquet, D.L. Rubin, M.-A. Storey, C.G. Chute, M.A. Musen, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Research* 37 (2009) 170–173. web server.
- [10] S. Handschuh, S. Staab (Eds.), *Annotation for the Semantic Web*, vol. 96 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2003.
- [11] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna, Semantic annotation for knowledge management: requirements and a survey of the state of the art, *Web Semantics: Science, Services and Agents on the World Wide Web* 4 (1) (2006) 14–28.
- [12] N.H. Shah, C. Jonquet, A.P. Chiang, A.J. Butte, R. Chen, M.A. Musen, Ontology-driven Indexing of Public Datasets for Translational Bioinformatics, *BMC Bioinformatics* 10 (2:S1).
- [13] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, T.O. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R.H. Scheuermann, N.H. Shah, P.L. Whetzel, S. Lewis, The OBO foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology* 25 (11) (2007) 1251–1255.
- [14] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research* 32 (2004) 267–270.
- [15] W. Xuan, M. Dai, B. Mirel, B. Athey, S.J. Watson, F. Meng, Interactive medline search engine utilizing biomedical concepts and data integration, in: *BiolINK: Linking Literature, Information and Knowledge for Biology*, SIG, ISMB'08, Vienna, Austria, 2007, pp. 55–58.
- [16] M. Dai, N.H. Shah, W. Xuan, M.A. Musen, S.J. Watson, B.D. Athey, F. Meng, An efficient solution for mapping free text to ontology terms, in: *American Medical Informatics Association Symposium on Translational Bioinformatics, AMIA-TBI'08*, San Francisco, CA, USA, 2008.
- [17] N.F. Noy, N.B. Griffith, M.A. Musen, Collecting community-based mappings in an ontology repository, in: *7th International Semantic Web Conference*, vol. 5318 of *LNCS*, Springer, Karlsruhe, Germany, 2008, pp. 371–368.
- [18] O. LePendou, N.F. Noy, C. Jonquet, P.R. Alexander, N.H. Shah, M.A. Musen, Optimize first, buy later: analyzing metrics to ramp-up very large knowledge bases, in: *Ninth International Semantic Web Conference*, vol. 6496 of *LNCS*, Springer, Shanghai, China, 2010, pp. 486–501.
- [19] J. Bhogal, A. Macfarlane, P. Smith, A review of ontology based query expansion, *Information Processing and Management* 43 (2007) 866–886.
- [20] L. Hollink, G. Schreiber, J. Wielemaker, B. Wielinga, Semantic annotation of image collections, in: *Knowledge Markup and Semantic Annotation Workshop*, Sanibel, FL, USA, 2003.
- [21] Q. Zou, W.W. Chu, C. Morioka, G.H. Leazer, H. Kangarloo, IndexFinder: a method of extracting key concepts from clinical texts for indexing, in: *American Medical Informatics Association Annual Symposium, AMIA'03*, Washington DC, USA, 2003, pp. 763–767.
- [22] W.R. Hersh, R.A. Greenes, SAPHIRE – an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships, *Computers and Biomedical Research* 23 (5) (1990) 410–425.
- [23] L.H. Reeve, H. Han, CONANN: an online biomedical concept annotator, in: *4th International Workshop Data Integration in the Life Sciences*, vol. 4544 of *LNCS*, Springer-Verlag, Philadelphia, PA, USA, 2007, pp. 264–279.
- [24] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: *American Medical Informatics Association Annual Symposium, AMIA'01*, Washington, DC, USA, 2001, pp. 17–21.
- [25] D. Hancock, N. Morrison, G. Velarde, D. Field, Terminizer – assisting mark-up of text using ontological terms, in: *Third International Biocuration Conference*, Berlin, Germany, 2009.
- [26] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, A. Jimeno, Text processing through Web services: calling Whatizit, *Bioinformatics* 24 (2) (2008) 296–298.
- [27] E. Pafilis, S.I. O'Donoghue, L.J. Jensen, H. Horn, M. Kuhn, N.P. Brown, R. Schneider, Reflect: augmented browsing for the life scientist, *Nature Biotechnology* 27 (2009) 508–510.

- [28] J.S. Simon N. Twigger, Joey Geiger, Using the NCBO Web services for concept recognition and ontology annotation of expression datasets, in: *Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS'09*, vol. 559 of *CEUR Workshop Proceedings*, Amsterdam, The Netherlands, 2009.
- [29] I.N. Sarkar, Leveraging biomedical ontologies and annotation services to organize microbiome data from mammalian hosts, in: *American Medical Informatics Association Annual Symposium, AMIA'10*, Washington DC, USA, 2010, pp. 717–721.
- [30] A.B. Can, N. Baykal, MedicoPort: a medical search engine for all, *Computer Methods and Programs in Biomedicine* 86 (1) (2007) 73–86.
- [31] M.N. Kamel-Boulos, A first look at HealthCyberMap medical semantic subject search engine, *Technology and Health Care* 12 (2004) 33–41.
- [32] N.C. Ide, R.F. Loane, D. Demner-Fushman, Essie: a concept-based search engine for structured biomedical text, *American Medical Informatics Association* 14 (3) (2007) 253–263.
- [33] K. Khelif, R. Dieng-Kuntz, P. Barbry, An ontology-based approach to support text mining and information retrieval in the biological domain, *Universal Computer Science, Special Issue on Ontologies and their Applications* 13 (12) (2007) 1881–1907.



# The Open Biomedical Annotator

Clement Jonquet, PhD<sup>1</sup>, Nigam H. Shah, M.B.B.S, PhD<sup>1</sup> and Mark A. Musen, MD, PhD<sup>1</sup>

<sup>1</sup>Stanford Center for Biomedical Informatics Research and the National Center for Biomedical Ontology, Stanford University, Stanford, CA

## Abstract

*The range of publicly available biomedical data is enormous and is expanding fast. This expansion means that researchers now face a hurdle to extracting the data they need from the large numbers of data that are available. Biomedical researchers have turned to ontologies and terminologies to structure and annotate their data with ontology concepts for better search and retrieval. However, this annotation process cannot be easily automated and often requires expert curators. Plus, there is a lack of easy-to-use systems that facilitate the use of ontologies for annotation. This paper presents the Open Biomedical Annotator (OBA), an ontology-based Web service that annotates public datasets with biomedical ontology concepts based on their textual metadata ([www.bioontology.org](http://www.bioontology.org)). The biomedical community can use the annotator service to tag datasets automatically with ontology terms (from UMLS and NCBO BioPortal ontologies). Such annotations facilitate translational discoveries by integrating annotated data.<sup>[1]</sup>*

## Introduction & background

The wealth of publicly accessible biomedical data is beginning to enable cross-cutting integrative translational bioinformatics studies.<sup>[2]</sup> However, translational discoveries that could be made by mining biomedical resources are hampered because most online resources typically do not use standard terminologies and ontologies to annotate their elements (i.e., experimental data sets, diagnoses, diseases, samples, experimental conditions, clinical-trial descriptions, published papers). Currently, a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene, and the clinical trials that have studied drugs or diseases related to that gene. The knowledge needed to address such questions is available in public biomedical resources; the problem is finding that information. The research community agrees that ontologies are essential for data integration and translational discoveries to occur.<sup>[3]</sup>

However, the variety of biomedical data is very large and the data are often annotated with free text meta-data by the researcher who created the dataset. The problem is that these text metadata are unstructured and rarely described using standard ontology terms available in the domains. This situation creates a challenge of producing consistent terminology or ontology labels for each element in public biomedical resources. Such labels would enable the identification of all related elements at a given level of granularity. For example, the Gene Ontology (GO) is widely used to describe the molecular functions, cellular location, and biological processes of gene products and allows the integration of these descriptions across several databases. A similar query on the disease dimension is currently not possible because of the lack of a common terminology to describe disease involvements for gene products.

One mechanism of achieving ontology-based annotation is map existing textual metadata describing the resource element to ontology terms allowing formulation of refined or coarse search criteria.<sup>[4,5]</sup>

The annotation of biomedical data with biomedical ontology concepts is not a common practice for several reasons:<sup>[6]</sup>

- Annotation often needs to be done manually either by expert curators or directly by the authors of the data (e.g., when a new Medline entry is created, it is manually indexed with MeSH terms);
- The number of biomedical ontologies available for use is large and ontologies change often and frequently overlap. The ontologies are not in the same format and are not always accessible via application programming interfaces (APIs) that allow users to query them programmatically;
- Users do not always know the structure of an ontology's content or how to use the ontology to do the annotation themselves;
- Annotation is often a boring additional task without immediate reward for the user.

We have previously reported on a system for ontology-driven indexing of public resources for translational bioinformatics.<sup>[1]</sup> In this paper, we

present an annotator Web service that allows scientists to utilize available biomedical ontologies for annotating their datasets automatically. The *Open Biomedical Annotator* (OBA) Web service processes the raw textual metadata and tags them with relevant biomedical ontology concepts and returns the annotations to the users. Annotations are scored according to the context from which they have been generated. The OBA Web service utilizes ontologies for annotation of biomedical data in order to facilitate interoperation, search and translational discoveries.

