# Profile-based Datas and Recommendation for RDF Data Linking

Mohamed Ben Ellefi

Collège Doctoral du Languedoc-Roussillon

# THÈSE

Pour obtenir le grade de

**Docteur**

Délivrée par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S***
Et de l'unité de recherche **LIRMM**

Spécialité: **Informatique**

Présentée par **Mohamed BEN ELLEFI**
benellefi@lirmm.fr

# Profile-based Dataset Recommendation for RDF Data Linking

Soutenue le 01/12/2016 devant le jury composé de

| | | | |
|---|---|---|---|
| Mathieu d'Aquin | Senior Researcher | Open University, UK | Rapporteur |
| Nathalie Pernelle | Maître de Conférences (HDR) | Université Paris Orsay | Rapporteur |
| Mohand Boughanem | Professeur | Université Paul Sabatier | Examinateur |
| Nabil Layaïda | Directeur de Recherche | INRIA Grenoble | Examinateur |
| Zohra Bellahsene | Professeur | Université Montpellier | Directrice |
| Konstantin Todorov | Maître de Conférences | Université Montpellier | Co-Encadrant |

Université Montpellier

*À ma chère Maman Saloua*
*À mes tantes et oncles*
*À mon frère Mostfa et à Takwa, la petite nouvelle de la famille*
*À tous mes amis notamment, Nader (mon bro) et Bilel (mon pote)*
*et à tous ceux que je ne nomme pas, mais qui se reconnaîtront.*

# Remerciements

Je tiens à exprimer ma gratitude à Monsieur **Mathieu d'Aquin**, Senior Researcher à l'Open University du Royaume Uni et Madame **Nathalie Pernelle**, Maître de Conférences à l'Université de Paris Orsay pour avoir accepté de rapporter sur ma thèse, ainsi qu'à Monsieur **Mohand Boughanem**, Professeur à l'Université Paul Sabatier et Monsieur **Nabil Layaïda**, Directeur de Recherche à l'INRIA de Grenoble pour avoir accepté de faire partie du jury.

Je tiens à exprimer mon respect, ma gratitude et mes plus chaleureux remerciements à Madame **Zohra Bellahsene**, Professeur à l'Université de Montpellier pour avoir été ma directrice de thèse. Je la remercie vivement pour ses conseils pertinents, sa gentillesse et sa présence à mes côtés. Je n'oublierai jamais son soutien et sa disponibilité dans les moments de doute. Son très bon encadrement a été l'une des clés de réussite de cette thèse.

Je voudrai aussi remercier Monsieur **Konstantin Todorov**,Maître de Conférences à l'Université de Montpellier pour avoir co-encadré cette thèse. J'ai notamment apprécié la grande confiance qu'il a placée en moi et en mes recherches, bien que je fusse son premier thésard. Ayant partagé avec lui mon bureau pendant la quasi-totalité de ma thèse, j'ai ainsi profité de ses conseils, de son bon jugement mais également de sa bonne humeur et de sa générosité. J'ai été fier d'avoir mené à bien la première thèse

Je remercie également mon laboratoire d'accueil le **LIRMM** ainsi que le projet **DATALYSE**, qui a financé mes travaux de recherche dans le cadre de cette thèse.

Je tiens à remercier tous ceux qui ont participé de près ou de loin à la réalisation de ce travail, notamment mes collègues du LIRMM et tous ceux que j'ai pu rencontrer au gré des réunions du projet ou des conférences auxquelles j'ai participé.

# Abstract

With the emergence of the Web of Data, most notably Linked Open Data (LOD), an abundance of data has become available on the web. However, LOD datasets and their inherent subgraphs vary heavily with respect to their size, topic and domain coverage, the schemas and their data dynamicity (respectively schemas and metadata) over the time. To this extent, identifying suitable datasets, which meet specific criteria, has become an increasingly important, yet challenging task to support issues such as entity retrieval or semantic search and data linking. Particularly with respect to the interlinking issue, the current topology of the LOD cloud underlines the need for practical and efficient means to recommend suitable datasets: currently, only well-known reference graphs such as DBpedia (the most obvious target), YAGO or Freebase show a high amount of in-links, while there exists a long tail of potentially suitable yet under-recognized datasets. This problem is due to the semantic web tradition in dealing with "finding candidate datasets to link to", where data publishers are used to identify target datasets for interlinking.

While an understanding of the nature of the content of specific datasets is a crucial prerequisite for the mentioned issues, we adopt in this dissertation the notion of "dataset profile" — a set of features that describe a dataset and allow the comparison of different datasets with regard to their represented characteristics. Our first research direction was to implement a collaborative filtering-like dataset recommendation approach, which exploits both existing dataset topic profiles, as well as traditional dataset connectivity measures, in order to link LOD datasets into a global dataset-topic-graph. This approach relies on the LOD graph in order to learn the connectivity behaviour between LOD datasets. However, experiments have shown that the current topology of the LOD cloud group is far from being complete to be considered as a ground truth and consequently as learning data.

Facing the limits the current topology of LOD (as learning data), our research has led to break away from the topic profiles representation of "learn to rank" approach and to adopt a new approach for candidate datasets identification where the recommendation is based on the intensional profiles overlap between different datasets. By intensional profile, we understand the formal representation of a set of schema concept labels that best describe a dataset and can be potentially enriched

by retrieving the corresponding textual descriptions. This representation provides richer contextual and semantic information and allows to compute efficiently and inexpensively similarities between profiles. We identify schema overlap by the help of a semantico-frequential concept similarity measure and a ranking criterion based on the tf*idf cosine similarity. The experiments, conducted over all available linked datasets on the LOD cloud, show that our method achieves an average precision of up to 53% for a recall of 100%. Furthermore, our method returns the mappings between the schema concepts across datasets, a particularly useful input for the data linking step.

In order to ensure a high quality representative datasets schema profiles, we introduce *Datavore*— a tool oriented towards metadata designers that provides ranked lists of vocabulary terms to reuse in data modeling process, together with additional metadata and cross-terms relations. The tool relies on the Linked Open Vocabulary (LOV) ecosystem for acquiring vocabularies and metadata and is made available for the community.

**Titre en Français**: La Recommandation des Jeux de Données -basée sur le Profilage- pour le Liage des Données RDF.

# Résumé

Avec l'émergence du Web de données, notamment les données ouvertes liées (Linked Open Data - LOD), une abondance de données est devenue disponible sur le web. Cependant, les ensembles de données LOD et leurs sous-graphes inhérents varient fortement par rapport à leur taille, le thème et le domaine, les schémas et leur dynamicité dans le temps au niveau des données (respectivement les schémas et les métadonnées). Dans ce contexte, l'identification des jeux de données appropriés, qui répondent à des critères spécifiques, est devenue une tâche majeure, mais difficile à soutenir, surtout pour répondre à des besoins spécifiques tels que la recherche d'entités centriques et la recherche des liens sémantique des données liées. Notamment, en ce qui concerne le problème de liage des données, le besoin d'une méthode efficace pour la recommandation des jeux de données est devenu un défi majeur, surtout avec l'état actuel de la topologie du LOD, dont la concentration des liens est très forte au niveau des graphes populaires multi-domaines tels que DBpedia (le plus fortement ciblé au niveau des liens) et YAGO, alors qu'une grande liste d'autre jeux de données considérés comme candidats potentiels pour le liage est encore ignorée. Ce problème est dû à la tradition du web sémantique dans le traitement du problème de "l'identification des jeux de données candidats pour le liage", où les éditeurs de données sont tenus à identifier les jeux de données appropriés à être avec un jeu de données source.

Bien que la compréhension de la nature du contenu d'un jeu de données spécifique est une condition cruciale pour les cas d'usage mentionnés, nous adoptons dans cette thèse la notion de "profil de jeu de données" - un ensemble de caractéristiques représentatives pour un jeu de données spécifique, notamment dans le cadre de la comparaison avec d'autres jeux de données. Notre première direction de recherche était de mettre en oeuvre une approche de recommandation basée sur le filtrage collaboratif, qui exploite à la fois les profils thématiques des jeux de données, ainsi que les mesures de connectivité traditionnelles, afin d'obtenir un graphe englobant les jeux de données du LOD et leurs thèmes. Cette approche a besoin d'apprendre le

comportement de la connéctivité des jeux de données dans le LOD graphe. Cependant, les expérimentations ont montré que la topologie actuelle de ce nuage LOD est loin d'être complète pour être considéré comme des données d'apprentissage.

Face aux limites de la topologie actuelle du graphe LOD (en tant que données d'apprentissage), notre recherche a conduit à rompre avec cette représentation de profil thématique et notamment du concept "apprendre pour classer" pour adopter une nouvelle approche pour l'identification des jeux de données candidats basée sur le chevauchement des profils intensionnels entre les différents jeux de données. Par profil intensionnel, nous entendons la représentation formelle d'un ensemble d'étiquettes extraites du schéma du jeu de données, et qui peut être potentiellement enrichi par les descriptions textuelles correspondantes. Cette représentation fournit l'information contextuelle qui permet de calculer la similarité entre les différents profils d'une manière efficace. Nous identifions le chevauchement de différentes descriptions textuelles à l'aide d'une mesure de similarité sémantico-fréquentielle qui se base sur un classement calculé par le *tf*idf* et la mesure cosinus. Les expériences, menées sur tous les jeux de données liés disponibles sur le LOD, montrent que notre méthode permet d'obtenir une précision moyenne de 53% pour un rappel de 100%. Par ailleurs, notre méthode peut retourner également des correspondances entre les concepts sématiques des différents jeux de données, ce qui est particulièrement utile dans l'étape de liage.

Afin d'assurer des profils intensionnels de haute qualité, nous introduisons *Datavore* - un outil orienté vers les concepteurs de métadonnées qui recommande des termes de vocabulaire à réutiliser dans le processus de modélisation des données. *Datavore* fournit également les métadonnées correspondant aux termes recommandés ainsi que des propositions des triplés utilisant ces termes. L'outil repose sur l'écosystème des Vocabulaires Ouverts Liés (Linked Open Vocabulary- LOV) pour l'acquisition des vocabulaires existants et leurs métadonnées.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## Contents

## 1.1 The Semantic Web Vision

With the advent of the World Wide Web (WWW), accessing information has become quicker and simpler via Web documents which are part of a "global search space". Knowledge in this version of Web is accessible by traversing hypertext links using Web browsers. On the other hand, search engines index documents on the Web and return potential results based on the introduced user query, i.e., the returned results are ordered by a ranking function based on the structure of links between the documents on the Web [2].

In recent years, this global information space of connected documents is currently evolving to one global Web of data—the Semantic Web—where both data and documents are linked. Tim Berners-Lee, the WWW inventor, defined the Semantic Web as "not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [3]. In other words, instead of supporting a distributed Web at the data level, this information space will be modeled in form of linked graph and represented using Semantic Web standards, i.e., Resource Description Framework standard[1] (as detailed in Section 1.1.1). By this way, the Semantic Web will allow the machines not only to represent data but also to process it as one global knowledge graph, which is as envisioned by the WWW inventor - "*with linked data, when you have some of it, you can find other, related, data.*"

Moreover, during the past few years, dedicated team of people at the World Wide Web consortium (W3C)[2] worked towards improving, extending and standardizing the Semantic Web, which they define as "*the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration, and reuse of data across various applications*" [4]. Different definitions of the Semantic Web reach agreement on the fact that it is **THE machine-readable Web** and can be thought of as an efficient way of representing knowledge on the WWW or as globally linked data. Semantic technologies (i.e., natural language processing (NLP) and semantic search) make use of this machine-readable Web in order to bring meaning to all the disparate and raw data that surround us. Perhaps we can sum up in a summarized Sir Berners-Lee observation: "The Semantic Technology isn't inherently complex. The Semantic Technology language, at its heart, is very, very simple. It's just about the relationships between things.".

In the following we will describe the core Semantic Web format, the Resource Description Framework (RDF), and two basic Semantic Web languages - Resource Description Framework Schema (RDFS) and Web Ontology Language (OWL).

[1] `https://www.w3.org/RDF/`
[2] `https://www.w3.org/Consortium/`

### 1.1.1 RDF Data Model

RDF is the basic building block for supporting the Semantic Web, almost like the significance of HTML for the conventional Web. It is an XML-based language for describing information contained in a Web resource (i.e., a Web page, a person, a city, ...). RDF relies heavily on the infrastructure of the Web, using many of its familiar and proven features, while extending them to provide a foundation for a distributed network of data. In the following we summarize briefly the main properties of RDF:

- RDF (/triples) is a language recommended by W3C [5], which is the format the Semantic Technology uses to store data in graph databases.
- RDF is structured i.e. it is machine-readable allowing interoperability among applications on the Web.
- RDF provides a basis for coding, exchanging, and reusing (meta)data.
- RDF has several basic elements, namely Resource, Property and Statement, which are discussed in the following subsections.

An RDF resource can be anything that can refer to, a Web page, a location, a person, a concept, etc. A resource is described by a set of triples in the form $< s, p, o >$ where the subject $s$ denotes a given resource that has a value $o$ for the property $p$. For example, the triple $< s_1, : firstName, "\text{Mohamed}" >$ states that the resource $s_1$ has as first name "Mohamed". These triples can also be called—RDF statements—and can be expressed logically as binary predicates $p(s, o)$, which is called RDF fact in this case.

An RDF resource is uniquely identified by a Uniform Resource Identifier (URI). For example, the Uniform Resource Locator (URL) `http://www.lirmm.fr/cMohamedBenEllefi` is a URI representing an unique resource "MohamedBenEllefi" under the domain `http://www.lirmm.fr/`. In fact, all URLs, which are commonly used for accessing Web sites, are URIs but not vice versa, example: "tel:+33638641164" is a URI but not a URL. Furthermore, URIs are extended to an Internationalized Resource Identifiers (IRIs) as globally unique identification mechanism [6]. It should be noted that URIs are not just a name but also a means of accessing information of the identified entity.

In an RDF triple, while the subject and the property are URIs, the object can be either a URI or a literal value. For instance, as depicted in Figure 1.1, while `http://www.lirmm.fr` is an object URI, "Mohamed Ben Ellefi" is a literal value.

In this dissertation, we are mainly dealing with RDF datasets, which are represented as a set of RDF triples. This collection of RDF triples can be represented as a

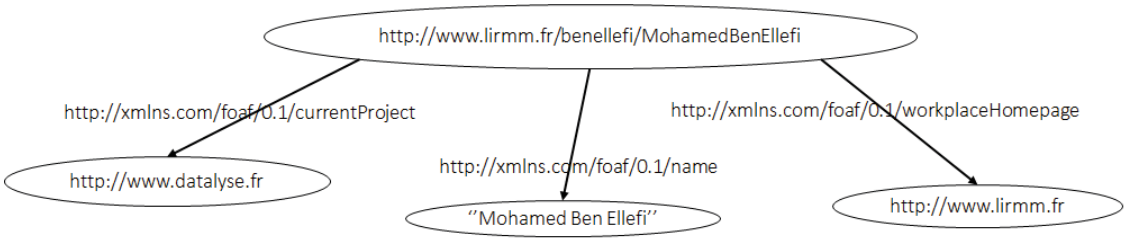


Figure 1.1: RDF triple represented as a directed graph.

labeled, directed multi-graph[3], where a node represents the subject or the object, while an edge represents a property. It should be noted here that a node can be an anonymous resource - known as "blank node". Blank nodes are used basically when the key purpose of a specific resource is to provide a context for some other properties to appear. This graph structure can be easily extended by new knowledge about an identified resource.