## Methods

The OBA Web service's workflow is composed of two main steps (Figure 1). First, the user's free text is given as input to a *concept recognition tool* along with a dictionary. The dictionary (or lexicon) is a list of strings that identifies ontology concepts. The dictionary is constructed by accessing biomedical ontologies and pooling all concept names or other string forms, such as synonyms or labels that syntactically identify concepts.<sup>1</sup> The choice of the set of ontologies used to create the dictionary depends of the type of biomedical data the OBA Web service is used to annotate. For instance, if a user wants to annotate gene-expression datasets with disease names, then SNOMED-CT and the NCI Thesaurus could be used. The tool recognizes concepts by using string matching on the dictionary.<sup>2</sup> The output is a set of *direct annotations*.

This primary set of annotations serves as input for the *semantic expansion components*, which expand the annotations extracted from the first step using the structure and/or semantics of one or more ontologies. For example:

- An *is\_a transitive closure* component traverses an ontology parent-child hierarchy to create new annotations with parent concepts of the concepts involved in direct annotations. For instance, if data are directly annotated with the concept *melanoma* from NCI Thesaurus, this semantic expansion component can generate new annotations with concepts *skin tumor* and *neoplasms* because NCI Thesaurus provides the knowledge that *melanoma is\_a skin tumor* and *skin tumor*

*is\_a* neoplasms. The maximum level in the hierarchy to use is parameterizable.

- A *semantic distance* component uses a given notion of concept similarity (or semantic distance)<sup>[7,8]</sup> to obtain related concepts and create new annotations. For instance, if a text is directly annotated with the concept *melanoma* from Mesh, this semantic expansion component can generate new annotations with concepts *apudoma* and *neurilemmoma* because Mesh specifies these three concepts as siblings in the hierarchy. The maximum distance (threshold) and the type of semantic distance (path/graph based or information content based) to use are parameterizable.
- An *ontology-mapping* component creates new annotations based on existing mappings between different ontologies. For instance, if a text is directly annotated with the concept NCI/C0025202 (*melanoma* in NCI Thesaurus), this semantic expansion component can generate new annotations with concepts SNOMEDCT/C0025202 (*melanoma* in SNOMED-CT) and 38865/DOID:1909 (*melanoma* Human disease) because the UMLS and the NCBO BioPortal provides the mapping information. The type of mapping to use is parameterizable.

The OBA web service is designed in manner that allows multiple semantic expansion components to be plugged-in, selected, and parameterized by a user when requesting the service.<sup>3</sup> As the result of the second step, the direct annotations and several sets of *semantically expanded annotations* are extracted, scored and returned.

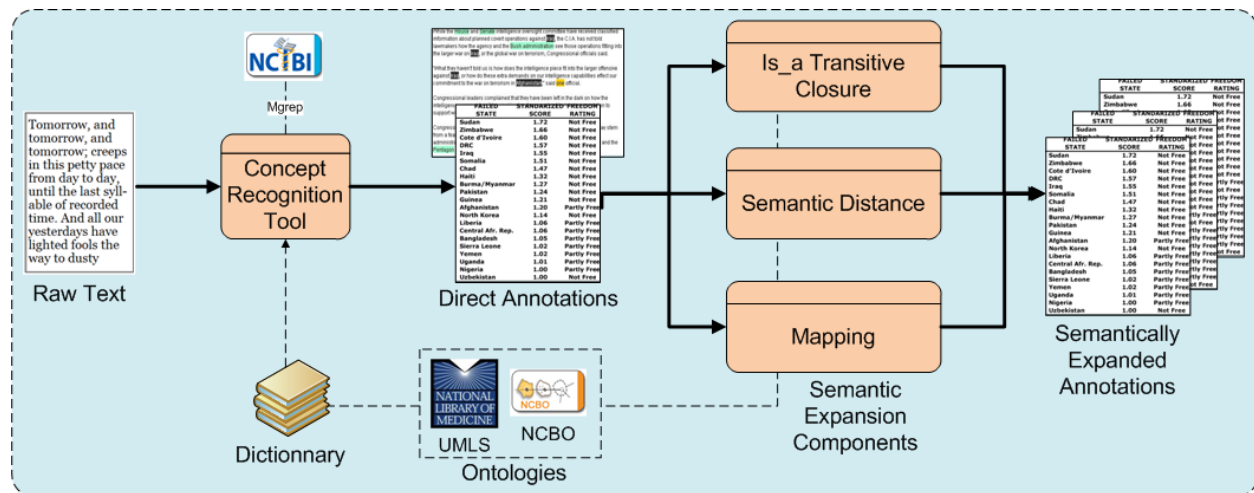
Annotations performed with the OBA service have implicit semantics that say *this dataset is about (or deals with) this concept*. Concepts are identified by UMLS Concept Unique Identifier (CUI)<sup>4</sup> or National Center for Biomedical Ontology (NCBO) Uniform Resource Identifier (URI). An annotation *context* asserts whether the annotation is direct or semantically expanded. In the latter case, the component used to produce the expanded annotation is described along with the concept from which the new annotation is derived. For example, the annotation [C0431097-ISA\_CLOSURE-C0025202] states that given text was annotated with the concept

<sup>1</sup> A *concept* is unique in an ontology (class). A *term* is a particular string form that identifies a concept. Usually, a concept has several terms (e.g., name, synonyms, label).

<sup>2</sup> Note that the concept recognizer does not execute any natural language processing techniques (stemming, spell-checking, morphological variants). However, this is not a major drawback as biomedical terminologies often contain syntactic variants for concepts as synonyms/terms.

<sup>3</sup> The service response time depends on the selected components as each consumes resources at a different level.

<sup>4</sup> NCBO is collaborating with National Library of Medicine to implement a license checking mechanism for UMLS licensed terminologies.



**Figure 1.** OBA web service workflow. First, direct annotations are created from raw text based on syntactic concept recognition according to a dictionary that use terms (concept names and synonyms) from both UMLS and NCBO ontologies. Second, different components expand the first set of annotations using ontologies semantics.

C0431097 ('malignant melanocytic lesion') using the is\_a relations of the concept C0025202 ('melanoma'). The scoring algorithm takes into account the context (direct, expanded, level, distance, etc.) and the frequency of annotations to evaluate which concepts annotates the best the given data. Annotations can be returned to the user in different formats (text, tab delimited, XML, or OWL). The description of the results returned by the OBA Web service is available.<sup>[9]</sup>

## Results

We have implemented the service using (at the moment of writing), all the (English) ontologies in UMLS (more than 94) and a subset of the NCBO BioPortal ontologies (more than 36).<sup>5</sup> Those ontologies offer a dictionary of 2,627,933 concepts and 5,177,973 terms. The service uses *Mgrep*,<sup>[10]</sup> a concept recognizer with a high degree of accuracy (>95%) in recognizing disease names<sup>[11]</sup> developed by the National Center for Integrative Biomedical Informatics (NCIBI) at the University of Michigan. *Mgrep* implements a novel radix-tree-based data-structure that enables fast and efficient matching of text against a set of dictionary terms. *Mgrep* was parameterized to match all the possible concepts.<sup>6</sup> We have conducted<sup>[12]</sup> a comparative evaluation of *Mgrep* with the gold standard in the biomedical

community MetaMap. For space reason, the results of this evaluation (in terms of precision, speed of execution, scalability and customizability) are described in another publication.<sup>[13]</sup> In the second step of the workflow, our biomedical annotator currently uses an is\_a transitive-closure component and leverages UMLS metathesaurus CUI-based mappings in order to expand the annotations created by *Mgrep*. The service is publicly available. It is deployed as a SOAP (Simple Object Access Protocol) and RESTful (REpresentational State Transfer) Web service.

We evaluated our biomedical annotator for the purpose of annotating a wide range of open biomedical resources.<sup>[1,14]</sup> For example, we annotated a set of 1,050,000 PubMed citations (title, abstract and other metadata), creating 174,840,027 annotations (18% direct, 82% expanded with is\_a relations). We obtained an average of 160 annotating concepts per citation and approximately 99% of our set was annotated (with at least 1 concept), demonstrating the service's utility.

We have used the annotator service internally to process several online datasets and have constructed an *Open Biomedical Resources* (OBR) index that allows a user to search for biomedical data annotated with ontology concepts.<sup>[1,14]</sup> The OBR index is directly queriable in the NCBO BioPortal ontology repository (<http://bioportal.bioontology.org/>). For example, searching for "melanoma" in BioPortal returns, among others, the concept DOID:1909 from the human disease ontology. A user can access the 13 ArrayExpress experiments, the 673 clinical trials, the

<sup>5</sup> Not all the NCBO BioPortal ontologies are fully usable through the REST web services API.

<sup>6</sup> If a text contains "cutaneous melanoma," two annotations are generated: one with 'melanoma' one with 'cutaneous melanoma' because the dictionary contains the two terms.

960 articles in PubMed, or the 10 GEO datasets related to this concept that OBA has annotated.

### Use cases supported

There are many use cases for the OBA Web service. The first use was to create the OBR index, which is described in a separate publication<sup>[1,14]</sup>. The service is currently being evaluated for use in several external workflows: (1) Researchers working on Trialbank ([www.trialbank.org](http://www.trialbank.org)) at the University of California, San Francisco, create annotations for HIV/AIDS clinical trials in order to provide a Web application for visualizing, and comparing the trials. They are evaluating the use of OBA to process the 'health condition', 'intervention' and 'outcomes' fields for trial records from clinicaltrials.gov. (2) Researchers at the University of Indiana are evaluating the utility of embedding the service in their research management system called Laboratree (<http://laboratree.org>); so that any textual annotation created in Laboratree would also have corresponding ontology term annotations. (3) Developers at Collabrx (<http://collabrx.com>) are embedding the service in their Rex platform for processing user generated content; and will evaluate the suitability of using medical dictionaries for processing such content. (4) Researchers at the Jackson Lab ([www.jax.org](http://www.jax.org)) are evaluating the utility of the OBA service in triaging articles for curation based on the ontology terms recognized in their title and abstract. Each of these groups get better interoperability of their data by using ontology annotations created with OBA. We are currently working on specific evaluations of OBA when used by each of these groups.

There are many other groups who are potential users of the annotator service. For example, Cancer nanoparticle research groups at Stanford and Washington Universities aim to use the annotator service for creating ontology-based annotations for the caNanoLab. And in the Ontology Development Information Extraction project, researchers at the University of Pittsburgh are developing a set of tools for extracting meaning and codifying medical documents that can enhance the annotator service (<http://www.bioontology.org/collaboration.html>).

### Related work

In the biomedical domain, automatic annotation or indexing of online resources is an important topic. A number of publicly available concept recognizers identify entities from ontologies or terminologies in text. For examples, see IndexFinder<sup>[15]</sup>, MetaMap<sup>[13]</sup>, CONANN<sup>[16]</sup>, and Mgrep<sup>[10,11]</sup>. MetaMap, which

identifies UMLS metathesaurus terms in text, is generally used as the gold standard for evaluating these tools. Our choice for Mgrep was made based on criteria for flexibility, speed and scalability as described before. Note that CONANN is very similar to OBA and is also available online. CONANN aims to identify the best possible matches, whereas Mgrep in the OBA identifies the greatest number of concepts. Plus, CONANN uses term frequency to filter results. However, CONANN is limited to UMLS and does not perform any semantic expansion step. Indeed, the knowledge contained in ontologies is rarely used to expand annotations, which gives to OBA a significant advantage. Note that the use of ontology semantics to enhance search is an active area of research.<sup>[5,18]</sup>

### Discussion and future work

The OBA Web service distinguishes itself from previous efforts for several reasons:

- It is a Web service that can be integrated in current programs and workflows;
- It uses public ontologies both to create annotations and to expand them;
- It has access to one of the largest available sets of publicly available biomedical ontologies from the UMLS metathesaurus and the NCBO BioPortal repository.

Current response times performed by the OBA Web service are ~20–25 seconds for 500 words. However, we are performing further technical improvements to OBA, such as: (1) keeping the dictionary loaded into memory between service calls (Mgrep constraint) and (2) loading the pre-computed hierarchy table into memory – in order to ensure fast response times for users

Future work will concentrate on three main issues that will determine the continued success of OBA Web service: (1) enhancement of the concept-recognition step by using natural languages processing techniques and eventually recognize 'relations,' (2) enhancement of the customizability of the service (parameters and ontologies used), and (3) enhancement of the semantic-expansion step by developing new components that use the knowledge in ontologies to relate concepts.

### Conclusion

Ontology-based annotation of biomedical data plays a crucial role for enabling data interoperability and the making of translational discoveries.<sup>[1]</sup> This situation is also true for e-science generally. The need to switch from the current Web to a semantic

Web with semantically rich content annotated using ontologies has been clearly identified.<sup>[19]</sup> Meeting this need requires services (usable by humans and software agents) that can be integrated into existing data curation and annotation workflows.

We have presented a service for ontology-based annotation of biomedical data. Our biomedical annotator has access to a large dictionary, which is composed of UMLS and NCBO ontologies. OBA is not limited to the syntactic recognition of terms, but also leverages the structure of the ontologies to expand annotations.

The service workflow is currently used in a project within NCBO to annotate a large number of public biomedical resources.<sup>[14]</sup> The OBA Web service is available to (and is already being used by) the community to evaluate its utility for creating ontology-based annotation of their data. The service can be customized to their specific needs (in terms of annotations parameters and biomedical ontologies used).

## Acknowledgments

This work is supported by NIH grant U54 HG004028 in support of the National Center for Biomedical Ontology, one of the National Centers for Biomedical Computing. We also acknowledge the assistance of Manhong Dai and Fan Meng (NCIBI).

## References

- Shah, N. H., Chiang, A. P., Butte, A. J., Chen, R., Musen, M. A.: Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. AMIA STB, San Francisco (Mar 2008)
- Butte, A., Chen, R.: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annual Symp., Washington DC (2006) 106
- Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 7(3) (Aug 2006) 256–274
- Shah, N.H., Rubin, D.L., Supekar, K.S., Musen, M.A.: Ontology-based Annotation and Query of Tissue Microarray Data. AMIA Annual Symp., Washington DC (Nov 2006) 709–713
- Moskovitch, R., Martins, S.B., Behiri, E., Weiss, A., Shahar, Y.: A Comparative Evaluation of Full-text, Concept-based, and Context-sensitive Search. AMIA 14(2) (Mar-Apr 2007) 164–174
- Shah, N.H.: Biomedical Data/Content Acquisition, Curation, Chapter in Encyclopedia of Database Systems, Springer-Verlag (2009)
- Lee, W.J., Raschid, L., Srinivasan, P., Shah, N.H., Rubin, D., Noy, N.: Using Annotations from Controlled Vocabularies to Find Meaningful Associations. 4<sup>th</sup> Int. Work. on Data Integration in the Life Sciences. Philadelphia, PA, (Jun 2007) 264–279
- Caviedesa, J.E., Cimino, J.J.: Towards the development of a conceptual distance metric for the UMLS. Biomedical Informatics 37(2) (Apr 2004) 77–85
- Jonquet, C., Musen, M.A., Shah, N.H.: Help will be provided for this task: Ontology-Based Annotator Web Service. Res. Report, BMIR-2008-1317, Stanford University (May 2008)
- Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B. Meng, F.: An Efficient Solution for Mapping Free Text to Ontology Terms. AMIA Summit on Translational Bioinformatics, San Francisco (March 2008)
- Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. BioLINK SIG: Linking Literature, Information and Knowledge for Biology, Vienna, Austria (Jul 2007) 55–58
- Bhatia, N., Shah, N.H., Rubin, D.L., Chiang, A.P., Musen, M.A.: Comparing Concept Recognizers for Ontology-Based Indexing: MGREP vs. MetaMap. AMIA STB, San Francisco (March 2009)
- Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annual Symp., Washington DC (Nov 2001) 17–21
- Jonquet, C., Musen, M.A., Shah, N.H.: A System for Ontology-Based Annotation of Biomedical Data. 5<sup>th</sup> Int. Work. on Data Integration in the Life Sciences. Evry, France, (Jun 2008) 144–152
- Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangarloo, H.: IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. AMIA Annual Symp., Washington DC (Nov 2003) 763–767
- Reeve, L.H., Han, H.: CONANN: An Online Biomedical Concept Annotator. 4<sup>th</sup> Int. Work. on Data Integration in the Life Sciences, Philadelphia, PA, (Jun 2007) 264–279
- Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: A Concept-based Search Engine for Structured Biomedical Text. AMIA 14(3) (2007) 253–263
- Handschuh, S., Staab, S., eds.: Annotation for the Semantic Web. Frontiers in Artificial Intelligence and Applications (96) (2003)