Publishing RDF data on the Web requires selecting a serialization format such as N-Triples[4], RDF/XML[5], N3[6] and Turtle[7]. In Figure 1.2, we give an example of RDF data in the N-Triples format.

### 1.1.2 Ontology Languages

In philosophy the term "ontology" refers to the "whatness" question, or in other words "what kinds of things are there?"[8].

In Semantic Web field, the term ontology can be defined as "the terms used to describe and represent an area of knowledge.", according to the W3C vision[9]. In other words, ontologies define the concepts and relationships in order to describe and represent a topic of concern. Moreover, the real added-value of ontologies, with respect to RDF, can be characterized by the introduction of inference techniques on the Semantic Web, i.e., the automatic generation of new (named) relationships based on the data itself and some ontology information.

It should be noted that certain ontologies can referred to as "vocabularies". Basca *et al.* [7] defined vocabularies as simple, lightweight ontologies, that usually comprise

[3]In graph theory, a labeled directed graph is a graph (that is a set of vertices connected by edges), where the edges are labelled and have a direction associated with them. A multi-graph is a graph where two vertices may be connected by more than one edge.

[4]`https://www.w3.org/2011/rdf-wg/wiki/N-Triples-Format`

[5]`https://www.w3.org/TR/REC-rdf-syntax/`

[6]`https://www.w3.org/TeamSubmission/n3/`

[7]`https://www.w3.org/TR/turtle/`

[8]`https://en.wikipedia.org/wiki/Ontology`

[9]`https://www.w3.org/TR/Webont-req/`

```
1 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://www.w3.org
    /1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/
    Person> .
2 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/age> "29"<http://www.w3.org/2001/XMLSchema#int> .
3 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/skypeID> "benellefi" .
4 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/birthday> "01-03-1987"<http://www.w3.org/2001/XMLSchema#
    date> .
5 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/workplaceHome> <http://www.lirmm.fr/> .
6 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/currentProject> <http://www.datalyse.fr/> .
7 <http://lirmm.fr/benellefi/MohamedBenEllefi> <http://xmlns.com/foaf
    /0.1/name> "Mohamed Ben Ellefi"@en .
8 <http://www.datalyse.fr/> <http://xmlns.com/foaf/0.1/name> "
    Datalyse Project"@en .
```

Figure 1.2: An example of RDF data in N-Triples format.

less than 50 terms. In the following, we will present the two main languages for encoding ontologies, RDFS and OWL.

RDFS, which is basically an extension of RDF, is used to create a vocabulary for describing classes, subclasses and properties of RDF resources, i.e., *rdfs:subClassOf*, *rdfs:range* or *rdfs:domain*. Furthermore, based on these relations, a semantic reasoner[10] can understand the RDFS semantics by expanding the number of triples. For instance *Mohamed Ben Ellefi rdf:type :Person* and *:Person rdfs:subClassOf :Human* than a triple *Mohamed Ben Ellefi a :Human* is generated. RDFS is a *W3C* recommendation[11]

OWL is an ontology language with highest level of expressivity than RDFS, and is also a W3C recommendation[12]. The advantage of OWL over RDFS is mainly its capability to express more complex relationships such restriction between classes (e.g. use union of classes as a range of relation) or chained properties. Due to its expressiveness power, most metadata designers use OWL to build ontologies or schema on the top of RDF datasets.

Finally, we note that RDF data are queried using the W3C standard "The SPARQL Protocol And RDF Query Language" - SPARQL[13]. SPARQL has four query forms, specifically SELECT, CONSTRUCT, ASK and DESCRIBE. For example, assume

[10] A semantic reasoner is software able to infer logical consequences from a set of asserted facts (`https://en.wikipedia.org/wiki/Semantic_reasoner`)

[11] `https://www.w3.org/TR/rdf-schema`

[12] `https://www.w3.org/TR/owl-ref/`

[13] `https://www.w3.org/TR/rdf-sparql-protocol/`

that we want to ask the query "what is the current project of Mohamed Ben Ellefi" to our small knowledge base in Figure 1.2. Figure 1.3 depicts a SPARQL query to get information about the current project of Mohamed Ben Ellefi.

```
PREFIX benellefi:<http://lirmm.fr/benellefi/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf
-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?projectname where{
benellefi:MohamedBenEllefi foaf:currentProject ?project.
?project foaf:name ?projectname
}
```

Figure 1.3: SPARQL query to retrieve the name of the current project of Mohamed Ben Ellefi

## 1.2 Current Challenges in Linked Data

The transition of the current document-oriented Web into a Web of interlinked Data - the Semantic Web - has lead to the creation of a global data space that contains many billions of information, the Web of Linked Data[14] (cf. Figure 1.4). In other words, Linked Data can be seen as a deployment path for the Semantic Web.

In this section, we start by an overview of existing linked data life-cycles going to different challenges to achieve this linked knowledge graph, while respecting Linked Data best practices [2].

### 1.2.1 Linked Data Life-Cycles

We start by providing the Tim Berners-Lee vision of the five star Linked Open Data system[15], which is also a $W3C$ recommendation[16], where the system has to ensure: *(1)* availability of data on the Web, in whatever format. *(2)* availability as machine-readable structured data, (example available, as CSV not as scanned image of table). *(3)* availability in a non-proprietary format, (i.e, CSV format instead of Microsoft

[14]The LOD cloud diagram is published regularly at `http://lod-cloud.net`.
[15]`http://5stardata.info/en/`
[16]`https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html`

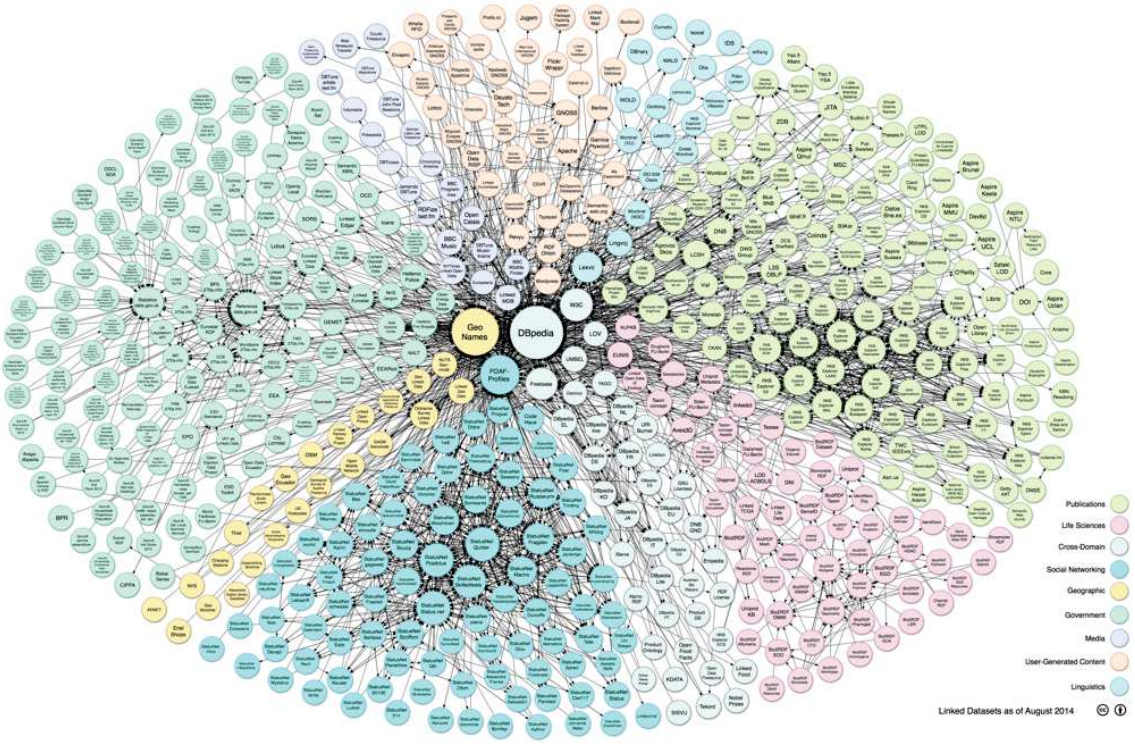


Figure 1.4: Overview of Linked Data knowledge bases in the LOD cloud (2014)

Excel). *(4)* publishing using open standards from (the W3C recommendation, RDF and SPARQL). *(5)* linking to other Linked Open Data, whenever feasible.

For this purpose, a number of Linked data life-cycle visions have been adopted by the Semantic Web community, where data goes through a number of stages to be considered as Linked Data. Figure 1.5 depicts five visions of different governmental Linked data life-cycles: Hyland *et al.* [8], Hausenblas *et al.* [9], Villazon-Terrazas *et al.* [10], the DataLift vision [11] and the LOD2 Linked Open Data Lifecycle [12]. Since there is no standardised Linked Data Life-Cycle, the main stages of publishing a new dataset as Linked Data can be summarized as follows:

- Extracting and transforming information from the raw data source to RDF data. Mainly, the data source can be in unstructured, structured or semi-structured format, i.e., CSV, XML file, a relational database, etc. In this stage, the cast does not include vocabulary modeling, namespaces assignment nor links to existing datasets.

- Assigning namespaces to all resources, notably make them accessible via their URIs.

- Modeling the RDF data by reusing relevant existing vocabularies terms whenever possible. Linked Data modeling process is detailed in Section 6.1.

- Hosting the linked dataset and its metadata publicly and make it accessible.
- Linking the newly published dataset to other datasets already published as Linked Data on the Web.

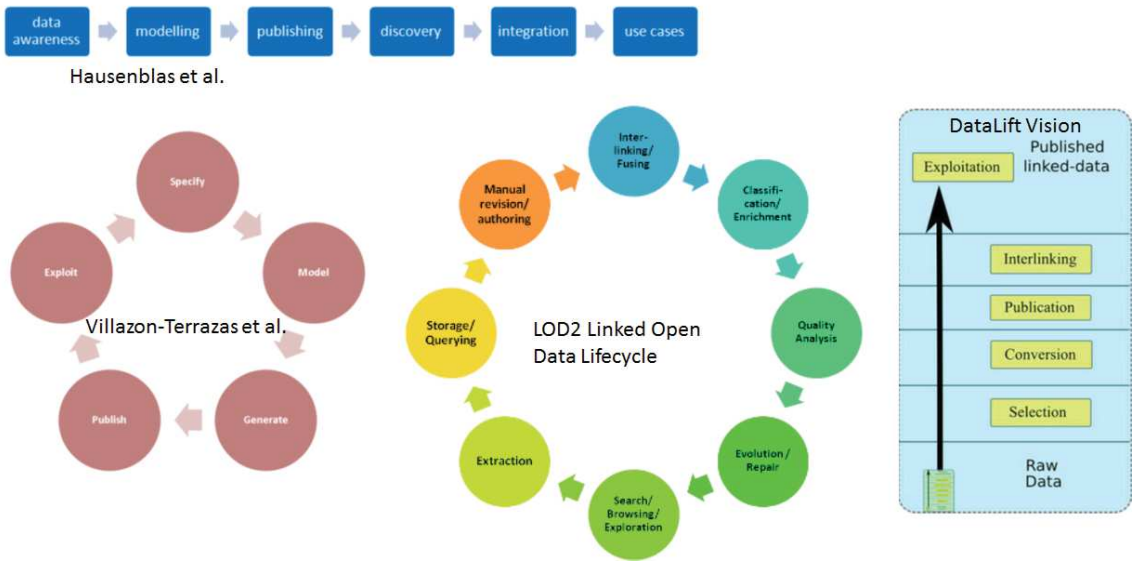


Figure 1.5: Five Examples of Governmental Linked Data Life-Cycles.

It should be noted that different stages of the linked data life-cycle are not in a strict sequence nor exist in isolation, but are in mutual enrichment.

### 1.2.2 Current Challenges

Some stages in the Linked Data life-cycles are not completely automated and may need human effort either for Linked data publishing, maintenance or consuming. Hence, in recent years, an increasing number of works have shown interest in the development of research approaches, standards, technology and tools for supporting different stages of the Linked Data life-cycle.

For instance, Linked Data publishers face a major challenge of: how to ensure a data store (i.e., large triple stores[17]) to be highly effective and ensure performance with a large scale data ? For this purpose, currently, there exists a wealth of works contributing to this central problem by developing distributed data storage solutions (e.g., Hadoop Distributed File System (HDFS), Google File System (GFS)) which can work in server clusters scattered in the cloud and handle massive amounts of triples.

[17]https://www.w3.org/wiki/LargeTripleStores

Another current challenge concerns the Linked Data modeling process which requires a huge effort from metadata designers, in particular on the issues raised by: how to identify suitable terms from published vocabularies in order to reuse them following the Linked data modeling best practices.

Further challenges include semantic links discovery between different RDF datasets, which is becoming manually unfeasible, considering the large amount of data available on the Web. Usually, among the different kinds of semantic links that can be established, the default option is to set *owl:sameAs* links between different URIs that refer to the same real objects. For example, DBpedia uses the URI `http://dbpedia.org/page/Montpellier` to identify the city of *Montpellier*, while Geonames uses the URI `http://www.geonames.org/2992166` to identify *Montpellier*. Data linking techniques and tools are often used to deal with this problem. However, this task still requires human involvements, notably on: *(i)* the identification of candidate datasets to link to, where the search for target ones should be done almost by an exhaustive search of all datasets in the different catalogues, which is rather manually not feasible; and *(ii)* the strenuous task of instance mappings configuration between different datasets.

We note that some stages in the Linked Data life-cycles are proceeded automatically, notably the automatic extraction and transformation of raw data which has lead to the publication of large amount of knowledge in the Web as Linked Datasets. However, automatic approaches have raised many questions regarding the quality, the currentness and the completeness of the contained information. Hence, the major challenge during this process concerns mainly the assessment of the data quality.

Some issues can arise after publishing Linked Data knowledge facing both data publishers and data consumers. Regarding data publishers (or rather maintainers), they have the responsibility to ensure a continued maintenance of the published dataset in terms of quality, i.e., the access, the dynamicity (different versions, mirrors), etc. On the other hand, the efforts of Linked Data Consumers revolve around the following requirements:

- finding and retrieving suitable information from different linked open datasets.
- integrating and reusing this knowledge.
- ensuring continued feedback for data maintainers.

In the light of the above, this thesis will address two main challenges: *(i)* the Linked Data modeling task, more precisely, we focus on the vocabulary items recommendations; and *(ii)* the data linking challenge, notably, we address the problem of candidate datasets discovery aiming at finding target datasets to link to. In other words, the two challenges that we will address will be like "***looking for a needle in a haystack***", whether in looking for suitable terms in vocabulary catalogues[18],

[18]vocabulary catalogues such as `http://lov.okfn.org/`

or in searching for target datasets in large amount of data available on Linked Data catalogues[19].

Linked datasets vary heavily with respect to their size, topic and domain coverage, the resource types and schemas or the dynamics and currency. To this extent, the discovery of suitable datasets which satisfy specific criteria face the major challenge of "the understanding of the nature of the content of specific datasets". To this end, this dissertation starts by dealing with the issue of: *what is the set of representative features that better describe an RDF dataset for a given task?*

The different contributions of this thesis will be outlined in the following section.

## 1.3 Contributions and Objectives

As stated in the previous section, the broad question that arises is *How to reduce human effort in linking and modeling RDF data?* To do so, our contributions aim to validate the following thesis:

> Profile-based recommendation methods are able to significantly reduce human effort in linking and modeling RDF data.