# References

1. Berners-Lee, T., Handler, J., Lassila, O., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American*. 34–43 (2001).
2. Gandon, F., Corby, O., Faron-Zucker, C.: Le web sémantique: Comment lier les données et les schémas sur le web ? Dunod (2012).
3. Gandon, F.: A Survey of the First 20 Years of Research on Semantic Web and Linked Data. *Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information*. (2018).
4. Gruber, T.R.: A translation approach to portable ontologies. *Knowledge Acquisition*. 5, 199–220 (1993).
5. Vrandeć, D.: Handbook on Ontologies. 293–313 (2009).
6. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*. 4, 14–28 (2006).
7. Handschuh, S., Staab, S. eds: Annotation for the Semantic Web. IOS Press (2003).
8. McCool, R.G.R., Miller, E.: Semantic Search. In: 12th International Conference on World Wide Web, WWW'03. pp. 700–709. ACM, Budapest, Hungary (2003).
9. Finin, T., Reddivari, P., Cost, R.S., Sachs, J.: Swoogle : A Search and Metadata Engine for the Semantic Web. In: ACM conference on Information and knowledge management. pp. 652–659 (2004).
10. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *Semantic Web and Information Systems*. 5, 1–22 (2009).
11. Ding, Y., Fensel, D.: Ontology Library Systems: The key to successful Ontology Re-use. In: 1st Semantic Web Working Symposium, SWWS'01. pp. 93–112. CEUR-WS.org, Stanford, CA, USA (2001).
12. Hartmann, J., Palma, R., Gómez-Pérez, A.: Ontology Repositories. *Handbook on Ontologies*. 551–571 (2009).
13. D'Aquin, M., Noy, N.F.: Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Web semantics*. 11, 96–111 (2012).
14. Gertjan van Heijst, Sabina Falasconi, Ameen Abu-Hanna, Guus Schreiber, M.S.: A case study in ontology library construction. *Artificial Intelligence in Medicine*. 7, 227–255 (1995).
15. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Barend Mons: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 3, (2016).
16. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*. 39, 541–545 (2011).
17. Sabou, M., Lopez, V., Motta, E.: Ontology Selection on the Real Semantic Web: How to Cover the Queens

- Birthday Dinner? In: Staab, S. and Svátek, V. (eds.) 15th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, EKAW'06. pp. 96–111. Springer, Podebrady, Czech Republic (2006).
18. Group, S.C.: Towards the Multilingual Semantic Web. (2014).
  19. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., Gàmez-Pérez, A.: A note on ontology localization. *Applied Ontology*. 5, 127–137 (2010).
  20. Euzenat, J., Shvaiko, P.: *Ontology matching*. Springer-Verlag, Berlin Heidelberg, DE (2007).
  21. D'Aquin, M., Castro, A.G., Lange, C., Viljanen, K. eds: 1st Workshop on Ontology Repositories and Editors for the Semantic Web, ORES'10. In: 1st Workshop on Ontology Repositories and Editors for the Semantic Web, ORES'10. CEUR-WS.org, Hersonissos, Greece (2010).
  22. Baclawski, K., Schneider, T.: The open ontology repository initiative: Requirements and research challenges. In: Tudorache, T., Correndo, G., Noy, N., Alani, H., and Greaves, M. (eds.) *Workshop on Collaborative Construction, Management and Linking of Structured Knowledge*, CK'09. p. 10. CEUR-WS.org, Washington, DC., USA (2009).
  23. Whetzel, P.L., Team, N.: NCBO Technology: Powering semantically aware applications. *Biomedical Semantics*. 4S1, 49 (2013).
  24. Till, M., Kutz, O., Codescu, M.: Ontohub: A semantic repository for heterogeneous ontologies. In: *Theory Day in Computer Science, DACS'14*. p. 2. , Bucharest, Romania (2014).
  25. Patel, C., Supekar, K., Lee, Y., Park, E.K.: OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification. In: 5th ACM International Workshop on Web Information and Data Management, WIDM'03. pp. 58–61. ACM, New Orleans, LA, USA (2003).
  26. Zhang, Y., Vasconcelos, W., Sleeman, D.: OntoSearch: An Ontology Search Engine. In: Bramer, M., Coenen, F., and Allen, T. (eds.) 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, AI'04. pp. 58–69. Springer, Cambridge, UK (2004).
  27. D'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E.: Watson: A Gateway for Next Generation Semantic Web Applications. In: *Poster & Demonstration Session at the 6th International Semantic Web Conference, ISWC'07*. p. 3. , Busan, Korea (2007).
  28. McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., Sansone, S.-A.: BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database*. 2016, (2016).
  29. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, T.O.B.I., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N.H., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*. 25, 1251–1255 (2007).
  30. Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., He, Y.: Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic acids research*. 45, D347–D352 (2016).
  31. Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H.: The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics*. 7, 97 (2006).
  32. Hoehndorf, R., Slater, L., Schofield, P.N., Gkoutos, G. V: Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*. 16, 1–9 (2015).
  33. Grosjean, J., Merabti, T., Griffon, N., Dahamna, B., Darmoni, S.: Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. In: 9th International Conference on Terminology and Artificial Intelligence, TIA'11. pp. 119–122. , Paris, France (2011).
  34. Vandenbussche, P.-Y., Atemez, G.A., Poveda-Villalon, M., Vatan, B.: Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. *Semantic Web*. 1, 1–5 (2014).
  35. Rueda, C., Bermudez, L., Fredericks, J.: The MMI Ontology Registry and Repository: A Portal for Marine Metadata Interoperability. In: *MTS/IEEE Biloxi - Marine Technology for Our Future: Global and Local Challenges*, OCEANS'09. p. 6. , Biloxi, MS, USA (2009).
  36. Naskar, D., Dutta, B.: Ontology Libraries : A Study from an Ontofier and an Ontologist Perspectives. In: 19th International Symposium on Electronic Theses and Dissertations, ETD'16. pp. 1–12. , Lille, France (2016).

37. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation. 18, W3C (2009).
38. Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J., Baker, T., Caracciolo, C., Retterath, A.: Publishing SKOS vocabularies with Skosmos. Manuscript submitted for review. (2015).
39. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*. 32, 267–270 (2004).
40. Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F.: BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. *Semantic Web*. 4, 277–284 (2013).
41. Annane, A.: Enhancing Ontology Matching with Background Knowledge Resources - Application to the Biomedical Domain, (2018).
42. Pouchard L., Huhns M., D.A.: Lessons learned in deploying a cloud-based knowledge platform for the ESIP Federation. American Geo-physical Union Fall Meeting, poster session. 22725 (2012).
43. Goble, C., Stevens, R.: State of the nation in data integration for bioinformatics. *Biomedical Informatics*. 41, 687–693 (2008).
44. Blake, J.A.: Bio-ontologies - fast and furious. *Nature Biotechnology*. 22, 773–774 (2004).
45. Bodenreider, O., Stevens, R.: Bio-ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*. 7, 256–274 (2006).
46. Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: A functional perspective. *Briefings in Bioinformatics*. 9, 75–90 (2008).
47. Névéal, A., Grosjean, J., Darmoni, S.J., Zweigenbaum, P.: Language Resources for French in the Biomedical Domain. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.) 9th International Conference on Language Resources and Evaluation, LREC'14. pp. 2146–2151. European Language Resources Association, Reykjavik, Iceland (2014).
48. Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungall, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P.: The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses. *Plant and Cell Physiology*. 54, e1 (2012).
49. Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G.F., Hancock, D., Morrison, N., Bruskiewich, R., McLaren, G.: Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. *AoB plants*. 2010, plq008 (2010).
50. Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., and Lewis, S.E.: The environment ontology: contextualising biological and biomedical entities. *Biomedical Semantics*. 4, 43 (2013).
51. Devare, M., Aubert, C., Laporte, M.-A., Valette, L., Arnaud, E., Buttigieg, P.L.: Data-driven Agricultural Research for Development - A Need for Data Harmonization Via Semantics. In: Jaiswal, P. and Hoehndorf, R. (eds.) 7th International Conference on Biomedical Ontologies, ICBO'16. p. 2. , Corvallis, Oregon, USA (2016).
52. Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougnot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönsch, G., Bunker, D.E., Cornelissen, J.H.C., Diaz, S., Enquist, B.J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., Pérez-Harguindeguy, N., Poorter, H., Schildhauer, M., Shipley, B., Violle, C., Weiher, E., Wirth, C., Wright, I.J., Klotz, S., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönsch, G., Bunker, D.E., Cornelissen, J.H.C., Díaz, S., Enquist, B.J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., Pérez-Harguindeguy, N., Poorter, H., Schildhauer, M., Shipley, B., Violle, C., Weiher, E., Wirth, C., Wright, I.J., Klotz, S.: Towards a thesaurus of plant characteristics: an ecological contribution. *Ecology*. 105, 298–309 (2016).
53. Griffiths, E., Brinkman, F., Buttigieg, P.L., Dooley, D., Hsiao, W., Hoehndorf, R.: FoodON: A Global Farm-to-Fork Food Ontology - The Development of a Universal Food Vocabulary. In: Jaiswal, P. and Hoehndorf, R. (eds.) 7th International Conference on Biomedical Ontologies, ICBO'16. p. 2. , Corvallis, Oregon, USA (2016).
54. Ibanescu, L., Dibie-Barthelemy, J., Dervaux, S., Guichard, E., Raad, J.: PO2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels. In: Emmanouel Garoufallou, Imma Subirats Coll, Armando Stellato, and Jane Greenberg (eds.) 10th International Conference on Metadata and Semantics Research. pp. 155–165. Springer, Göttingen, Germany (2016).