In this section, we start by representing our first contribution which consists in introducing a new notion of features-based RDF dataset profiling which leads to provide a classification of an extensive state of the art on dataset profiling methods. Our next contributions aim to deal with the challenge of candidate datasets identification for the interlinking. For this purpose, we have developed two different approaches:

- Topic Profile-based Dataset Recommendation for Data Linking
- Dataset Recommendation for Data Linking: an intensional approach

In line with the datasets intensional profiles quality, our further contribution address the Linking Data modeling process where we developed a new vocabulary recommendation tool - *Datavore*.

### 1.3.1 Dataset Profiling

This contribution aims to be a guide for dataset profiles features extraction methods with respect to a given application scenario. A dataset profile can be broadly defined as a comprehensive set of commonly investigated features that describe an RDF dataset. Furthermode, we provide a taxonomic classification of existing approaches in the field of profiles extraction systems with respect to the following profiles features:

[19]Linked Data catalogues such as `http://ckan.org/`

- Semantic: domain/topic, context, index elements and schema/instances
- Qualitative: trust, accessibility, context, degree of connectivity and representation.
- Statistical: schema level and instance level
- Temporal: global, instance-specific and semantics specific

To the best of knowledge, this is the first study of its kind. We note here that the choice of the feature is essentially based on what is known by the community as a representative feature with respect to a given application scenario.

### 1.3.2 Identifying Candidate Datasets for Data Linking

Data linking is the main challenge that we address in this dissertation, notably we focus on the question of: what is the suitable set of datasets to select as target candidate to be linked with a given RDF data source?

In the following, we present our two different approaches of candidate dataset recommendation:

#### 1.3.2.1 Topic Profile-based Dataset Recommendation for Data Linking

Our research direction has been to develop a dataset recommendation system aiming to reduce human effort in the data linking process. To this extent, we adopt a well-known and efficient technique in the field of recommendation systems, the "Collaborative Filtering (CF)" based recommendation [13]. To do so, our system will learn knowledge from two graphs:
*(1)* The topic-dataset-graph, produced through the profiling method of Fetahu *et al.* [1], as a representative profiles for datasets. *(2)* The already established connectivity graph between datasets which is measured by the existence or not of links.

By this way, we are able to learn the connectivity behavior of datasets using their topics profiles in the current topology of the LOD (considered as evaluation data) and subsequently, to provide an efficient CF-based candidate dataset recommendation for interlinking. Furthermore, a subsequential contribution of the learning step is the introduction of new method for topic profiles propagation to the entire LOD graph in inexpensive manner.

An extensive experimental evaluation of the approach has been conducted, where we used real world LOD datasets. To the best of our knowledge there is not a common benchmark in this field, hence, we developed three simple baselines. The

proposed approach showed a superior performance compared to the introduced baselines. Moreover, we demonstrate a global performance of our technique with respect to the evaluation data by achieving the following results:

- Average recall: 81%.
- Average precision: 19%.
- A reduction of the original (LOD) search space of up to 86% on average.

#### 1.3.2.2 Dataset Recommendation for Data Linking: an Intensional approach

As can be observed in the previous approach, the performance in terms of precision needs improvement, we explain that by the amount of explicitly declared links in the LOD cloud as learning data. Hence, our next contribution consists of providing a new approach for candidate dataset recommendation that adopts the notion of *intensional profile* and skips the learning step (unlike the first approach).

By intensional profile, each dataset is represented by a set of schema information in the form of concept labels enriched by their corresponding textual descriptions. The main strength of this profiling method is the fact that a profile can be easily constructed for any RDF dataset in a simple and inexpensive way. Based on the intensional profile representation, our approach provides a candidate dataset recommendation by the help of a semantico-frequential similarity measure and a ranking criterion based on the tf*idf cosine similarity model.

Similarly to the first approach, we conducted experiments using current LOD topology as evaluation data where the intensional approach shows a high recommendation efficiency by achieving an average precision of up to 53% for a recall of 100%. Furthermore, we note that our method is able to return mappings between the schema concepts across datasets. This mapping is particularly useful for the configuration of linking tools.

### 1.3.3 Vocabulary Selection for Linked Data Modeling

This contribution aims to ensure a high quality datasets intensional profiles by assisting metadata designers in the Linked Data modeling process. To this extent, we introduce our tool - *Datavore* - that provides ranked lists of vocabulary terms to reuse together with the descriptive metadata via its interactive graphical user interface (GUI). Furthermore, *Datavore* is able to detect existing possible cross-terms relations, providing to the user a list of suitable triples. For acquiring existing

vocabularies and metadata, the tool relies on the Linked Open Vocabulary (LOV) [20] as a trusty ecosystem. *Datavore* is made available as prototype for the community.

It should be noted that the totality of contributions in this thesis are targeted to be integrated in the LOD life-cycle of the DATALYSE project[21].

## 1.4 Thesis Overview

Below we provide an overview of how the thesis is organized, along with the main contributions of each chapter.

Chapter 2 provides the necessary background on the datasets profiles and introduces our vision of datasets profiles features. In the second part of this chapter, we present different profiling techniques notably the profile extraction techniques and their LOD use-cases. The different techniques are classified in a big taxonomy of profiles features.

Chapter 3 presents some prominent application scenarios of RDF dataset profiles where used features are identified explicitly and aligned with the introduced profiles taxonomy features. The second part of this chapter highlights the linked data application scenario, and in particular, the candidate dataset identification which is the main challenge of this dissertation. Finally, we cite the existing works on this thematic.

Chapter 4 presents our topic profile-based approach that we have developed to discover candidate datasets for interlinking. We first describe the topic profile approach that we adopt for this recommendation. Then, we present the workflow of the proposed Collaborative Filtering-like recommendation approach. Finally, we describe the used evaluation framework followed by an extensive experimental evaluation for the proposed approach.

Chapter 5 describes our intension-based approach for candidate dataset recommendation. First, we start by introducing the notion of intensional profiles with respect to the profiles taxonomy features. Then, we depict different steps in the workflow of the introduced recommendation approach. After that, we present the experimentations that have been conducted to evaluate the efficiency and effectiveness of our approach compared to the evaluation data and the baselines, which have been developed for this purpose. Finally, we discuss the positioning of intentional-based recommendation approach with respect to the first recommendation approach and current related works.

[20] `lov.okfn.org`
[21] `http://www.datalyse.fr/`

Chapter 6 focuses on increasing metadata designers awareness regarding vocabulary reuse in the context of LOD datasets. First, we provide a detailed discussion of the various types of support features for modeling LOD datasets. Next, we present our prototype *Datavore* and illustrate the general concept of our approach.

Chapter 7 provides a conclusion and discusses various proposals for future work.

# Chapter 2

# A Guide to RDF Dataset Profiling Features

## Contents

# Introduction

Linked open data (LOD), as provided by a quickly growing number of sources, constitutes a wealth of easily accessible information. Usually, these LOD datasets evolve over time when more facts and entities are added. However, the "rigid" schema and ontology that model these datasets do not follow the data dynamicity and lose validity over all entities of the dataset. Hence, in order to consume these data, it is vital to thoroughly examine and understand each dataset, its structure, and its properties. Since manual inspection requires a huge effort besides than been a limited extent in term of information completeness, different profiling techniques are proposed to deal with these challenges.

This chapter contains literature overview on dataset profiling approaches. The aim of this chapter is to give the reader a bird's-eye view of the different profiling techniques notably the used terminologies the profile extraction techniques and their LOD use-cases. First, we describe the basic concepts of LOD dataset profiling by providing a comprehensive set of commonly investigated dataset features (*cf.* Section 2.1). These features are based on the existing literature in the field of dataset profiling (whether or not referred to explicitly by using this term) and are organized into a taxonomy, formally represented as an RDF vocabulary of dataset profiling features and made available to the public. Further, we provide a broad overview of the existing approaches and tools for the automatic extraction of such features (*cf.* Section 2.2). Finally, we conclude in Section 2.3.

## 2.1 RDF Dataset Profiling Conceptualization and Features

In this dissertation, we are dealing with RDF datasets which can be formally defined as follow:

**Definition 1 (RDF dataset)** . *Let $U$ be a set of URIs, $L$ be a set of literals and $B$ be the set of blank nodes. An RDF dataset is a set of triples $< s, p, o >$ where $s \in (U \cup B), p \in U$ and $o \in (U \cup B \cup L)$.*

The main focus of this section is to stress on RDF data profiling terminology and definition, hence, we start by citing a broad definition of data profiling from Wikipedia[1]: *"Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that data".*

In line with the Semantic Web principles, there are several definitions of dataset

[1] `https://en.wikipedia.org/wiki/Data_profiling`

profiling, among which we cite, the vision of [14] where data profiling is *"an umbrella term for methods that compute meta-data for describing datasets"*.

Furthermore, when dealing with LOD dataset profiling, independently on the intended application, one is faced with the question of "what are the features that describe a dataset and allow to define it?"; one needs to be able to determine these characteristics that, gathered together, provide a descriptive summarization of a dataset, with regard to a certain aspect or application purpose. In this dissertation, we provide a formal definition of LOD dataset profile:

**Definition 2 (Dataset Profile)** . *A dataset profile can be seen as the formal representation of a* ***set of features*** *that describe a dataset and allow the comparison of different datasets with regard to their characteristics.*

Usually, the relevant feature set is dependent on a given application scenario and task.

The present chapter makes an inventory of dataset features that can be considered in the dataset profiling process. We take into account features that have been studied in the literature, although often referred to by using different names or defined differently. We propose a common terminology and definitions of (and distinction between) the key terms used in the field. We organize the features in a (non-strict) hierarchy and we introduce the notion of an *basic feature*, understood as one that has no descendants in this hierarchy.

### 2.1.1 Semantic Characteristics

We present a set of features that carry semantic information about dataset.

1. **Domain/Topic** – A domain refers to the field of life or knowledge that the dataset treats (e.g., music, people). It describes and englobes the topics covered by a dataset [15] (e.g., life sciences or media), understood as more granular, structured metadata descriptions of a dataset, as the one found in [16]. The cross-domain or multi-topical nature of a dataset is separate feature that indicates of its potential connectivity properties and its specificity.
2. **Context** – We identify two members of this group:
   (a) ***connectivity properties***, meaning concretely the set of RDF triples shared with other datasets, and
   (b) ***domain/topical overlap with other datasets***. Important information, especially with regard to user queries, can be made available by the overlap of the domains or topics covered by a dataset and other datasets. This overlap can be expressed, for instance, by the presence of shared

topics between two datasets [17], [18]. A contextual profile provides additional information to the selected dataset for user queries based on the overlap with other datasets on schema or instances levels. the contextual profile is mainly intended to find relevant datasets for user queries.

3. **Index elements** – Index models have been introduced in order to retrieve information from the LOD. An index is defined as a set of key elements (e.g., types), which are used to lookup and retrieve data items. These elements can be defined on schema level or on entity level. A dataset, therefore, can be inversely described by the set of index elements that are pointing to it in a given index or a set of indices [19]. In that sense, a set of index elements is viewed as a descriptive semantic dataset characteristic.

4. **Representative Schema/Instances** – This group of features is found on schema and on instance level and is understood as a set of types (schema concepts) or a set of key properties/values, or a representative sample of instances [20], [21], [22].

### 2.1.2 Qualitative Characteristics

According to Mendes et al. [23], the problem of data quality is related to values being in conflict between different data sources as a consequence of the diversity of the data. For Bizer et al. [24] relate data quality problems to those arising in web-based information systems, which integrate information from different providers. Hogan et al. [25] discuss about errors, noise, difficulties or modelling issues, which are prone to the non-exploitations of the data from the applications. According to Zaveri et al. [26], the term data quality problem refers to a set of issues that can affect the potentiality of the applications that use the data. Thus the study of data quality has a strong and on-going tradition. Data quality assessment involves the measurement of quality dimensions that are relevant to the consumer. The dimensions can be considered as the characteristics-Profile- of a dataset. Here, we provide a list of features related to several of these dimensions. Many of these features apply to data quality in general and are directly issued from [27]. However, some of them have been defined particularly in the context of linked data [26].

1. **Trust** – Trust is a major concern when dealing with LOD data. Data trustworthiness can be expressed by the following features.

   (a) ***verifiability:*** the "degree and ease with which the information can be checked for correctness", according to [28].

   (b) ***believability:*** the "degree to which the information is accepted to be correct, true, real and credible" [29]. This can be verified by the presence of the provider/contributor in a list of trusted providers [26].

(c) ***reputation:*** a judgement made by a user to determine the integrity of a source [26]. Two aspects are to take into consideration:

i. ***reputation of the data publisher:*** an indice coming from a survey in a community that determines the reputation of a source,

ii. ***reputation of the dataset:*** scoring the dataset on the basis of the references to it on the web.

(d) ***licensing policy:*** the type of license under which a dataset is published indicates whether reproduction, distribution, modification, redistribution are permitted. This can have a direct impact on data quality, both in terms of trust and accessibility (see below).

(e) ***provenance:*** "the contextual metadata that focuses on how to represent, manage and use information about the origin of the source" [26].

2. **Accessibility** – This family of characteristics regards various aspects of the process of accessing data.

(a) ***availability:*** the extent to which information is available and easily accessible or retrievable [28].

(b) ***security:*** refers to the degree to which information is passed securely from users to the information source and back [26].

(c) ***performance:*** the response time in query execution [26].

(d) ***versatility of access:*** a measure of the provision of alternative access methods to a dataset [26].

3. **Representativity** – The features included in this group provide information in terms of noisiness, redundancy or missing information in a given dataset.

(a) ***completeness:*** the degree to which all required information regarding schema, properties and interlinking is present in a given dataset [26]. In the Linked Data context, [28] defines the following sub-features:

i. ***schema completeness (ontology completeness)*** – the degree to which the classes and properties of an ontology are represented,

ii. ***property completeness*** – measure of the missing values for a specific property,

iii. ***population completeness*** – the percentage of all real-world objects of a particular type that are represented in the datasets, and

iv. ***interlinking completeness*** – refers to the degree to which links are not missing in a dataset.

(b) ***understandability:*** refers to expression, or, as defined by [29], the extent to which data is easily comprehended.

(c) ***accuracy / correctness:*** the equivalence between a value in a dataset instances and the actual real world value of it.

(d) ***conciseness:*** the degree of redundancy of the information contained in a dataset.

(e) ***consistency:*** the presence of contradictory information.

(f) ***versatility:*** whether data is available in different serialization formats, or in different formal and/or natural languages.

4. **Context / task specificity** – This category comprises features that tell something about data quality with respect to a specific task.

   (a) ***relevance:*** the degree to which the data needed for a specific task is appropriate (applicable and helpful) [29], or the importance of data to the user query [28].

   (b) ***sufficiency:*** the availability of enough data for a particular task. [28] uses the term "amount-of-data".

   (c) ***timeliness:*** the availability of timely information in a dataset with regard to a given application. For example, is there enough data on a timely subject in biological studies at the present moment?.

5. **Degree of connectivity** – Connectivity here is understood as simply the number of datasets, with which a dataset is interlinked, or as the number of triples in which either the subject or the object come from another dataset (note the difference with contextual connectivity in the class of semantic features and interlinking completeness in the representation class of features).

### 2.1.3 Statistical Characteristics

This group of characteristics comprises a set of statistical features, such as size and coverage or average number of triples, property co-occurrence, etc. [30], [31].

1. **Schema-level** – According to schema, we can compute statistical features such as *class / properties usage count*, *class / properties usage per subject and per object* or *class / properties hierarchy depth*.

2. **Instance-level** – Features on this level are computed according to the data only, i.e.,*URI usage per subject (/object), triples having a resource (/blanks) as subject (/object), triples with literals, min(/max/avg.) per data type (integer / float / time, etc.), number of internal and external links, number of ingoing*

*(/outgoing) links per instance, number of used languages per literal, classes distribution as subject (/object) per property, property co-occurrence*

### 2.1.4 Temporal Characteristics

This class of features concerns the dynamicity of a dataset (as identified in a catalogue like Datahub) [32], [33]. Every dataset feature is dynamic, i.e., changing over time (take for example data quality). Inversely, the dynamics of a dataset can be seen as a feature of, for example, quality. For that reason, this family of features is seen as transversal (spanning over the three groups of features described above). A profile characteristic can, therefore, be based on the dynamicity estimation of a dataset over an extended period of time measured over a set of aspects that we have classified in the following groups.

1. **Global** –
   (a) ***lifespan:*** measured on an entire dataset or parts of it.
   (b) ***stability:*** an aggregation measure of the dynamics of all dataset characteristics.
   (c) ***update history:*** a feature with multiple dimensions regarding the dataset update behavior, divided into:
      i. ***frequency of change:*** the frequency of updating a dataset, regardless to the kind of update.
      ii. ***change patterns:*** the existence and kinds of categories of updates, or change behavior.
      iii. ***degree of change:*** to what extent the performed updates impact the overall state of the dataset.
      iv. ***change triggers:*** the cause or origine of the update as well as the propagation effect reinforced by the links.
2. **Instance-specific** –
   (a) ***growth rate:*** the level of growth of a dataset in terms of data entities (instances).
   (b) ***stability of URIs:*** the level of stability of URIs i.e., an URI can be moved, modified or a removed.
   (c) ***stability of links:*** the level of broken links between resources, i.e., a links is considered as broken link if the a target URIs changes [34]
3. **Semantics-specific** [35] [36] –

(a) ***structural changes:*** evaluation of the degree of change in the structure (internal or external) of a dataset.

(b) ***domain-dependent changes:*** this feature reflects the dynamics across different domains that impacts the data.

(c) ***vocabulary-dependent changes:*** a measure of the dynamics of vocabulary usage.

(d) ***vocabulary changes:*** a measure of the impact of a change in a vocabulary to the dataset that uses it.

(e) ***stability of index models:*** the level of change in the original data after having been indexed.

Finally, we would like to draw the readers' attention to the fact that we provide a representation of the profile features hierarchy in the form of a new RDF vocabulary of dataset profiles (VoDP) and make it available to the community via the following link: `http://data.data-observatory.org/vocabs/profiles/ns`.

## 2.2 RDF Dataset Profiling Methods Classification-Literature Review

In this section, we provide a pivotal guideline for the approaches for dataset profiling, as well as the systems and tools for dataset features extraction, following the categorization introduced in the previous section. An overview of the dataset features and the corresponding extraction systems is shown in Figure 2.1 and described in detail below.

### 2.2.1 Semantic Characteristics Extraction

**FluidOps Data Portal**[2] [17] is a framework for source contextualization. It allows the users to explore the space of a given source, i.e., search and discover data sources of interest. Here, the contextualization engine favors the discovery of relevant sources during exploration. For this, entities are extracted/clustered to give for every source a ranked list of contextualization sources. This approach is based on well-known data mining strategies and does not require schema information or data adhering to a particular form.

[2]The FluidOps Data Portal is currently tested by a pilot customer and is available on data.fluidops.net and.

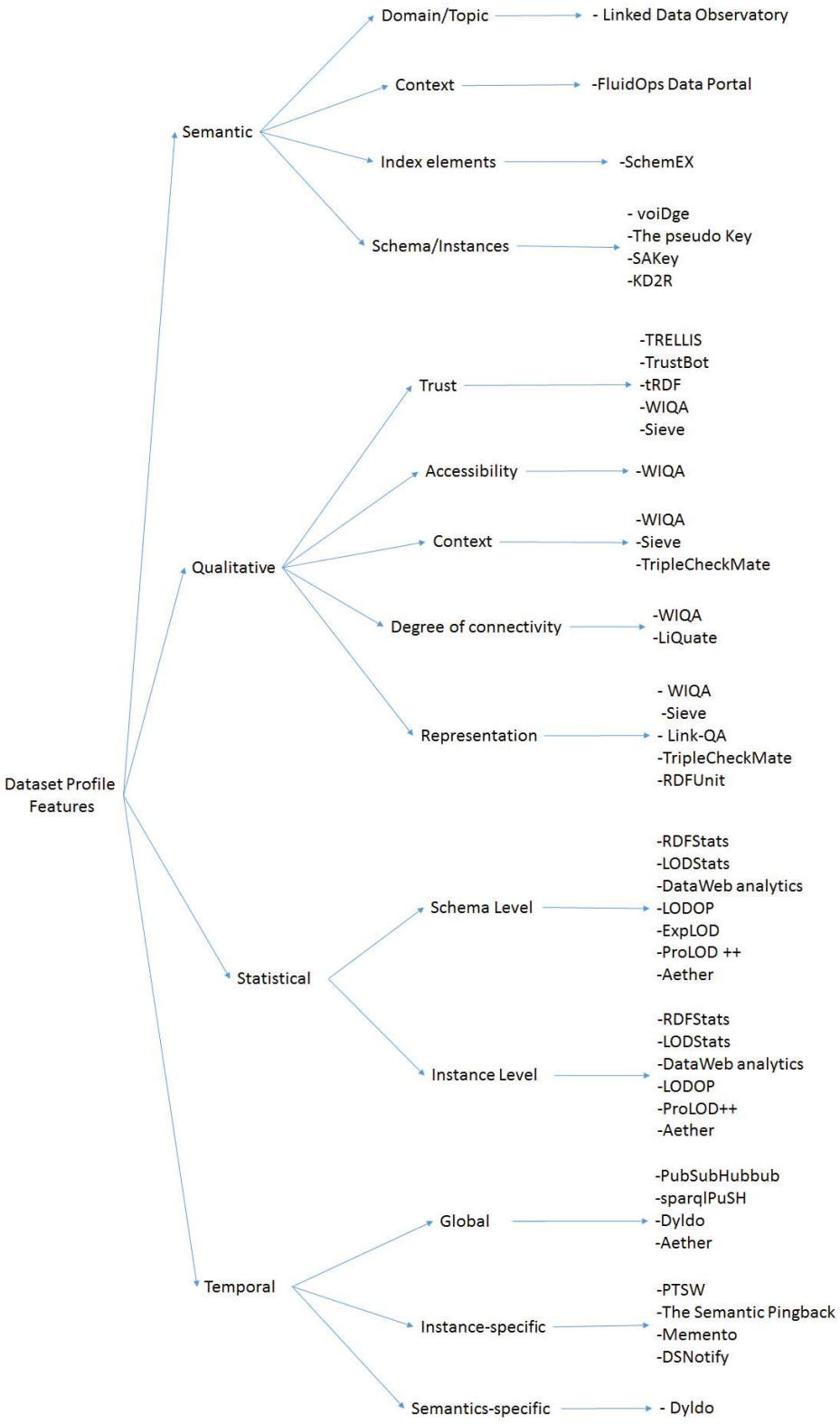


Figure 2.1: Overview of dataset profile features and extraction systems.

**Linked Data Observatory**[3] [16] provides an explorative way to browse and search through existing datasets in the LOD Cloud according to the topics which are covered. By deploying entity recognition, sampling and ranking techniques, the Linked Data Observatory allows to find datasets providing data for a given set of topics or to discover datasets covering similar fields. This Structured Dataset Topic Profiles are represented in RDF using the VoID vocabulary in tandem with the Vocabulary of Links (VoL)[4].

**voiDge**[5] is a tool that automatically generates VoID (Vocabulary of Interlinked Data) descriptions for large datasets. This tool allows users to compute the various VoID informations and statistics on dumps of LOD as illustrated in [37]. Additionally, the tool identifies (sub)datasets and annotates the derived subsets according to the VoID specification.

**The keys discovery approaches** aim at selecting the smallest set of relevant predicates representing the dataset in the instance comparison task. We review two keys discovery approaches: ***(i)*** The *pseudo-Key* [21], a relaxed version of a key that tolerates a few instances having the same values for the properties, and ***(ii)*** *SAKey*[38] – an approach to discover *almost keys* in datasets where erroneous data or duplicates exist. *SAKey* is an extension of *KD2R*[39] which aims to derive exact composite keys from a set of non keys discovered on RDF data sources. The pseudo-Key and the almost keys approaches mainly differ on the level of the semantic discovery of identity links.

**SchemEX**[19] is a stream-based indexing and schema extraction approach over the LOD. The schema extraction abstracts RDF instances to RDF schema concepts that represent instances with the same properties. The index is each schema concept that maps to data sources containing instances with corresponding properties.

### 2.2.2 Quality Assessment Systems

Zaveri *et al.* [26] provide an extensive survey of 21 works on linked data quality assessment based on quality dimensions, the respective metric, types of data and

[3]The Linked Data Observatory demo is publicly available according to LOD principles at http://data-observatory.org/lod-profiles/index.htm

[4]VOL (`http://data.linkededucation.org/vol/index.htm`) provides a general vocabulary to describe metadata about links or linksets, within or accross specific datasets. VoL was designed specifically to represent additional metadata about computed links which cannot be expressed with default RDF(S) expressions and enable a qualification of a link or linkset.

[5]The source code and the documentation of the *voiDge* tool can be downloaded on http://hpi.de/naumann/projects/btc/btc-2010.html

level of automation. In this section, we focus on tools that are implemented and available.

**TRELLIS**[6] [40] is an interactive environment that examines the degree of trust of datasets based on user annotation. The user can provide Trellis with semantic markup of annotations through interaction with the ACE tool[7] [41]. The tool allows several users to add and store their observations, viewpoints and conclusions. The annotations made by the users with ACE can be used in TRELLIS to detect conflicting information or handle incomplete information.

**TrustBot** [42] is an Internet Relay Chat bot that make trust recommendations to users based on the trust network it builds. It allows users to transparently interact with the graph by making a simple serie of queries. Users can add their own URIs to the bot at any time, and incorporate the data into the graph. The bot keeps a collection of these URIs that are spidered when the bot is launched or called upon to reload the graph. TrustBot is a semi-automatic tool to gauge the trustworthiness of the data publisher.

**tRDF**[8] [43] is a framework that provides tools to represent, determine, and manage trust values that represent the trustworthiness of RDF statements and RDF graphs. it contains a query engine for tSPARQL, a trust-aware query language. *tSPARQL* is an extension of the RDF query language SPARQL in two clauses; TRUST AS clause and the ENSURE TRUST clause. The trust values are based on subjective perceptions about the query object. The users can query the dataset and access the trust values associated to the query solutions in a declarative manner.

**WIQA**[9] [24] is a set of components to evaluate the trust of a dataset using a wide range of different filtering policies based on quality indicators like provenance information, ratings, and background information about information providers. This framework is composed of two components: a Named Graph Store for representing information together with quality related meta-information, and an engine which enables applications to filter information and to retrieve explanations about filtering decisions. WIQA policies are expressed using the WIQA-PL syntax, which is based on the SPARQL query language.

[6]TRELLIS is an open-source tool and available online at http://www.isi.edu/ikcap/trellis/demo.html

[7]Annotation Canonicalization through Expression synthesis

[8]This tools are available online on http://trdf.sourceforge.net/tsparql.shtml

[9]WIQA is an open-source tool and available on http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/impl

**Sieve**[10] [23] is the quality evaluation module in *LDIF* (Linked Data Integration Framework). To assess the quality of a dataset, the user can choose which characteristics of the data indicate higher quality, how this quality is quantified and how should it be stored in the system. This is enabled by a conceptual model composed of assessment metrics, indicators and scoring functions (TimeCloseness, Preference, SetMembership, Threshold and Interval Membership). Sieve aimed mainly to perform data fusion (integration) based on quality assessment.

**Link-QA**[11] [44] is a framework for detection of the quality of linksets using five network metrics (degree, clustering coefficient, open *sameAs* chains, centrality, description richness through *sameAs*). This framework is completely automatic and takes as input a set of resources, SPARQL endpoints and/or dereferencable resources and a set of triples. The workflow consists of five components: Select of set of, Construct, Extend, Analyse and Compare.

**LiQuate**[12] [45] is a tool to assess the quality related to both incompleteness of links, and ambiguities among labels and links. This quality evaluation is based on queries to a Bayesian Network that models RDF data and dependencies among properties.

**TripleCheckMate**[13] [46] is a user-driven quality evaluation tool. The system will provide the user with a list of classes wherein he can choose the ones he is most familiar with. There are three options: ***(i)*** Any: where a completely random resource will be retrieved, ***(ii)*** Class: where one has the option to choose a random resource belonging to that class will be retrieved, ***(iii)*** Manual: where you can manually put in the *DBpedia* URI of a resource of your choice. After selecting a resource, the user will be shown each triple belonging to that resource. The system allow to evaluate each triple whether it contains an error or not. If it contains an error, the user can select an error type from a suggested list.

**RDFUnit**[14] [47] is a framework for the data quality tests of RDF knowledge based on Data Quality Test Pattern, DQTP. A pattern can be:***(i)*** a resource of a specific type should have a certain property, ***(ii)*** a literal value should contain at most one literal for a certain language. The user can select and instantiate existing DQTPs. If the adequate test pattern for a given dataset is not available, the user has to write his own DQTPs, which can then become part of a central library to facilitate later re-use.

[10]This tool is open-source and available on http://sieve.wbsg.de/development
[11]Link-QA is Open-source and available on http://bit.ly/Linked-QA.
[12]The demo is published at http://liquate.ldc.usb.ve.
[13]TriplecheckMate is open source and available on https://github.com/AKSW/TripleCheckMate.
[14]This framework is on http://aksw.org/Projects/RDFUnit.html

### 2.2.3 Extraction of Statistical Characteristics

**RDFStats**[15] [48] is a framework for generating statistics from RDF data that can be used for query processing and optimisation over SPARQL endpoints. Thoses statistics include histograms for subjects (URIs, blank nodes) and histograms for properties and associated ranges. RDFStats can be integrated into user interfaces and other Semantic Web applications to provide this information but also to support tools to achieve a better performance when processing large amount of data.

**LODStats**[16] [30] is a statement-stream-based tool and framework for gathering comprehensive statistics about datasets adhering RDF. The tool calculates 32 different statistical criterions on LOD such as those covered by the VoID Vocabulary. It computes descriptive statistics such as the frequencies of property usage and datatype usage, the average length of literals, or the number of namespaces appearing at the subject URI position. It is available for integration with CKAN metadata repository, either as a patch or as an external web application using CKAN's API.

**Data Web analytics**[17] [49] examines the growth of the LOD Cloud since 2007. It provides statistics about the Cloud containing multiple aspects such as the usage of vocabularies as well as provenance and licensing information. The main difference to LODStats is that this information is partially entered manually in the Data Hub and updated infrequently, whereas with LODStats these calculations can be performed automatically.

**LODOP**[18] [31] is a framework for computing, optimizing, and benchmarking statistics for Linked Datasets. This system provides a total of 56 scripts, which compute 15 different statistical properties across different subsets of the input dataset. This statistical properties are determined via the following types of groupings : by resource, property, class, class and property, datatype, context URL, vocabulary, language, object URI, or no grouping. **ExpLOD** [50] creates usage summaries from RDF graphs including meta-data about the structure of a RDF graph, such as the sets of instantiated RDF classes of a resource or the sets of used properties. This structure information is aggregated with statistics like the number of instances per class or the number of property usage.