55. Musker, R., Lange, M., Hollander, A., Huber, P., Springer, N., Riggle, C., Quinn, J.F., Tomich, T.P.: Towards designing an ontology encompassing the environment-agriculture-food-diet-health knowledge spectrum for food system sustainability and resilience. In: Jaiswal, P. and Hoehndorf, R. (eds.) 7th International Conference on Biomedical Ontologies, ICBO'16. p. 5. , Corvallis, Oregon, USA (2016).
56. Hughes, L.M., Bao, J., Hu, Z.-L., Honavar<sup>†</sup>, V., Reecy, J.M.: Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *Animal Science*. 86, 1485–1491 (2014).
57. Lehmann, R.J., Reiche, R., Schiefera, G.: Future internet and the agri-food sector: State-of-the-art in literature and research. *Computers and Electronics in Agriculture*. 89, 158–174 (2012).
58. Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B.: Advancing ecological research with ontologies. *Trends in ecology & evolution*. 23, 159–68 (2008).
59. Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L., Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., Tuama, É.Ó., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wiczorek, J., Whitacre, J., Wooley, J.: Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. *PloS one*. 9, e89606 (2014).
60. Meng, X.: Special Issue – Agriculture Ontology. *Integrative Agriculture*. 11, 1 (2012).
61. Wang, Y., Wang, Y., Wang, J., Yuan, Y., Zhang, Z.: An ontology-based approach to integration of hilly citrus production knowledge. *Computers and Electronics in Agriculture*. 113, 24–43 (2015).
62. Dibie, J., Dervaux, S., Doriot, E., Ibanescu, L., Pénicaud, C.: MS2O – A Multi-scale and Multi-step Ontology for Transformation Processes: Application to Micro-Organisms. In: 22nd International Conference on Conceptual Structures, ICCS'16. pp. 163–176. Springer, Cham, Annecy, France (2016).
63. Lousteau-Cazalet, C., Barakat, A., Belaud, J.P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., Vialle, C.: A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*. 127, 351–367 (2016).
64. Jaiswal, P.: Plant Reverse Genetics: Methods and Protocols. Presented at the (2011).
65. Deans, A.R., Lewis, S.E., Huala, E., Anzaldo, S.S., Ashburner, M., Balhoff, J.P., Blackburn, D.C., Blake, J.A., Burleigh, J.G., Chanet, B., Cooper, L.D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T.A., Dettai, A., Diogo, R., Druzinsky, R.E., Dumontier, M., Franz, N.M., Friedrich, F., Gkoutos, G. V., Haendel, M., Harmon, L.J., Hayamizu, T.F., He, Y., Hines, H.M., Ibrahim, N., Jackson, L.M., Jaiswal, P., James-Zorn, C., Köhler, S., Lécointre, G., Lapp, H., Lawrence, C.J., Le Novère, N., Lundberg, J.G., Macklin, J., Mast, A.R., Midford, P.E., Mikó, I., Mungall, C.J., Oellrich, A., Osumi-Sutherland, D., Parkinson, H., Ramírez, M.J., Richter, S., Robinson, P.N., Ruttenberg, A., Schulz, K.S., Segerdell, E., Selmann, K.C., Sharkey, M.J., Smith, A.D., Smith, B., Specht, C.D., Squires, R.B., Thacker, R.W., Thessen, A., Fernandez-Triana, J., Vihinen, M., Vize, P.D., Vogt, L., Wall, C.E., Walls, R.L., Westerfeld, M., Wharton, R.A., Wirkner, C.S., Woolley, J.B., Yoder, M.J., Zorn, A.M., Mabee, P.: Finding Our Way through Phenotypes. *PLoS Biology*. 13, e1002033 (2015).
66. Matentzoglou, N., Malone, J., Mungall, C., Stevens, R.: MIRO: guidelines for minimum information for the reporting of an ontology. *Journal of biomedical semantics*. 9, 6 (2018).
67. Hartmann, J., Haase, P.: Ontology Metadata Vocabulary and Applications. 906–915 (2005).
68. Allocca, Carlo and d'Aquin, Mathieu and Motta, E.: DOOR - Towards a Formalization of Ontology Relations. In: International Conference on Knowledge Engineering and Ontology Development, KEOD'09. pp. 13–20. , Madera, Portugal (2009).
69. Vandenbussche, P.-Y., Vatan, B.: Metadata recommendations for linked open data vocabularies. (2012).
70. Keith, A., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets - on the design and usage of void, the “vocabulary of interlinked datasets.” In: Linked Data on the Web Workshop, LDOW'09. , Madrid, Spain (2009).
71. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. (1998).
72. Dutta, B., Nandini, D., Kishore, G.: MOD : Metadata for Ontology Description and Publication. In: International Conference on Dublin Core & Metadata Applications, DC'15. pp. 1–9. , Sao Paulo, Brazil (2015).
73. Park, J., Oh, S., Ahn, J.: Ontology selection ranking model for knowledge reuse. *Expert Systems with*

- Applications. 38, 5133–5144 (2011).
74. Malone, J., Stevens, R., Jupp, S., Hancocks, T., Parkinson, H., Brooksbank, C.: Ten Simple Rules for Selecting a Bio-ontology. *PLOS Computational Biology*. 12, 6 (2016).
  75. Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: Conference on Data Mining and Data Warehouses. *SiKDD'05*. p. 4. , Ljubljana, Slovenia (2005).
  76. Duque-Ramos, A., Fernández-Breis, J.T., Iniesta, M., Dumontier, M., Egaña Aranguren, M., Schulz, S., Aussenac-Gilles, N., Stevens, R.: Evaluation of the OQuaRE framework for ontology quality. *Expert Systems with Applications*. 40, 2696–2703 (2013).
  77. Gomez-Perez, A.: Some ideas and examples to evaluate ontologies. In: 11th Conference on Artificial Intelligence for Applications, CAIA'94. , San Antonio, TX, USA (1994).
  78. Gómez-Pérez, A.: From Knowledge Based Systems to Knowledge Sharing Technology: Evaluation and Assessment. , Technical Report KSL 94-73. Knowledge Systems Laboratory, Stanford University. Stanford, CA, USA (1994).
  79. Gruber, T.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In: International Workshop on Formal Ontology (1993).
  80. Cantador, I., Fernandez, M., Castells, P.: Improving Ontology Recommendation and Reuse in WebCORE by Collaborative Assessments. In: Noy, N., Alani, H., Stumme, G., Mika, P., Sure, Y., and Vrandečić, D. (eds.) Workshop on Social and Collaborative Construction of Structured Knowledge, CKC'07. CEUR-WS.org, Banff, Canada (2007).
  81. Buitelaar, P., Eigner, T., Declerck, T.: OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection. In: Demonstration Session at the 3rd International Semantic Web Conference, ISWC'04. pp. 3–6. , Hiroshima, Japan (2004).
  82. Alani, H., Brewster, C., Shadbolt, N.: Ranking Ontologies with AKTiveRank. 1–15 (2006).
  83. Tartir, S., Arpinar, I.B.B.B., Moore, M., Sheth, A.P., Aleman-meza, B.: OntoQA: Metric-Based Ontology Quality Analysis. In: IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources. p. 9. , Houston, TX, USA (2005).
  84. Martínez-Romero, M., Vázquez-Naya, J.M., Pereira, J., Pazos, A.: BiOSS: A system for biomedical ontology selection. *Computer methods and programs in biomedicine*. 114, 125–40 (2014).
  85. D'Aquin, M., Lewen, H.: Cupboard - A place to expose your ontologies to applications and the community. In: 6th European Semantic Web Conference, ESWC'09. pp. 913–918. Springer, Heraklion, Crete, Greece (2009).
  86. Butt, A.S., Haller, A., Xie, L.: RecOn: Ontology recommendation for structureless queries. *Applied Ontology*. 11, 301–324 (2016).
  87. Tan, H., Lambrix, P.: Selecting an Ontology for Biomedical Text Mining. In: Human Language Technology Conference, BioNLP Workshop. pp. 55–62. Association for Computational Linguistics, Boulder, CO, USA (2009).
  88. Alani, H., Noy, N.F.N.F., Shah, N., Shadbolt, N., Musen, M.A.: Searching Ontologies Based on Content: Experiments in the Biomedical Domain. In: 4th International Conference on Knowledge Capture, K-Cap'07. pp. 55–62. ACM, Whistler, BC, Canada (2007).
  89. Maiga, G.: A Flexible Biomedical Ontology Selection Tool. In: Kizza, J.M., Lynch, K., Nath, R., Aisbett, J., and Vir, P. (eds.) Strengthening the Role of ICT in Development. pp. 171–189. Fountain Publishers (2009).
  90. Martínez-romero, M., Vázquez-naya, J.M., Pereira, J., Pazos, A.: A Multi-criteria Approach for Automatic Ontology Recommendation Using Collective Knowledge. 89–103.
  91. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., DeLeenheer, P., and Pan, J.Z. (eds.) 8th Extended Semantic Web Conference, ESWC'11. pp. 245–259. Springer, Heraklion, Crete, Greece (2011).
  92. Montiel-Ponsoda, E., de Cea, G.A., Suarez-Figueroa, M.C., Palma, R., Gomez-Pérez, A., Peters, W.: LexOMV: an OMV extension to capture multilinguality. In: Buitelaar, P., Choi, K., Gangemi, A., Huang, C., and Oltramari, A. (eds.) Lexicon/Ontology Interface Workshop, OntoLex'07. p. 10. , Busan, South-Korea (2007).
  93. Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., Cea, G.A.: Enabling Language Resources to Expose