[15] http://rdfstats.sourceforge.net/
[16] https://github.com/AKSW/LODStats/wiki/LODStats-clean-install
[17] http://lod-cloud.net/state/
[18] https://github.com/bforchhammer/lodop/

**ProLOD ++**[19] [51] is an interactive user interface, which is divided into a cluster tree view and a details view. The cluster view enables users to explore the cluster tree and to select a cluster for further investigation for statistics. *ProLOD ++* is extension of *ProLOD*[52] which generated basic statistics. In addition to the mining and the cleansing tasks of *ProLOD ++*, the tool generates profiling features like related to key analysis, predicate and value distribution, string pattern analysis, link analysis and data type analysis.

**Aether**[20] [53]is a tool that generates automatically an extended VoID statistical profile from a *sparql*1.1 endpoint This statistical profile can be viewed in a graphical interface with the viewer module. In addition, *Aether* provides a temporal profile by allowing the comparison between datasets versions and a qualitative profile by detecting outliers and errors. The generated extensions of the VoID description represent statistics in both schema and instance level.

### 2.2.4 Temporal Characteristics Extraction Systems

**PubSubHubbub**[21] [54] is a decentralized real-time Web protocol that delivers data to subscribers when they become available. Parties (servers) speaking the PubSubHubbub protocol can get near-instant notifications when a topic (resource URL) they're interested in is updated.

**sparqlPuSH**[22] [55] is an interface that can be plugged on any SPARQL endpoint and that broadcasts notifications to clients interested in what is happening in the store using the PubSubHubbub protocol i.e., $SPARQL + pubsubhubbub = sparqlPuSH$. Practically, this means that one can be notified in real-time of any change happening in a SPARQL endpoint. A resource can ping a PubSubHubbub hub when it changes, then, the notifications will be broadcasted to interested parties. sparqlPuSH consists in two steps: *(i)* register the SPARQL queries related to the updates that must be monitored in a RDF store, *(ii)* broadcast changes when data mapped to these queries are updated in the store.

**Ping the Semantic Web (PTSW)**[56] is a web service archiving the location of recently created/updated RDF documents. If a document is created or updated, its author can notify PTSW that the document has been created or updated by pinging

[19]https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/app.html, https://github.com/HPI-Information-Systems/ProLOD

[20]A demo of *Aether* is available on http://demo.seco.tkk.fi/aether/

[21]https://code.google.com/p/pubsubhubbub/

[22]https://code.google.com/p/sparqlpush/

the service with the URL of the document. This protocol is used by crawlers or other types of software agents to know when and where the latest updated RDF documents can be found. PTSW is dedicated to Semantic Web documents and all the sources they may come from: blogs, databases exported in RDF, hand-crafted RDF files, etc.

**The Semantic Pingback**[57] is a mechanism that allows users and publishers of RDF content, of weblog entries or of an article to obtain immediate feedback when other people establish a reference to them or their work, thus facilitating social interactions. It also allows to publish backlinks automatically from the original WebID profile (or other content, e.g. status messages) to comments or references of the WebID (or other content) elsewhere on the Web, thus facilitating timeliness and coherence of datasets. It is based on the advertisement of a lightweight RPC[23] service.

**Memento**[58,59] is a protocol-based time travel that can be used to access archived representations resources. The current representation of a ressource is named the *Original Resource*, whereas resources that provide prior representations are named *Mementos*. This system provides relationships like the *first-memento, last-memento, next-memento* and *prev-memento*. Mementos are available both in HTML and RDF/XML. For example, a demo of Memento for the DBpedia URI `http://dbpedia.org/data/Tim_Berners-Lee` can be found in `http://mementoarchive.lanl.gov/dbpedia/memento/20090701/http://dbpedia.org/data/Tim_Berners-Lee`.

**The Dynamic Linked Data Observatory (Dyldo)**[32], [33] is a framework to achieve a comprehensive overview of how LOD changes and evolves on the Web. It is an observatory of the dynamicity on the Web of Data (snapshots) over time. The dataset provides weekly crawls of LOD data sources starting from the 2nd of November 2008 and contains 550K RDF/XML documents with a total of 3.3M unique subjects with 2.8M locally defined entities. The system examines, firstly, the usage of Etag and Last-Modified HTTP header fields, followed by an analysis of the various dynamic aspects of a dataset (change frequency, change volume, etc).

**DSNotify**[24][34] is a Link monitoring and maintenance framework, which attenuates the problem of broken links due to the URI instability. When remote resources are created, removed, changed, updated or moved, the system revises links to these resources accordingly. This system can easily be extended by implementing custom crawlers, feature extractors, and comparison heuristics.

[23]Remote procedure call.
[24]http://www.cibiv.at/ niko/dsnotify

## 2.3 Conclusion

In this chapter, we have introduced the definition of RDF dataset profiling that we adopt in this dissertation. Furthermore, we have provided a comprehensive survey of existing research aimed at supporting the dataset profiling task, a central challenge when facilitating dataset discovery in tasks such as entity retrieval, distributed search or entity linking. It should be noted that given the complexity of the topic, we have focused on first providing an exhaustive taxonomy of dataset features, also available as a structured RDF vocabulary, and then surveyed methods for assessing and extracting such features from arbitrary datasets.

The following chapter will outline the main applications that make use of the hierarchy of dataset profiles features and argue the effectiveness of datasets profiling methods in the datasets comparison challenge.

# Chapter 3

# Datasets Profiling Applications

## Contents

# Introduction

The adoption of the linked data best practices [2] has led to the extension of the web with a global data space connecting data from diverse domains (e.g., people, scientific publications, music). Furthermore, we can observe three types of generic linked data applications:

**(i) Generic linked data browsers** which navigate between HTML pages by following hypertext links, Linked Data browsers allow the navigation between data sources by following RDF triples links.

**(ii) Linked data search engines** that crawl the Web of Data by following links between data sources and provide expressive query capabilities over aggregated data, similar to how a local database is queried today. In other words, while browsers provide the mechanisms for navigating, search engines offer the place at which that navigation process begins. Example of search engines are Falcon [60] and Swoogle[1].

**(iii) Domain-specific linked data applications** that offer specific functionality by merging data from various Linked Data sources for a better domain linked data consumption.

This chapter will provide an overview of several linked data applications, which make use of the dataset profiles notably in the fields of data curation, query applications and information retrieval leading to the main aim of this dissertation profiles-based RDF datasets comparison for the data linking.

In line with the Definition. 2, we recall that by a *dataset profile* we understand the formal representation of a set of features that describe a dataset and allow the comparison of different datasets with regard to their represented characteristics. However, this relevant feature set is completely dependent on a given linked data application scenario.

We start by identifying explicitly subsets of features that are considered relevant in particular prominent application scenarios, discussed and analysed in detail in Section 3.1. Furthermore, the section will put a special focus on the data linking application and notably the features to be used in the datasets comparison task. Section 3.2 will provide a general overview of existing candidate datasets identification techniques with respect to the data linking task preceded by a motivation for this process. Finally, we conclude in Section 3.3.

[1] `http://swoogle.umbc.edu/`

## 3.1 Current Applications for Datasets Profiling

In this section, we outline the use of different profiles features by different applications with respect to the approaches:
(1) *IR.*
(2) Data curation, cleansing, maintenance, etc.
(3) Schema inference
(4) Query federation, optimization, reformulation, etc.
(5) Data linking.

It should be noted that the work described in Sections [ 3.1.1 - 3.1.4 ] are extracted from our under-review paper [61].

### 3.1.1 Information Retrieval (*IR*)

In information retrieval applications, Linked Data is mostly used in the context of semantic search, typically aiming to solve the limitations of keyword-based models which is limited to the understanding of literal strings rather than the mining. For instance, keyword-based models does not distinguish between search terms "books about recommender systems" *vs.* "systems that recommend books", as demonstrated in [62].

Semantic search has been one of the motivations of the Semantic Web since it was envisioned. According to [63], Information Retrieval on the Semantic Web, a search engine returns documents rather than, or in addition to, exact values in response to user queries. For this purpose, this approach includes an ontology-based scheme for the semi-automatic annotation of documents and a retrieval system. The retrieval model is based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm. Semantic search is combined with conventional keyword-based retrieval to achieve tolerance to knowledge base incompleteness.

In Information retrieval a statistical methods are used to measure the strength of the semantic relation between words. They are able to capture only small fraction of important relations such as part-of-time dimensions.

The datasets providing semantic features to enrich a text with additional information, not given there explicit, allow one to go beyond standard Bag of Words representation [64]. Wide range of methods based on linking to external, domain-oriented resources has been proposed, e.g., [65], [66], [67]. They also employ statistical features extracted from large-scale text corpora [68] and allow one to expand the user queries to increase recall [69].

Enriching a text for *IR* and related tasks (such as text categorization, eg. [70]) shows improvement as measured by formal evaluation frameworks, i.e., at the annual TREC conference[2] [62]. Providing lexical information to text processing tasks involve dataset features related to representation

In addition, geographical and temporal contexts play an increasingly important role in *IR* applications. These contexts enable retrieval of information relevant with respect to the spatial [71] and temporal [72] dimensions of the query.

In many retrieval tasks a geographical context is needed to add facilities that can rank the relevance of candidate information with respect to geographical closeness as well as semantic relatedness according to the topic of interest. In [71] an example of spatial features usage has been shown to support *IR* tasks.The approach employs an ontology of places that combines limited coordinate data with qualitative spatial relationships between places. Beside spatial information, a time is also an important dimension of any information space. It is also used in retrieval to determine a documents' credibility [73]. The overview of the methods and applications aiming at satisfying search needs by combining the traditional notion of document relevance with temporal relevance is given in [72]. The demonstration of such approach has been shown in YAGO2 [74], where search interface allows one to seek temporal and spatial knowledge facts.

Early stage *IR* systems used bilingual dictionaries to support a user in selecting terms in the language of the documents being searched. To allow better control of the generation of the new query in the target language bilingual dictionaries have been replaced by a pair of language-ontology lexicons. It provides the user a view on definitions of the senses in an ontology and then select matching terms in the target language [75].

The majority of the semantic search applications are domain-oriented, and a large number of practical cases have been shown for repositories related to biomedical sciences. For example, the concept-based search mechanism [76] allows biologists to describe the topics of the search interest more specifically and retrieve the information with higher precision (in comparison to usage of keywords only). It should be stressed here that the concept-based search requires linking to high-quality external resources (such as, e.g., UMLS [77]), which involves features related to trust (cf Section 1), especially verifiability and believability.

However, methods for extracting biomedical facts from the scientific literature have improved considerably, and the associated tools will probably soon be used in many laboratories to automatically annotate and analyzes the growing number of system-wide experimental datasets [78]. Owing to the increasing body of text and the open-access policies of many journals, literature mining is also becoming useful for both hypothesis generation and biological discovery. However, the latter will require

[2] Text REtrieval Conference (TREC), http://trec.nist.gov/

the integration of literature and high-throughput data, which should encourage close collaborations between biologists and computational linguists. One of the examples of applications that employ domain features in medical area is Textpresso [79] – a text-mining system for scientific literature whose capabilities go far beyond those of a simple keyword search engine. Textpresso's two major elements are a collection of full texts of scientific articles split into individual sentences, and the implementation of categories of terms, based on Gene Ontology [80], using which a repository of articles and individual sentences can be searched. The categories are classes of biological concepts (e.g., gene, allele, cell or cell group, phenotype, etc.) and classes that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., biological process, etc.). Together they form a catalog of types of objects and concepts called ontology. After this ontology is populated with terms, the whole corpus of articles and abstracts is marked up to identify terms of these categories. The current ontology comprises 33 categories of terms. A search engine enables the user to search for one or a combination of these tags and/or keywords within a sentence or document, and as the ontology allows word meaning to be queried, it is possible to formulate semantic queries. Full text access increases recall of biological data types from 45% to 95%. Extraction of particular biological facts, such as gene-gene interactions, can be accelerated significantly by ontologies, with Textpresso [79] automatically performing nearly as well as expert curators to identify sentences; in searches for two uniquely named genes and an interaction term, the ontology confers a 3-fold increase of search efficiency. Textpresso currently focuses on Caenorhabditis elegans literature, with $3,800$ full text articles and $16,000$ abstracts. The lexicon of the ontology contains $14,500$ entries, each of which includes all versions of a specific word or phrase, and it includes all categories of the Gene Ontology database [80]. Textpresso is a useful curation tool, as well as a search engine for researchers, and is ready to be extended to other domain-specific literature.

The other ontology-based information retrieval system [81] - MELISA (MEdical Literature Search Agent) is a prototype for medical literature sources. The model is based on an architecture with three levels of abstraction, the use of separated ontologies and query models, and the definition of some aggregation operators to combine results from different queries.

Text-mining in molecular biology [82] - the automatic extraction of information about genes, proteins and their functional relationships from text documents - has emerged as a hybrid discipline on the edges of the fields of information science, bioinformatics and computational linguistics. A range of text-mining applications have been developed recently that will improve access to knowledge for biologists and database annotators.

**Features for Information Retrieval applications:** In line with the the dataset profile features taxonomy, depicted in Figure2.1, *IR* involves qualitative profile features related to trust (cf. Section 1) (i.e., verifiability and believability) and the

accessibility of data(cf. Section 3). In addition, to preserve the semantic search, *IR* implies profiles features like (cf. Section 1) topical domains, and context (cf. Section 2).

### 3.1.2 Data Curation, Cleansing, Maintenance

The advent of semantic web technologies has has invaded the Web by large amount of knowledge, represented as Linked Open Data (LOD) in form of RDF triples. However, many datasets were extracted from unstructured and semistructured information using automated extraction approaches, or are the result of some crowdsourcing process. This often raises questions about the quality, the currentness and the completeness of the contained information. For example, a recent user study on DBpedia - a hub node of the Linked Data Cloud - uncovered a large number of quality problems with respect to accuracy, relevancy, representational consistency and interlinking [83].

In the context of Linked Data, statistical approaches to error detection and type prediction are shown to be more effective than the standard ontology reasoning techniques due to their independence of the background knowledge and robustness to noise [84]. Therefore, a number of recent works focus on statistical methods for: (1) Outlier detection to detect errors in numerical values [85], [84], [86]; (2) Automatic prediction of missing types of instances [84]; and (3) Identification of wrong links between datasets [87]. A further line of research in Linked Data quality is related to the discovery of errors in the data based on the existing interlinking (e.g., [88], [89]). Thereby some works go beyond error detection and attempt to automatically determine correct data values in case of inconsistencies [88].

In the following we discuss selected recent approaches and features they use in more detail.

Fleischhacker et al. [85] focus on identifying errors in numerical linked data using statistical outlier detection. In the first step, the authors determine the properties and their sub-populations to which numerical outlier detection can be applied. This step includes collection of statistical information such as the number of instances and properties in the dataset along with the property-specific statistics. The latter provide insides in numerical value usage and distribution for a property. In the second step, distributions of numerical values in the properties are analyzed to detect outliers. Finally, the authors verify identified outliers against the values in other datasets using *owl:sameAs* references. This final verification step helps to differentiate natural outliers (such as the highest mountain) from real errors in numerical data. Wienand et al. [86] address the same problem by grouping instances by their types before the outlier detection is applied.

Data inconsistency can also be frequently observed in multilingual DBpedia [88].