- Translations as Linked Data on the Web. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.) 9th International Conference on Language Resources and Evaluation, LREC'14. pp. 409–4013. European Language Resources Association, Reykjavik, Iceland (2014).
94. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero Bruno, Ayme, S.: Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Human mutation*. 33, 803–808 (2012).
  95. de Melo, G.: Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web*. 6, 8 (2015).
  96. Palma, R., Hartmann, J., Haase, P.: *Ontology Metadata Vocabulary for the Semantic Web*. (2008).
  97. Farrar, S., Langendoen, D.: Markup and the GOLD ontology. In: *Proceedings of workshop on digitizing and annotating text and field recordings* (pp. 845–862). 1–12 (2003).
  98. Buitelaar, P., Cimiano, P., Haase, P., Sintek, M.: Towards linguistically grounded ontologies. In: 6th. pp. 111–125. Springer, Heraklion, Crete, Greece (2009).
  99. Gracia, J., Montiel-Ponsoda, E., Gómez-Pérez, A.: Cross-lingual linking on the multilingual web of data (position statement). *CEUR Workshop Proceedings*. 936, (2012).
  100. Fu, B., Brennan, R., O'Sullivan, D.: Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. In: Buitelaar, P., Cimiano, P., and Montiel-Ponsoda, E. (eds.) 1st Workshop on the Multilingual Semantic Web. pp. 13–20. CEUR-WS.org, Raleigh, NC, USA (2010).
  101. Euzenat, J., Shvaiko, P.: *Ontology matching*, Second edition. Springer-Verlag, Berlin Heidelberg, DE (2013).
  102. David, J., Euzenat, J., Cássia Trojahn dos Santos, F.S.: The Alignment API 4.0. *Semantic Web*. 2, 3–10 (2011).
  103. Jupp, S., Liener, T., Sarntivijai, S., Vrousseau, O., Burdett, T., Parkinson, H.: OxO – A Gravy of Ontology Mapping Extracts. In: Horridge, M., Lord, P., and Warrender, J.D. (eds.) 8th International Conference on Biomedical Ontology, ICBO'17. p. 2. , Newcastle, UK (2017).
  104. Bellahsene, Z., abnd Duyhoa Ngo, V.E., Todorov, K.: YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation. In: 14th European Semantic Web Conference, ESWC'17. pp. 137–142. Springer, Portoroz, Slovenia (2017).
  105. Falconer, S.M., Storey, M.-A.: A cognitive support framework for ontology mapping. In: *The Semantic Web*. pp. 114–127. Springer (2007).
  106. Lanzenberger, M., Sampson, J.: Alviz-a tool for visual ontology alignment. In: *Information Visualization, 2006. IV 2006. Tenth International Conference on*. pp. 430–440 (2006).
  107. Ivanova, V., Bach, B., Pietriga, E., Lambrix, P.: Alignment Cubes: Towards Interactive Visual Exploration and Evaluation of Multiple Ontology Alignments. In: *International Semantic Web Conference*. pp. 400–417 (2017).
  108. Noy, N.F., Griffith, N.B., Musen, M.A.: Collecting Community-Based Mappings in an Ontology Repository. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T.W., and Thirunarayan, K. (eds.) 7th International Semantic Web Conference, ISWC'08. pp. 368–371. Springer, Karlsruhe, Germany (2008).
  109. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating Mappings For Ontologies in Biomedicine: Simple Methods Work. In: *American Medical Informatics Association Annual Symposium, AMIA'09*. pp. 198–202. , Washington DC, USA (2009).
  110. Faria, D., Jiménez-Ruiz, E., Pesquita, C., Santos, E., Couto, F.M.: Towards Annotating Potential Incoherences in BioPortal Mappings. In: 13th International Semantic Web Conference, ISWC'13. pp. 17–32. Springer, Riva del Garda, Italy (2014).
  111. Pathak, J., Chute, C.G.: Debugging Mappings between Biomedical Ontologies: Preliminary Results from the NCBO BioPortal Mapping Repository . In: Smith, B. (ed.) *International Conference on Biomedical Ontology*. pp. 95–98. , Buffalo, NY, USA (2009).
  112. Kamdar, M.R., Tudorache, T., Musen, M.A.: A Systematic Analysis of Term Reuse and Term Overlap across Biomedical Ontologies. *Semantic web*. 8, 853–871 (2017).
  113. Cheatham, M., Hitzler, P.: String Similarity Metrics for Ontology Alignment. In: 12th International Semantic Web Conference, ISWC'13. pp. 294–309. Springer, Sydney, Australia (2013).

114. Ngo, D., Bellahsene, Z.: YAM++ : A Multi-strategy Based Approach for Ontology Matching Task. In: ten Teije, A., Volker, J., Handschuh, S., Stuckenschmidt, H., D'Acquin, M., Nikolov, A., Aussenac-Gilles, N., and Hernandez, N. (eds.) 18th International Conference on Knowledge Engineering and Knowledge Management, EKAW'12. pp. 421–425. Springer, Galway City, Ireland (2012).
115. Locoro, A., David, J., Euzenat, J.: Context-Based Matching: Design of a Flexible Framework and Experiment. *Journal on Data Semantics*. 3, 25–46 (2014).
116. Faria, D., Pesquita, C., Santos, E., Cruz, I.F., Couto, F.M.: Automatic Background Knowledge Selection for Matching Biomedical Ontologies. *PLoS ONE*. 9, e111226 (2014).
117. Hartung, M., Gross, A., Kirsten, T., Rahm, E.: Effective Composition of Mappings for Matching Biomedical Ontologies. In: 9th European Semantic Web Conference, ESWC'12. pp. 176–190. Springer, Heraklion, Crete, Greece (2015).
118. Quix, C., Roy, P., Kensch, D.: Automatic selection of background knowledge for ontology matching. In: International Workshop on Semantic Web Information Management , SWIM '11. pp. 1–7. ACM Press, New York, NY, USA (2011).
119. Groß, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping Composition for Matching Large Life Science Ontologies. In: 2nd International Conference on Biomedical Ontology, ICBO, Buffalo, NY, USA. pp. 109–116 (2011).
120. Sabou, M., d'Aquin, M., Motta, E.: Exploring the Semantic Web as Background Knowledge for Ontology Matching. Presented at the (2008).
121. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*. 25, 158–176 (2013).
122. Faria, D., Pesquita, C., Balasubramani, B.S., Martins, C., Cardoso, J., Curado, H., Couto, F.M., Cruz, I.F.: OAEI 2016 Results of AML. In: 11th International Workshop on Ontology Matching, OM, Kobe, Japan. pp. 138–145 (2016).
123. Jiménez-Ruiz, E., Grau, B.C., Cross, V.: LogMap family participation in the OAEI 2016. In: 11th International Workshop on Ontology Matching, OM, Kobe, Japan. pp. 185–189 (2016).
124. Mascardi, V., Locoro, A., Rosso, P.: Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation. *IEEE Transactions on Knowledge and Data Engineering*. 22, 609–623 (2010).
125. Leo, B.: Random forests. *Machine learning*. 45, 5–32 (2001).
126. Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., Witten, I.H.: The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 11, 10–18 (2009).
127. Ngo, D., Bellahsene, Z.: YAM++ results for OAEI 2013. In: 8th International Workshop on Ontology Matching, OAEI'13. pp. 211–218. , Sydney, Australia (2013).
128. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*. 6, 239–251 (2005).
129. Hollink, L., Schreiber, G., Wilemaker, J., Wielinga, B.: Semantic Annotation of Image Collections. In: Handschuh, S., Koivunen, M., Dieng-Kuntz, R., and Staab, S. (eds.) Knowledge Markup and Semantic Annotation Workshop. , Sanibel, FL, USA (2003).
130. Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S.: Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*. 9, 509–515 (2008).
131. Butte, A.J., Chen, R.: Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. In: American Medical Informatics Association Annual Symposium, AMIA'06. pp. 106–110. , Washington DC, USA (2006).
132. Griffon, N., Soualmia, L.F., Névél, A., Massari, P., Thirion, B., Dahamna, B., Darmoni, S.J.: Evaluation of Multi-Terminology Super-Concepts for Information Retrieval. In: Moen, A., Andersen, S.K., Aarts, J., and Hurlen, P. (eds.) 23rd International Conference of the European Federation for Medical Informatics, MIE'11. pp. 492–496. IOS Press, Oslo, Norway (2011).
133. Dieng-kuntz, R., Antipolis, I.S., Barbry, P., Antipolis, S., Khelif, K.: An ontology-based approach to support text mining and information retrieval in the biological domain. *Universal Computer Science, Special Issue on Ontologies and their Applications*. 13, 1881–1907 (2007).
134. Zou, Q., Chu, W.W., Morioka, C., Leazer, G.H., Kangaroo, H.: IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. In: American Medical Informatics Association Annual