Whereas DBpedia offers a rich source of entity-related data in multiple languages, each DBpedia language edition evolves in isolation often leading to mutual inconsistency in entity representations across languages. Bryl et al. [88] consider a use case of resolving the conflicts in these datasets for the entity type *dbpedia-owl:PopulatedPlace*. The proposed conflict resolution strategy involves usage of the provenance metadata of the statements in question extracted from the Wikipedia revision history. According to [88] the most effective features in this use case include frequency of values, update frequency of properties in a specific language as well as the overall activity of the editors.

**Features for error detection in numerical values:** In [85] the authors detect errors in numerical values using outlier detection. To identify the properties to which numerical outlier detection can be applied, the following statistical characteristics (discussed in Section 2.1.3) are used: (1) total number of instances, (2) names of the properties used in the dataset, (3) frequency of usage with numerical values in the object position for each property, and (4) total number of distinct numerical values for each property.

**Features for conflict resolution in multilingual DBpedia:** The features used in conflict resolution in [88] include provenance metadata at the statement, property and author levels. The temporal dataset profile 1 includes in particular: (1) recency of the specific statement (measured using the time of the last edit), (2) overall editing frequency of the property in the dataset, and (3) the overall number of edits performed by the specific editor.

### 3.1.3 Schema Inference

Many existing Linked Data sources do not explicitly specify schemas, or only provide incomplete specifications. However, many real-world applications (e.g., answering queries over distributed data [90]) rely on the schema information. Recently, approaches aimed at automatic inference of missing schema information have been developed (e.g., [84],[19]).

For example, planning and answering queries over distributed data typically includes creation of a mediated schema [90].

Paulheim et al. [84] infer missing type statements for instances based on the property-specific distribution of types. First, for each property, the frequencies of types in the subject and object positions are collected. Second, to assign a type to a non-typed instance, these property-specific statistics are aggregated over the incoming and outgoing properties of this instance using a probabilistic model. Thereby the properties can be weighted to reflect their predictive power. The results of a comparison to a standard reasoning approach in [84] suggest that such statistical

type inference is more robust with respect to the noise in the data. Finally, in [84] the authors use type distributions to identify incorrect statements.

**Features for type inference:** Statistical characteristics of datasets (see Section 2.1.3) play an important role in the type inference applications. For example, in [84] statistics on the completeness of type statements as well as property-specific type distributions are required (i.e. the types of resources appearing in subject and object positions of each property including their frequencies).

### 3.1.4 Distributed Query Applications

Linked Data Cloud can be queried either through direct HTTP URI lookups or using distributed SPARQL endpoints [91] that can include full-text search extensions (see e.g., [92]). Also combinations of both query paradigms are possible [93]. Typically, the first step of query answering over distributed data is the generation of ordered query plans against the mediated schema on a number of data sources [94]; In this step, dataset profiling plays an important role.

In order to guide distributed query processing, existing applications rely on indexes of varying granularity including *Schema-level Indexes* and *Data Summaries*. *Schema-level Indexes* contain information about properties and classes occurring at certain sources. *Data Summaries* use a combined description of instance- and schema-level elements to summarise the content of data sources [91]. The majority of existing federated query approaches for LOD (e.g., [93], [91], [95], [96]) are aimed to optimize for efficient query processing and do not (yet) take quality parameters of LOD sources into account. Therefore, existing *Data Summaries* mostly contain frequencies and interlinking statistics of varying granularity.

For example, data summaries for on-demand queries over linked data in [91] employ counts of data items such as subjects, objects and predicates to optimize the queries. [95] performs top-k query processing over Linked Data and assumes existence of a ranking function determining importance of triples, using e.g., triple-level interlinking statistics or triple counts. For another example, Elbassuoni et al. [97] proposes language-based model for ranking query results of exact, relaxed, and keyword-augmented graph-structured queries over RDF graphs. To estimate query result probability [97] uses witness count (i.e. number of times a triple is observed in external sources). The authors obtain such witness counts using the number of hits of a web search engine.

A more recent approach [94] takes quality parameters into account and builds a *Conditional Data Quality Profile* estimating quality parameters of the data relevant for the specific query dynamically.

**Granularity of profiling for quality-aware query applications:** An important

aspect of data profiling for query answering is the granularity of statistics in the profile.

One of the earliest approaches addressing the problem of data quality in query answering for integrated systems is presented by Naumann et al. [98]. Here, the quality profiles are collected at the data source level, i.e. a data source is associated with a vector of quality metrics, such as overall completeness, accuracy, etc. within the source. These vectors do not reflect possible quality differences across properties; However, in the context of Linked Data such property-specific differences are essential as properties within a source oft differ with respect to the quality aspects. The property-specific quality variations in the data source profiles are considered in [99]. Here, a vector with quality metrics is associated with each property. Although such property-specific vectors provide more insides, they still do not take into account possible quality differences within each property. For example, in the context of Linked Data such quality differences may arise across entities of different types sharing a common property.

Recently, [94] presented a query framework that takes into account query-specific aspects of the data sources along with the user preferences with respect to data quality. To return the most relevant results, this framework uses data quality preferences expressed in the user profile and an estimation of the result quality for specific query. The focus of this work is the user- and query-specific quality estimation as opposed to overall quality estimation of the source and its properties. The authors propose a *Conditional Data Quality Profile* to improve quality estimation for queries bound with conditions. This work is focused on equality comparisons (but does not cover range selection conditions).

In this context, the authors discuss the trade-offs between the accuracy of the quality estimation and the overhead incurred to generate accurate statistics for conditional data quality profiles.

**Features for efficient and quality-aware query applications:** The majority of existing query applications rely on semantic and statistical characteristics (see Sections 2.1.1 and 2.1.3) at the schema-level, i.e. properties and classes occurring at certain sources for effective query interpretation. In addition, applications that optimize for efficient query processing require data-level statistics (including frequency and interlinking) either on triple level or for each subject, object and predicate individually [91]. Finally, quality-aware query applications also take into account qualitative characteristics (see Section 2.1.2) (e.g., completeness and accuracy) at different granularity levels. This includes overall data source statistics [98], as well as property-specific [99] and type-specific statistics [94].

### 3.1.5 Data Linking Applications

Data linking applications aim at generating links between entities from different data sources that refer to the same real world object. This task is referred as instance matching (IM), entity resolution, interlinking, reference reconciliation, also, it can be in line with record linkage for databases and entity annotation in the text mining context. For entity annotations, we cite some popular tools and services using NLP techniques such as *DBpedia Spotlight* [100], *Illinois Wikifier* [65], and the multilingual tools *Babelfy* which is a dedicated multilingual annotation service supporting 50 languages [101].

On the other hand, some data linking approaches are numerical[3] by using complex similarity measures, aggregation functions and thresholds to build linkage rules. For example, interlinking systems like LIMES [102] or SILK [103], require link specification files where the user is expected to provide the names of the classes, in which to look for instances to match and the properties whose values to compare. In this context, data linking can be seen as a linking operation taking as input two different datasets and produces as output a set of links between similar instances with different descriptions linked with the "owl:sameAs" link or the set of SKOS "Match" properties. Hence, an important step towards the data linking task is the discovery of relevant datasets that may contain similar resources to be linked with the identity or other type of relations. For this purpose, candidate datasets discovery task requires a comparison framework that rank different target datasets with respect to a given source. In particular, we need to identify the set of profiles features that describe a dataset and allow the comparison of different datasets with regard to their represented characteristics.

**Features for data linking applications:** Data linking applications described above typically use semantic features discussed in Section 2.1.1 such as topics, domains, languages (versatility) 3 and location coverage 2, as well as representative parts of schema/instances, and specifically the key candidates extracted with the keys discovery approaches that is particularly useful in the configuration step. For candidate dataset discovery, we will go through different used profile features in the remainder of this chapter.

## 3.2 Candidate Datasets Identification for Data Linking

The purpose of this section is to stress on the challenge of "finding candidate datasets to link to".

[3]Numerical approachs are used to find numerical approximations to the solutions

### 3.2.1 Motivation

In line with the data linking applications, the first challenge facing the data linker is to identify potential target datasets that may contain similar instances as a given source dataset. For this purpose, let us take a step back from the naive methods which have been usually adopt, i.e., one of two following solutions: (i) applying the brute force for combining all the possible pairs of datasets to the interlinking tool; and (ii) requesting the user for selecting the most suitable datasets following his beliefs.

However, with the huge growth of the web of data, and notably LOD cloud, an exhaustive search of all existing datasets in available catalogues, is manually unfeasible. Hence, the most common linking tradition is limited to target popular and cross domain datasets such as DBpedia [104], whereas many other potential LOD datasets has been ignored. This led to an inequitable distribution of links and consequently, a limited semantic consumption in the Linked Data graph.

Recommender systems can provide effective solutions to reduce human efforts in searching candidate datasets challenges. Recommender systems have been an active research area for more than a decade. Typically, recommendation approaches are based on explicit score ratings for candidate items. This scoring can produce an ordered list of suitable results or even more to reduce the search space to the *Top N* most suitable recommendation.

In this dissertation, we adopt this recommendation concept in order to find suitable target datasets for the interlinking. This task is, also, known under the names of: candidate datasets "identification", respectively, "selection" and "recommendation".

### 3.2.2 Current Works on Candidate Datasets Selection

With respect to finding relevant datasets on the Web, we cite briefly several studies on discovering relevant datasets for query answering have been proposed. Based on well-known data mining strategies, the works in [18] and [17] present techniques to find relevant datasets, which offer contextual information corresponding to the user queries. A used feedback-based approach to incrementally identify new datasets for domain-specific linked data applications is proposed in [105]. User feedback is used as a way to assess the relevance of the candidate datasets. Also, we cite the LODVader [106], a framework for LOD Visualisation, Analytics and Discovery, which proposes to compare datasets using Jaccard coefficient based on *rdf:type*, *owl:Classes* and general predicates.

In the following, we cite approaches that have been devised for candidate dataset recommendation for the interlinking task and which are directly relevant to our work.

Nikolov *et al.* [107] propose a keyword-based search approach to identify candidate sources for data linking. The approach consists of two steps: (i) searching for potentially relevant entities in other datasets using as keywords randomly selected instances over the literals in the source dataset, and (ii) filtering out irrelevant datasets by measuring semantic concept similarities obtained by applying ontology matching techniques.

Leme *et al.* [108] present a ranking method for datasets with respect to their relevance for the interlinking task. The ranking is based on Bayesian criteria and on the popularity of the datasets, which affects the generality of the approach (*cf.* the cold-start problem discussed previously). The authors extend this work and overcome this drawback in [109] by exploring the correlation between different sets of features—properties, classes and vocabularies— and the links to compute new rank score functions for all the available linked datasets.

Mehdi *et al.* [110] propose a method to automatically identify relevant public SPARQL endpoints from a list of candidates. First, the process needs as input a set of domain-specific keywords which are extracted from a local source or can be provided manually by an expert. Then, using natural languages processing techniques and queries expansion techniques, the system generates a set of queries that seek for exact literal matches between the introduced keywords and the target datasets, i.e., for each term supplied to the algorithm, the system runs a matching with a set of eight queries: {original-case, proper-case, lower-case, upper-case} * {no-lang-tag, @en-tag}. Finally, the produced output consists of a list of potentially relevant SPARQL endpoints of datasets for linking. In addition, an interesting contribution of this technique is the bindings returned for the subject and predicate query variables, which are recorded and logged when a term match is found on some particular SPARQL endpoint. The records are particularly useful in the linking step.

A recent approach is presented by Röder et al. [111], where authors present *Tapioca*, a linked dataset search engine for topical similarity of datasets. Topics are frequent schema classes and properties extracted from the dataset metadata. The similarity of two datasets is defined as the similarity of their topic distributions which are extracted using the Latent Dirichlet Allocation – a generative model for the creation of natural language documents.

As stated before, to the best of our knowledge, only few existing approaches aim to deal with the candidate datasets identification challenge. Furthermore, none of the studies outlined above have been evaluated in term of real world LOD datasets, except for [109] approach which, according to the authors, achieves a mean average precision of around 60%.

Hence, in this dissertation, we have chosen to face the challenge of providing a greater efficiency when dealing with real world LOD datasets. To this end, we will introduce a new candidate dataset recommendation approach for the interlinking

process.

## 3.3 Conclusion

In the beginning of this chapter, we gave an overview of what profiles features that can be used in application like information retrieval, distributed query, data curation and notably data linking. We should retain that, given the continuous evolution and expansion of the Web of data, the problem of dataset profiling will become an increasingly important one, and corresponding methods will form a crucial building block for enabling reuse and take-up of datasets beyond established and well-understood knowledge bases and reference graphs.

In the second part of this chapter, we addressed the task of candidate datasets identification for data linking where a profiling methods are crucial for the datasets comparison process and, consequently for datasets ranking. In the remainder of this dissertation, we will present two different approaches for datasets identification for the interlinking process showing the impact of adopting different profile features. The proposed approaches will be evaluated via relevant experiments based on real world datasets from the LOD cloud.

# Chapter 4

# Dataset Recommendation for Data Linking based on Topic Profiles

## Contents

# Introduction

The wide variety and heterogeneity of web datasets characteristics, in particular Linked Open Data (LOD) [112], pose significant challenges for data consumers when attempting to find useful datasets without prior knowledge of available datasets. Dataset registries such as Datahub[1] or DataCite[2] aim at addressing this issue, for instance, by enabling users and data providers to annotate their datasets with some basic metadata, for instance, descriptive tags and access details. However, due to the reliance on human annotators, such profile are often sparse and outdated [1]. This has contributed to the fact that, the majority of data consumption, linking and reuse focuses on established datasets and knowledge graphs such as DBpedia or YAGO, while a long tail of datasets has hardly been reused and adopted.

However, with the huge growth of the Linked Open Data Cloud, this task is becoming harder and harder to proceed manually. Hence, the proposed solution for easing this task is the automatic recommendation of a set of candidate datasets to be linked, which consequently, reduces considerably the search space.

This chapter will meet the data linking challenge, presented in Section 3.1.5, in the process of candidate datasets identification for interlinking. Our recommendation approach relies on the notion of a dataset profile, with its large definition, providing comparable representations of the datasets by the help of characteristic features.

In this chapter, we provide a new recommendation method based on the direct relatedness of datasets as emerging from the topic-dataset-graph produced through the profiling method of Fetahu *et al.* [1]. In line with our datasets profiles features taxonomy (*cf.* Fig.2.1), the adopted topic profile is fitted in the semantic characteristics part, as described in Section and 1 . Furthermore, we adopt established *collaborative filtering* practices by considering the topic relationships emerging from the global topic-dataset-graph to derive specific dataset recommendations. We exploit dataset connectivity measures to relate non-profiled datasets to datasets in the dataset-topic-graph, enabling us to consider arbitrary datasets as part of our recommendations. The intuition is that this leads to more robust and less error-prone recommendations, since the consideration of global topic connectivity provides reliable connectivity indicators even in cases where the underlying topic profiles might be noisy. Our assumption is that even poor or incorrect topic annotations will serve as reliable relatedness indicator when shared among datasets.

In our experiments, we apply our approach to the LOD cloud as one scenario and use case, where dataset recommendation is of particular relevance. Our experiments show superior performance compared to three simple baselines, namely based on shared key-words, shared topics, and shared common links. In a series of exper-

[1] `http://datahub.io`
[2] `http://datacite.org`

iments, we demonstrate the performance of our technique compared to the current version of the LOD as an evaluation data, achieving a reduction of the original (LOD) search space of up to 86% on average.

The contributions of the dissertation that will be presented in this chapter are:

- A dataset recommendation technique based on topic-profiles.
- An efficient approach of propagating dataset profiles over the entire LOD cloud.
- A set of baseline recommendation approaches, made available to the community as a benchmark.