- Symposium, AMIA'03. pp. 763–767. , Washington DC, USA (2003).
135. Baumgartner, W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.A., Jr, W.A.B., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.A.: Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*. 23, 41–48 (2007).
  136. Goble, C., Stevens, R., Hull, D., Wolstencroft, K., Lopez, R.: Data curation + process curation=data integration + science. *Briefings in Bioinformatics*. 9, 506–517 (2008).
  137. Dowell, K.G., McAndrews-Hill, M.S., Hill, D.P., Drabkin, H.J., Blake, J.A.: Integrating text mining into the MGI biocuration workflow. *Database*. 2009, 1–11 (2009).
  138. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, AMIA'01. pp. 17–21. , Washington DC, USA (2001).
  139. Altman, R.B., Bergman, C.M., Blake, J.A., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen, L.J., Krallinger, M., Mons, B., O'Donoghue, S.I.S.I., Peitsch, M.C., Rebholz-Schuhmann, D., Shatkay, H., Valencia, A.: Text mining for biology - the way forward: opinions from leading scientists. *Genome Biology*. 9, (2008).
  140. Hancock, D., Morrison, N., Velarde, G., Field, D.: Terminizer -- Assisting Mark-Up of Text Using Ontological Terms. In: 3rd International Biocuration Conference. p. 22. , Berlin, Germany (2009).
  141. Pavlopoulos, G.A., Pafilis, E., Kuhn, M., Hooper, S.D., Schneider, R.: OnTheFly: A Tool for automated document-based text annotation, data linking and network generation. *Bioinformatics*. 25, 977–978 (2009).
  142. Pafilis, E., O'Donoghue, S.I.S.I., Jensen, L.J., Horn, H., Kuhn, M., Brown, N.P., Schneider, R.: Reflect: augmented browsing for the life scientist. *Nature biotechnology*. 27, 508–510 (2009).
  143. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Jimeno, A.: Text processing through Web services: Calling Whatizit. *Bioinformatics*. 24, 296–298 (2008).
  144. Hersh, W.R., Greenes, R.A.: SAPHIRE - an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Computers and Biomedical Research*. 23, 410–425 (1990).
  145. Reeve, L.H., Han, H.: CONANN: An Online Biomedical Concept Annotator. In: Cohen-Boulakia, S. and Tannen, V. (eds.) 4th International Workshop Data Integration in the Life Sciences, DILS'07. pp. 264–279. Springer-Verlag, Philadelphia, PA, USA (2007).
  146. Jovanović, J., Bagheri, E.: Semantic annotation in biomedicine: the current landscape. *Journal of Biomedical Semantics*. 8, 44 (2017).
  147. Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., Jacobson, R.S.: NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*. 17, 32 (2016).
  148. Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., Verspoor, K.: Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*. 15, (2014).
  149. Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*. 15, 327–340 (2014).
  150. Sakji, S., Gicquel, Q., Pereira, S., Kergoulay, I., Proux, D., SJ, D., Metzger, M.H.: Evaluation of a French Medical Multi-Terminology Indexer for the Manual Annotation of Natural Language Medical Reports of Healthcare-Associated Infections. In: et al., C.S. (ed.) 13th World Congress on Medical Informatics, MedInfo'10. pp. 252–256. IOS Press, Cape Town, South Africa (2010).
  151. Soualmia, L.F., Dahamna, B.: SIBM at CLEF eHealth Evaluation Lab 2016 : Extracting Concepts in French Medical Texts with ECMT and CIMIND. (2016).
  152. Aussenac-Gilles, N., Despres, S., Szulman, S.: The TERMINAE Method and Platform for Ontology Engineering from Texts. In: Buitelaar, P. and Cimiano, P. (eds.) Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. pp. 199–223. Ios Press (2008).
  153. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F.: An Efficient Solution for Mapping Free Text to Ontology Terms. In: American Medical Informatics Association Symposium on Translational Bioinformatics, AMIA-TBI'08. , San Francisco, CA, USA (2008).
  154. Twigger, S., Geiger, J., Smith, J., Simon N. Twigger Joey Geiger, J.S.: Using the NCBO Web Services for

- Concept Recognition and Ontology Annotation of Expression Datasets. In: Marshall, M.S., Burger, A., Romano, P., Paschke, A., and Splendiani, A. (eds.) Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS'09. , Amsterdam, The Netherlands (2009).
155. Sarkar, I.N.: Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts. In: American Medical Informatics Association Annual Symposium, AMIA'10. pp. 717–721. , Washington DC., USA (2010).
  156. Groza, T., Oellrich, A., Collier, N.: Using silver and semi-gold standard corpora to compare open named entity recognisers. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. pp. 481–485 (2013).
  157. Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK: Linking Literature, Information and Knowledge for Biology, SIG, ISMB'08. pp. 55–58. , Vienna, Austria (2007).
  158. McCray, A.T.: An Upper-Level Ontology for the Biomedical Domain. *Comparative and Functional Genomics*. 4, 80–84 (2003).
  159. Thomas, D.G., Pappu, R. V, Baker, N.A.: NanoParticle Ontology for Cancer Nanotechnology Research. *Biomedical Informatics*. 1–16 (2010).
  160. Tirrell, R., Evani, U., Berman, A.E., Mooney, S.D., Musen, M.A., Shah, N.H.: An Ontology-Neutral Framework for Enrichment Analysis. In: American Medical Informatics Association Annual Symposium, AMIA'10. pp. 797–801. , Washington DC, USA (2010).
  161. Odgers, D.J., Dumontier, M.: Mining Electronic Health Records using Linked Data. In: AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science. pp. 217–21. AMIA, San Francisco, CA, USA (2015).
  162. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*. 6, 26094 (2016).
  163. Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., Shah, N.H.: Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug Safety*. 37, 777–790 (2014).
  164. Rebholz-Schuhmann, D., Oellrich, A., Hoehndorf, R.: Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*. 13, 829–839 (2012).
  165. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*. 84, 216 (2001).
  166. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value Method. *Digital Libraries*. 3, 115–130 (2000).
  167. Goeuriot, L., Kelly, L., Suominen, H., Névél, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab Overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer (2017).
  168. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névél, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. In: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G., San Juan, E., Capellato, L., and Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. pp. 429–443. Springer International Publishing, Cham (2015).
  169. Kelly, L., Goeuriot, L., Suominen, H., Névél, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., and Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. pp. 255–266. Springer, Cham (2016).
  170. Névél, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In: 4th Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing, BioTxtM'14. pp. 24–30. , Reykjavik, Iceland (2014).
  171. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., others: Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*. 35, 44 (2008).
  172. Rothman, B., Leonard, J.C., Vigoda, M.M.: Future of electronic health records: implications for decision

- support. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*. 79, 757–768 (2012).
173. Liu, H., Bielinski, S.J., Sohn, S., Murphy, S., Wagholikar, K., Jonnalagadda, S., Ravikumar, K., Wu, S., Kullo, I., Chute, C.: An Information Extraction Framework for Cohort Identification Using Electronic Health Records. In: *American Medical Informatics Association Summits on Translational Science*. pp. 149–153 (2013).
  174. Herasevich, V., Tsapenko, M., Kojicic, M., Ahmed, A., Kashyap, R., Venkata, C., Shahjehan, K., Thakur, S.J., Pickering, B.W., Zhang, J., others: Limiting ventilator-induced lung injury through individual electronic medical record surveillance. *Critical care medicine*. 39, 34–39 (2011).
  175. Chapman, W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. *Biomedical Informatics*. 34, 301–310 (2001).
  176. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: ConText: an algorithm for determining negation, experienter, and temporal status from clinical reports. *Journal of biomedical informatics*. 42, 839–51 (2009).
  177. Lossio-Ventura, J.A.: *Towards the French Biomedical Ontology Enrichment*, (2015).
  178. Kageura, K., Umino, B.: Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*. 3, 259–289 (1996).
  179. Zhang, Z., Iria, J., Brewster, C., Ciravegna, F.: A Comparative Evaluation of Term Recognition Algorithms. In: *6th International Conference on Language Resources and Evaluation, LREC'08*. pp. 2108–2113. , Marrakech, Morocco (2008).
  180. Hliaoutakis, A., Zervanou, K., Petrakis, E.G.M.: The AMTE approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering*. 68, 380–392 (2009).
  181. Spasic, I., Greenwood, M., Preece, A., Francis, N., Elwyn, G.: FlexiTerm: a flexible term recognition method. *Biomedical Semantics*. 4, (2013).
  182. Golik, W., Bossy, R., Ratkovic, Z., Nédellec, C.: Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science*. 70, 157–172 (2013).
  183. Agirre, E., Soroa, A.: Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In: *4th International Workshop on Semantic Evaluations*. pp. 7–12 (2007).
  184. Manandhar, S., Klapaftis, I.P., Dligach, D., Pradhan, S.S.: SemEval-2010 task 14: Word sense induction & disambiguation. In: *5th international workshop on semantic evaluation*. pp. 63–68 (2010).
  185. Dehkordi, M.Y., Boostani, R., Tahmasebi, M.: A novel hybrid structure for clustering. In: *Advances in Computer Science and Engineering*. pp. 888–891. Springer (2008).
  186. Breton, D., Bringay, S., Marques, F., Poncelet, P., Roche, M.: Epimining: Using Web News for Influenza Surveillance. In: *3rd Workshop on Data Mining for Healthcare Management, DMHM'12*. , Kuala Lumpur, Malaysia (2012).
  187. Pletneva, N., Vargas, A., Boyer, C.: *How Do General Public Search Online Health Information?* , Geneva, Switzerland (2011).
  188. McCray, A.T., Loane, R.F., Browne, A.C., Bangalore, A.K.: Terminology issues in user access to Web-based medical information. In: *American Medical Informatics Association Annual Symposium, AMIA'99*. p. 107 (1999).
  189. Plovnick, R.M., Zeng, Q.T.: Reformulation of consumer health queries with professional terminology: A pilot study. *Journal of Medical Internet Research*. 6, 2008 (2004).
  190. MacLean, D.L., Heer, J.: Identifying medical terms in patient-authored text: A crowdsourcing-based approach. *Journal of the American Medical Informatics Association*. 20, 1120–1127 (2013).
  191. Doing-Harris, K.M., Zeng-Treitler, Q.: Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of medical Internet research*. 13, e37 (2011).
  192. Opitz, T., Azé, J., Bringay, S., Joutard, C., Lavergne, C., Mollevi, C.: Breast cancer and quality of life: medical information extraction from health forums. In: *Medical Informatics Europe*. pp. 1070–1074 (2014).
  193. Nzali, M.D.T., Bringay, S., Azria, D., Lavergne, C., Mollevi, C.: Acquisition du vocabulaire patient/médecin présent dans les forums de santé dédiés au cancer du sein. *Epidemiology and Public Health / Revue d'Epidémiologie et de Santé Publique*. 63, S66–S67 (2015).

194. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2, 49–79 (2004).
195. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. *Journal of Web Semantics*. 9, 365–401 (2011).
196. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Biomedical Informatics*. 41, 706–716 (2008).
197. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novere, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*. 30, 1338–1339 (2014).
198. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: Semantic interoperability for drug discovery, (2012).
199. Tissaoui, A., Aussenac-Gilles, N., Hernandez, N., Laublet, P.: EvOnto - Joint Evolution of Ontologies and Semantic Annotations. In: 3rd International Conference on Knowledge Engineering and Ontology Development, KEOD'11. , Paris, France (2011).
200. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. Presented at the (2013).
201. Momtchev, V., Peychev, D., Primov, T., Georgiev, G.: Expanding the Pathway and Interaction Knowledge in Linked Life Data. In: International Semantic Web Challenge (2009).
202. Jupp, S., Klein, J., Schanstra, J., Stevens, R.: Developing a kidney and urinary pathway knowledge base. *Journal of biomedical semantics*. 2 Suppl 2, S7 (2011).
203. Sneiderman, C.A., Demner-Fushman, D., Fiszman, M., Ide, N.C., Rindflesch, T.C.: Knowledge-based Methods to Help Clinicians Find Answers in Medline. *American Medical Informatics Association*. 14, 772–780 (2007).
204. Coulet, A., Garten, Y., Dumontier, M., Altman, R.B., Musen, M.A., Shah, N.H.: Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *J. Biomedical Semantics*. 2, S10 (2011).
205. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C recommendation. 30, (2013).
206. Surroca, G.: ViewointS : vers une émergence de connaissances collectives par élicitation de point de vue, (2017).
207. Mika, P.: Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*. 5, 5–15 (2007).
208. Karapiperis, S., Apostolou, D.: Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*. 1, 199–216 (2006).
209. Gruber, T.: Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*. 6, 4–13 (2008).
210. Freddo, A.R., Tacla, C.A.: Integrating social web with semantic web : ontology learning and ontology evolution from folksonomies. *KEOD 2009 proceedings*. 247–253 (2009).
211. Limpens, F., Gandon, F., Buffa, M.: Bridging ontologies and folksonomies to leverage knowledge sharing on the social Web: A brief survey. In: 2008 23rd IEEE/ACM International Conference on Automated Software Engineering - Workshops. pp. 13–18. IEEE (2008).
212. Gandon, F., Buffa, M., Cabrio, E., Corby, O., Faron-Zucker, C., Giboin, A., Le Thanh, N., Mirbel, I., Sander, P., Tettamanzi, A., Villata, S.: Challenges in Bridging Social Semantics and Formal Semantics on the Web. In: 15th International Conference on Enterprise Information Systems, ICEIS'13. pp. 3–15. Springer, Angers, France (2013).
213. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*. 30, 740–742 (2013).
214. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., others: Gene Ontology: tool for the unification of biology. *Nature genetics*. 25, 25–29 (2000).
215. Scharffe, F., Atemezeng, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., Hamdi, F., Bihanic, L., Képéklian, G., Cotton, F., Euzenat, J., Fan, Z., Vandenbussche, P.-Y., Vattant, B.: Enabling linked data publication with

the Datalift platform. Presented at the July 23 (2012).

216. Achichi, M., Lisena, P., Todorov, K., Troncy, R., Delahousse, J.: DOREMUS: A Graph of Linked Musical Works. Presented at the October 8 (2018).
217. Sachit Rajbhandari, J.K.: The AGROVOC Concept Scheme – A Walkthrough. *Integrative Agriculture*. 11, 694–699 (2012).
218. Dumontier, M., Gray, A.J.G., Marshall, M.S., Alexiev, V., Ansell, P., Bader, G., Baran, J., Bolleman, J.T., Callahan, A., Cruz-Toledo, J., Gaudet, P., Gombocz, E.A., Gonzalez-Beltran, A.N., Groth, P., Haendel, M., Ito, M., Jupp, S., Juty, N., Katayama, T., Kobayashi, N., Krishnaswami, K., Laibe, C., Novère, N. Le, Lin, S., Malone, J., Miller, M., Mungall, C.J., Rietveld, L., Wimalaratne, S.M., Yamaguchi, A.: The health care and life sciences community profile for dataset descriptions. *PeerJ*. 16, (2016).
219. Musen, M.A., Bean, C.A., Cheung, K.-H., Dumontier, M., Durante, K.A., Gevaert, O., Gonzalez-Beltran, A., Khatri, P., Kleinstein, S.H., O'Connor, M.J., Pouliot, Y., Rocca-Serra, P., Sansone, S.-A., Wiser, J.A., the CEDAR team: The center for expanded data annotation and retrieval. *American Medical Informatics Association*. 22, 1148–1152 (2015).
220. Wilkinson, M.D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L.O., Dumontier, M.: A design framework and exemplar metrics for FAIRness. *Scientific Data*. 5, 180118 (2018).
221. Baker, T., Suominen, O.: Global Agricultural Concept Scheme (GACS): A multilingual thesaurus hub for Linked Data. (2014).
222. Corby, O., Dieng-Kuntz, R., Faron-zucker, C., Gandon, F., Nria, I., Faron-zucker, C., Cnrs, U.M.R.: Searching the Semantic Web: Approximate Query Processing Based on Ontologies. *IEEE Intelligent Systems*. 21, 20–27 (2006).
223. Pesce, V., Pesce, V., Maru, A., Keizer, J.: The CIARD RING, an Infrastructure for Interoperability of Agricultural Research Information Services. *Agricultural Information Worldwide*. 4, 48–53 (2011).
224. Cooper, L., Meier, A., Laporte, M.-A., Elser, J.L., Mungall, C., Sinn, B.T., Cavaliere, D., Carbon, S., Dunn, N.A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J.H., Stevenson, D.W., Arnaud, E., Jaiswal, P.: The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic acids research*. 46, D1168–D1180 (2018).
225. Dastgheib, S., Whetzel, T., Zaveri, A., Afrasiabi, C., Assis, P., Avillach, P., Jagodnik, K.M., Korodi, G., Pilarczyk, M., Pons, J. de, Schürer, S.C., Terryn, R., Verborgh, R., Wu, C., Dumontier, M.: The SmartAPI Ecosystem for Making Web APIs FAIR. In: N. Nikitina, D. Song, A. Fokoue, and P. Haase (eds.) 16th International Semantic Web Conference, ISWC'17, Posters & Demonstrations and Industry Tracks. CEUR, Vienna, Austria (2017).
226. Zeng, M.L.: Knowledge Organization Systems (KOS). *Knowledge organization*. 35, 160–182 (2008).
227. McGuinness, D.L.: Spinning the semantic web: bringing the World Wide Web to its full potential. Presented at the (2003).
228. Deléger, L., Grouin, C.: Detecting negation of medical problems in French clinical notes. In: 2nd ACM SIGHIT International Health Informatics Symposium. pp. 697–702 (2012).