This chapter is organized as follows. Section 4.1 formalises the used notions and definitions of datasets topic profiling and datasets connectivity. Section 4.2 discuss the topic profiles approach that we adopted as a preprocessing step. Section 4.3 presents the theoretical grounds of our technique. Finally, Section 4.5 depicts the conducted experiments of our approach using real world LOD datasets before concluding in Section 4.6. This chapter is based on [113].

## 4.1 Preliminaries

We start by introducing notation and definitions. Let $T_1, ..., T_N$ be a number of topics from a set of topics $\mathcal{T}$ and let $\mathcal{D} = \{D_1, ..., D_M\}$ be a set of datasets.

### 4.1.1 Dataset Topic Profile

Topic modeling algorithms such as Latent Dirichlet allocation [114] are used to discover a set of topics from a large collection of documents, where a topic is a distribution over terms that is biased around those associated under a single theme. Topic modeling approaches have been applied to tasks such as corpus exploration, document classification, and information retrieval. Here, we will look into a novel application of this group of approaches, exploiting the topic structure in order to define and construct dataset profiles for dataset recommendation.

As a result of the topic modeling process, a bipartite—*profile*—graph is built, providing a relation between a document and a topic. Documents in our setting are the datasets to be considered, therefore the profile graph is induced by the relation between a dataset, $D_i$, and a topic, $T_k$, expressed by a weight, $w_{ik} \in [0, 1]$, for all $i = 1, ..., M$ and $k = 1, ..., N$. Formally, a profile graph is defined as follows.

**Definition 3 (Dataset Topic Profile Graph (DTPG))** *A dataset topic profile graph is a weighted directed bipartite graph $\mathcal{P} = (\mathcal{S}, \mathcal{E}, \Delta)$, where $\mathcal{S} = \mathcal{D} \cup \mathcal{T}$, $\mathcal{E}$ is a*

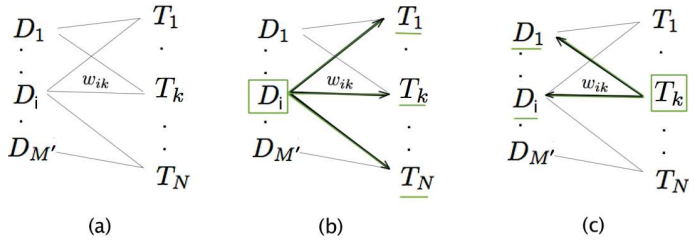


Figure 4.1: (a) An example of a bipartite profile graph with topics and datasets linked by weighted edges. (b) Representing a dataset, $D_i$, as a set of topics. (c) Representing a topic, $T_k$, as a set of datasets.

*set of edges of the form* $e_{ik} = (D_i, T_k)$ *such that* $D_i \in \mathcal{D}$ *and* $T_k \in \mathcal{T}$ *and*

$$\begin{aligned} \Delta \colon \mathcal{E} &\to [0,1] \\ e_{ik} &\mapsto w_{ik} \end{aligned}$$

*is a function assigning weights to the edges in* $\mathcal{E}$*.*

The bipartite property of $\mathcal{DTPG}$ allows to represent a given dataset by a set of topics—its *profile*. For the purposes of this study, it is worth noting that, inversely, a topic can be represented by a set of weighted datasets—what we will call the *signature* of a topic (see Figure 4.1). We will denote by Profile($D_i$) the function returning the topic profile of $D_i$, i.e., the set of topics together with their weights with respect to $D_i$. Inversely, we will denote by $\mathcal{D}_{T_k}$ the set of datasets together with their weights with respect to a topic $T_k$, derived again from the graph $\mathcal{DTPG}$.

### 4.1.2 Datasets Connectivity

The connectivity behavior of datasets is a central concept within the proposed recommendation framework. We consider the following definition of a measure of the strength of dataset connectedness.

**Definition 4 (Dataset inter-connectivity measure)** *Let* $D_i, D_j \in \mathcal{D}$ *be two datasets. We define a measure of their common degree of connectivity as follows.*

$$\mathcal{C}(D_i, D_j) = \frac{shared(D_i, D_j) \times [total(D_i) + total(D_j)]}{2 \times total(D_i) \times total(D_j)} \tag{4.1}$$

*where* $shared(.,.)$ *returns the number of links between two datasets and* $total(D_i)$ *returns the total number of links between* $D_i$ *and any other dataset in* $\mathcal{D}$*.*

Note that (4.1) is the symmetric version of the measure of connectivity of $D_i$ to $D_j$ given by

$$\mathcal{C}'(D_i, D_j) = \frac{shared(D_i, D_j)}{total(D_i)}.$$

Explicitly, (4.1) is obtained by taking the mean

$$\frac{\mathcal{C}'(D_i, D_j) + \mathcal{C}'(D_j, D_i)}{2} = \mathcal{C}(D_i, D_j).$$

The measure $\mathcal{C}$ is in the interval $[0, 1]$ and has the advantage of considering the relative connectivity between datasets instead of simply looking at the number of links. In our experimental setting, $shared(D_i, D_j)$ is taken as the sum of the links between two datasets in both directions: $D_i \rightarrow D_j$ and $D_j \rightarrow D_i$, resulting in the number of incoming and outgoing links between the datasets. A specific version of the measure $C$ can be defined by taking only certain types of links (or predicates) in consideration (in our application scenario, we have considered LOD datasets, therefore an example of a specific predicate can be `owl:sameAs`).

In a more general manner, it is possible to use any dataset connectivity measure of our choice. The measure given above is one that worked well in our experiments (see Section 4.5). In addition, one can define in a broader sense a measure of dataset relatedness incorporating semantic elements such as vocabulary and keywords overlap. Dataset complementarity can be of interest in certain scenarios, as well. However, in the current study we have focused on connectivity only, leaving the other possibilities out for future work.

## 4.2 Topic Modeling as a Preprocessing/Learning Step

In this section, we will describe the preprocessing step of our candidate dataset recommendation system. In order to profile our dataset we adopt on the topic profiling approach of [1] which will be described in Section 4.2.1. Then, in section 4.2.2, we introduce our proposition of the expanding this profile graph over the entire LOD.

### 4.2.1 Dataset Topic Profiling

The [1] approach consists of a processing pipeline that combines suitable techniques for dataset sampling, topic extraction and topic relevance ranking. As shown in Figure. 4.2 ,the main steps of the topic profiles extraction pipeline are the following:

**(i)** First the system extracts metadata from DataHub using the CKAN API. To extract the metadata for datasets part of the LOD-Cloud group in DataHub one can call the WEB-REST service: `http://datahub.io/api/action/group_show?id=lodcloud`. These metadatas are crucial to access LOD datasets, i.e., by SPARQL endpoint or a dump file.

**(ii)** For each LOD dataset, the system extracts resource types and a sample of instances. The system is able to use three techniques of sampling:

1. **random sampling**: resource instances are selected randomly.
2. **weighted sampling**: resource instances that carry more information (higher wight which is computed based on the number of datatype properties used to define a resource) have higher chances to be selected.
3. **resource centrality sampling**: The weight here is the ratio of the number of resource types used to describe a particular resource divided by the total number of types in a dataset. Such a strategy ensures that the selected resources are instances of the most important concepts (the more structured and linked to other concepts).

**(iii)** In this step, the system extracts all the literal values of the extracted sample of resource instances from the previous step. Then, it extracts named entities from this textual information using the NER tool DBpedia Spotlight [100] which results on a set of DBpedia resources. The topics $\mathcal{T}$ are the DBpedia categories instances assigned to the extracted entities through the datatype property **dcterms:subject**. Authors notes that the topics are expanded with related topic instances (via the property **skos:broader** up to two levels (determined experimentally).

Example, for a particular resource instance i.e. ¡`http://data.linkededucation.org/resource/lak/conference/edm2012/paper/21`¿ the system extracts all the literal values and extracts the corresponding entities from DBpedia Spotlight such as: ¡`http://dbpedia.org/resource/Learning`¿, ¡`http://dbpedia.org/resource/Student`¿ then their corresponding categories are extracted: ¡`http://dbpedia.org/resource/Category:Cognition`¿, ¡`http://dbpedia.org/resource/Category:Academia`¿, ¡`http://dbpedia.org/resource/Category:Education`¿, ¡`http://dbpedia.org/resource/Category:Intelligence`¿; etc.

**(iv)** The extracted topics from the previous step are assigned to their corresponding datasets to construct a weighted bipartite graph which we named DTPG as defined in Def. 3. Recall that $w_{ik}$ is the weight of the topic $T_k$ with respect to $D_i$ as given in Def. 3. For this purpose authors used PageRank with Priors [115], K-Step Markov [116] and HITS [117], in a combination with a Normalised Topic Relevance Score, in order to rank topics with respect to their relevance to the LOD datasets.

**(v)** The resulted datasets topic profile graph is represented via the VoID and the Vocabulary of Links (VoL)[3]. This graph is accessible via the following SPARQL endpoint: `http://data-observatory.org/lod-profiles/profile-explorer`.

[3] `http://data.linkededucation.org/vol/index.htm`

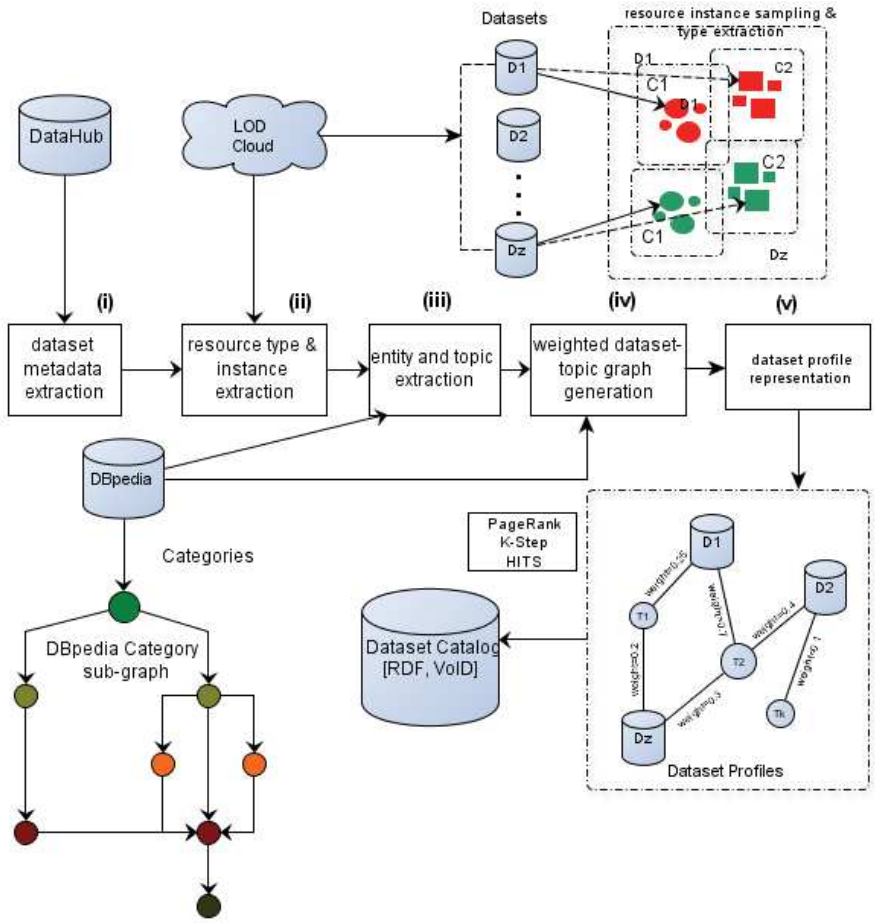


Figure 4.2: Linked Data graphs topic profiles extraction pipeline [1].

The current version of the topic dataset profile graph contains 76 datasets and 185,392 topics. Working with this already annotated subset of existing datasets is not sufficient and would limit the scope of our recommendations significantly. In addition, the number of the profiled datasets, compared to the number of topics is very small, which in turn appeared to be problematic in the recommendation process due to the high degree of topic diversity leading to a lack of discriminability.

One way of approaching this problem would be to index all LOD datasets by applying the original profiling algorithm [1]. However, given the complexity of this processing pipeline—consisting of resource sampling, analysis, entity and topic extraction for a large amount of resources—it is not efficient enough, specifically given the constant evolution of Web data, calling for frequent re-annotation of datasets. In the next section we will propose, one of the original contributions of this dissertation, an efficient method for automatic expansion of the initial profiles index given in [1] over the entire linked open data space based on dataset connectivity measures.

### 4.2.2 Expanding the Profiles Graph over the LOD

Let $\mathcal{DTPG}$ be a topic profile graph and let $D_j \in \mathcal{D}$ be a random dataset, which is not necessarily included in the original topic profile graph $\mathcal{DTPG}$. We assign topic weights to $D_j$ considering its degree of connectivity with respect to datasets from the topic profile graph by using the following measure of relatedness between linked datasets and topics (see Figure. 4.3, steps 1 and 2).

**Definition 5 (Connectivity-based dataset andtopic relatedness measure)** *Let $D_j \in \mathcal{D}$ and $T_k \in \mathcal{T}$. We define the following dataset and topic relatedness measure.*

$$\sigma(D_j, T_k) = \max_{D_i \in \mathcal{D}} \mathcal{C}(D_i, D_j) * w_{ik}. \tag{4.2}$$

Recall that $w_{ik}$ is the weight of the topic $T_k$ with respect to $D_i$ as given in Def. 3, taking a zero value in case $T_k$ is not in Profile($D_i$). $\mathcal{C}(D_i, D_j)$ is the connectivity measure between two datasets, as defined in (4.1). The dataset and topic relatedness measure $\sigma$ is a way to measure the datasets connectivity behavior using their profiles. We will use the notation $\sigma_{jk} = \sigma(D_j, T_k)$ as a shortcut. Note that $\sigma$ is in the $[0, 1]$ interval.

This new weighting scheme allows to propagate inexpensively the profile of $D_i$ to datasets that are connected to it. Hence, a new graph is created between target datasets and source datasets topics. Precisely, a topic $T_k \in$ Profile($D_i$) will be assigned to a dataset $D_j$ that has a non-zero value of $\mathcal{C}(D_i, D_j)$. The weight of this novel topic-dataset relation is now based on the connectivity order of $D_j$ with respect to $D_i$, scaled by the weight $w_{ik}$ of $T_k$ with respect to $D_i$. In that sense, $w_{ik}$ plays a penalization role: the novel weight $\sigma_{jk}$ of $T_k$ with respect to $D_j$ is penalized by the weight of $T_k$ in the original topic graph, i.e., datasets with high degree of connectivity to $D_i$ will get relatively low weights with respect to a topic, if that topic has a relatively low weight with respect to $D_i$. We consider the maximum value over all datasets in $\mathcal{D}$, the set of the originally profiled datasets. In this way, we avoid ambiguity when a non-indexed dataset $D_j$ is connected to a single topic $T_k$ via multiple already indexed datasets, assuring that the highest value of relation between $T_k$ and $D_j$ is preserved. Thus, the choice of a topic to be assigned to a dataset is not influenced, only its weight is, and no connectivity information is lost.

The topic-dataset relatedness measure (5) allows to construct a novel profile graph by computing $\sigma_{jk}$ for all possible values of $j$ and $k$ ($j = 1, ..., M$ and $k = 1, ..., N$). The novel graph, that we call *the Linked Dataset Topic Profile Graph (LDPG)*, includes new datasets and the original topics as its nodes and is defined as follows (see Figure. 4.3, step 2).

**Definition 6 (Linked Dataset Topic Profiles Graph (LDPG))** *The LDPG is a weighted directed bipartite graph $\mathcal{P}_l = (\mathcal{S}_l, \mathcal{E}_l, \Delta_l)$, where $\mathcal{S}_l = \mathcal{D} \cup \mathcal{T}$, $\mathcal{E}_l$ is a set of edges of the form $e'_{jk} = (D_j, T_k)$ such that $D_j \in \mathcal{D}$ and $T_k \in \mathcal{T}$ and*

$$\begin{aligned} \Delta_l &: \mathcal{E}_l \to [0,1] \\ & e'_{jk} \mapsto \sigma_{jk} \end{aligned}$$

*is a function assigning weights to the edges in $\mathcal{E}_l$.*

As this was the case within the original profiling scheme, the inherently bipartite nature of the graph $\mathcal{P}_l$ allows for a two-fold interpretation — either a dataset is modeled as a set of topics (a dataset's *profile*), or, inversely, a topic is modeled as a set of datasets assigned to it (a topic's *signature*). Therefore, it is easy to define a set of **significant** datasets with respect to a given topic, by thresholding on their weights in the Linked profiles graph with respect to the topic of interest. Note again that for the purposes of the recommendation task, we will be interested in keeping the weights of every dataset in the resulting topic representations and thus model every topic by a set of *(dataset, weight)* couples.

**Definition 7 (Dataset significance for a topic. Topic signature)** *A dataset $D_j \in \mathcal{D}$ is* **significant** *with respect to a topic $T_k \in \mathcal{T}$ if its weight in the LDPG $\sigma_{jk} = \sigma(D_j, T_k)$ is greater than a given value $\theta \in (0,1)$.*

*A topic $T_k$ is modeled by the set of its significant datasets together with their respective weights, given as*

$$\mathcal{D}^*_{T_k} = \{(D_j, \sigma_{jk}) | \sigma_{jk} > \theta\}_{j=1,...,M}, \tag{4.3}$$

*for $k = 1, ..., N$. We will call $\mathcal{D}^*_{T_k}$ the* **signature** *of the topic $T_k$.*

With this definition, the profile of a given dataset, Profile($D_i$), is modeled as a number of sets of significant datasets – one per topic in Profile($D_i$) coupled with their weights with respect to each topic (see Figure. 4.3, step 3), or otherwise – a set topic signatures.

For sake of generality, we draw the readers attention to the fact that the learning approach resulting in index (topic model) extension applies to any dataset profile definition that one might like to consider and not exclusively to the one based on the topic modeling paradigm.

## 4.3 Topic Profile-based Framework

Recall that, in line with [109], dataset recommendation is the problem of computing a rank score for each of a set of datasets $\mathcal{D}$ so that the rank score indicates the

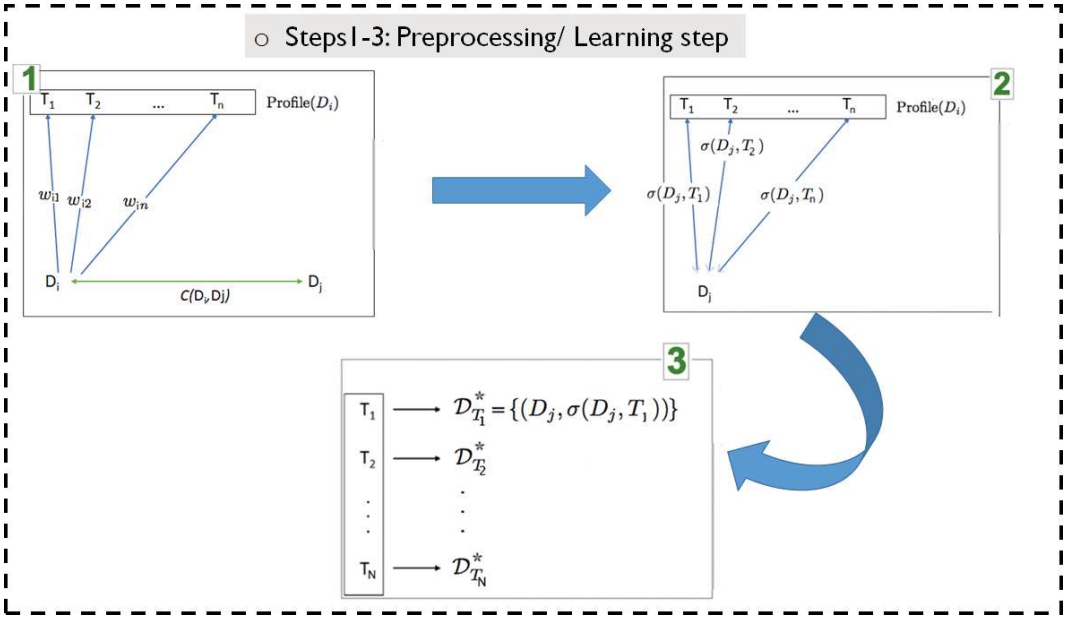


Figure 4.3: The preprocessing steps of the profile-based dataset recommendation framework.

relatedness of a dataset from $\mathcal{D}$ to a given dataset, $D_0$. In turn, this allows to determine the likelihood of datasets in $\mathcal{D}$ to contain linking candidates for $D_0$.

This section will introduce our topic-based recommendation approach which is mainly inspired by the "Collaborative Filtering (CF)" technique – one of the most used and successfully applied methods for personalized recommender systems, a large and continuously active literature exists(see [13]). Basically, the CF approaches tend to recommend for active user the items that similar users– **those with similar tastes**– liked in the past, i.e., some items must have been rated in the past. This is termed "Collaborative-filtering" methods because, we filter objects based on the similarities in behavior using *collaboration* between users or items. The intuition is that the recommendations coming from these similar users should be relevant as well and of interest to the active user because they had a **similar behaviour** in the past.

In the context of LOD dataset recommendation, in order to detects similar tastes of connectivity, we start by grouping datasets by their topics profiles - see Figure. 4.3, steps 1 to 3. Then, we detect the connectivity behaviour of the datasets based on their topics profiles using the existing linksets (connectivity) between datasets, as quantified in Eq. 4.1, to construct a recommendation datasets system (Figure. 4.4, step 4).

In other words, let $D_0$ be an unpublished and non-linked dataset. The aim of the recommendation task is to provide the user with an ordered list of datasets, potential

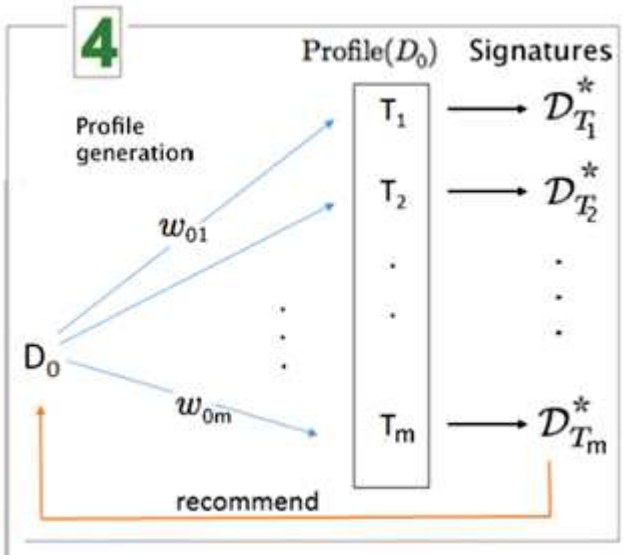


Figure 4.4: The recommendation step of the profile-based dataset recommendation framework.

candidates for interlinking with $D_0$, which narrows down considerably the original search space (i.e., the LOD cloud) and contains little popular and weakly linked datasets. As defined in [109], dataset recommendation is the problem of computing a rank score for each $D_j \in \mathcal{D}$ that indicates the likelihood of $D_j$ to be relevant to a dataset $D_0$. In the context of using topic-based dataset profiles for linking recommendation, we restate the problem in the following manner.

> *For a given non-linked dataset $D_0$, profile-based dataset recommendation is the problem of computing a rank score $r_{0j}$ for each $D_j \in \mathcal{D}$ based on topic overlap between $D_j$ and $D_0$, so that $r_{0j}$ indicates the relevance of $D_j$ to $D_0$ for the interlinking task.*

We start by generating the topic profile of $D_0$, Profile$(D_0) = \{(T_1, w_{01}), ..., (T_m, w_{0m})\}$. As a results of the expansion of the topic profiles over the LOD datasets, we have a set of *significant* LOD datasets for each topic in Profile$(D_0)$ together with their corresponding relevance values $\sigma$, namely the set of $m$ topic signatures $\{\mathcal{D}^*_{T_k}\}_{k=1}^m$. These datasets constitute the pool, from which we will recommend interlinking candidates to $D_0$. We will use $n$ to denote their number, that is $n = \sum_{j=1}^{m} |\mathcal{D}^*_{T_j}|$, or the sum of the numbers of datasets in each topic signature. The aim is to serve the user with the most highly ranked datasets from that pool. There are two ranking criteria to consider: the weight $w$ of each topic in Profile$(D_0)$ and the weight $\sigma$ of each dataset in each of the topic signatures in $\{\mathcal{D}^*_{T_k}\}_{k=1}^m$ (step 4 in Figure. 4.4). Since the ranking score in our setting depends on topic overlap, we define the interlinking relevance of a dataset $D_j$ with respect to $D_0$ in the following manner.

**Definition 8 (Dataset interlinking relevance)** *For all $j = 1, ..., n$, the relevance of a dataset $D_j \in \mathcal{D}$ to a dataset $D_0$ via the topic $T_k$ is given by*

$$r_j^0 = w_{0k} * \sigma_{jk}, \tag{4.4}$$

*with $k = 1, ..., m$.*

Note that $j$ covers the total number of datasets in the set of $m$ topic signatures, therefore the relevance value depends on $j$ only (i.e., a single relevance value per dataset from the pool of candidates). Similarly to the definition of $\sigma$ in Def. 8, $w$ has a penalization function, decreasing the ranks of datasets that have high values of their $\sigma$ weights, but are found in topic signatures of a low relevance to $D_0$ (expressed by a low value of $w$).

It is easy to define a mapping $f : \mathcal{R} \to \mathbb{N}$ from a space of interlinking relevance values $\mathcal{R}$ to the natural numbers such that $f(r_{j_1}^0) > f(r_{j_2}^0) \iff r_{j_1}^0 \leq r_{j_2}^0$, for any $j_1, j_2 \in [1, n]$ and $1 = \max_j f(r_j^0)$. With this definition, since there is a relevance value $r_j^0$ per dataset $D_j \in \mathcal{D}$, $f(r_j^0)$ returns the rank of the dataset $D_j$ with respect to $D_0$. The results of the recommendation process are given in a descending order with respect to these ranks.

## 4.4 Evaluation Framework

The quality of the outcome of a recommendation process can be evaluated along a number of dimensions. Ricci *et al.* [13] provide a large review of recommender system evaluation techniques and cite three common types of experiments: (i) offline setting, where recommendation approaches are compared without user interaction, (ii) user studies, where a small group of subjects experiment with the system and report on the experience, and (iii) online experiments, where real user populations interact with the system.

In our approach, we assume that the dataset connectivity behavior when data were collected (i.e., steps 1, 2 and 3 in Fig. 4.3) is similar enough to the profile connectivity behavior when the recommender system is deployed (i.e., step 4 in Fig. 4.4), so that we can make reliable decisions based on an *offline evaluation*. The offline experiment is performed by using pre-collected data as evaluation data (ED). Using these data, we can simulate the profiles connectivity behavior that impacts the recommendation results.

The most straightforward, although not unproblematic (see the discussion that follows below) choice of ED for the entity linking recommendation task is the existing link topology of the current version of links between web datasets. Since this evaluation data are the only available data that we have for both training (our preprocessing steps 1, 2 and 3 in Fig. 4.3) and testing (the actual recommendation

in step 4 of Figure. 4.4), we opted for a *5-fold cross-validation* [118] to evaluate the effectiveness of the recommendation system. As shown in Figure. 4.5, in 5-fold cross-validation, the ED was randomly split into two subsets: the first one, containing random 80% of the linked datasets in the ED, was used as training set while the second one, containing the remaining linked datasets (i.e., random 20% of the ED), was retained as the validation data for tests (i.e., the test set). We repeated these experiments five times changing at each time the 20% representing the test set in order to cover 100% of the whole data space. The evaluation is based on the capacity of our system to *reconstruct* the links from the ED in the recommendation process.

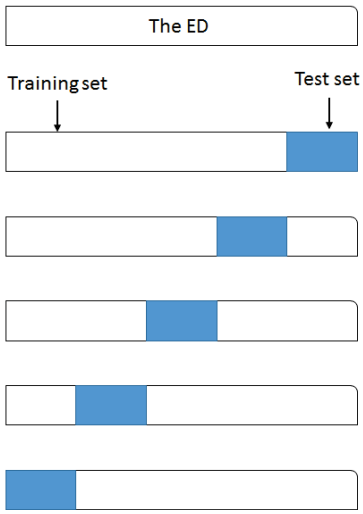


Figure 4.5: The 5-fold Cross-Validation.

Among the 5-fold cross-validation we evaluated the efficiency of our system with the Leave-one-out approach using most common measures to evaluate recommender systems. These measures are formalized as functions of the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) as follows.

**Precision:**

$$Pr = \frac{TP}{TP + FP}. \tag{4.5}$$

**Recall:**

$$Re = \frac{TP}{TP + FN}. \tag{4.6}$$

**F1-Score:**

$$F1 = \frac{2TP}{2TP + FN + FP}. \tag{4.7}$$

In addition, [13] present a measure of the false positive overestimation, particularly important in our offline evaluation case:

$$FalsePositiveRate = \frac{FP}{FP + TN}. \tag{4.8}$$

## 4.5 Experiments and Results

In this section, we start by a discussion on the evaluation setting then we proceed to report on the experiments conducted in support of the proposed recommendation method.

### 4.5.1 Experimental Setup

As defined in Section 4.1, we will consider two sets of datasets:

- $\mathcal{D}'$: All the datasets indexed by the topics profiles graph[4], which will be considered as source datasets (to be linked) in the testing set.
- $\mathcal{D}$: All datasets in the LOD cloud group on the Data Hub[5], which will be considered as target datasets (to be recommended) in the testing set.

We trained our system as described in steps 1, 2 and 3 in Figure. 4.3. We started by extracting the topic profiles graph from the available endpoint of Data Observatory[6]. Then we extracted VoID descriptions of all LOD datasets, using the *datahub2void* tool[7]. The constituted evaluation data corresponds to the outgoing and incoming links extracted from the generated VoID file (it is made available on `http://www.lirmm.fr/benellefi/void.ttl`).

Note that in the training set we used the actual values of `VoID:triples` (see Section 4.2) to compute dataset connectivity, while in the test set we considered binary values (two datasets from the evaluation data are either linked or not). For example, $shared(tip, linkedgeodata) = 6$, so in the training set we considered 6 as the shared triples value in Eq. (4.1), while in the test set we only consider the information that *tip* is connected to *linkedgeodata* and vice versa. Training is performed only once. The experiments have been executed on a single machine with 32GB RAM and processor $CPU@2.8Ghz$, Intel(R) core(TM) i7-4900MQ.

[4] http://data-observatory.org/lod-profiles/profile-explorer/
[5] http://datahub.io/group/lodcloud
[6] `http://data-observatory.org/lod-profiles/sparql`
[7] `https://github.com/lod-cloud/datahub2void`

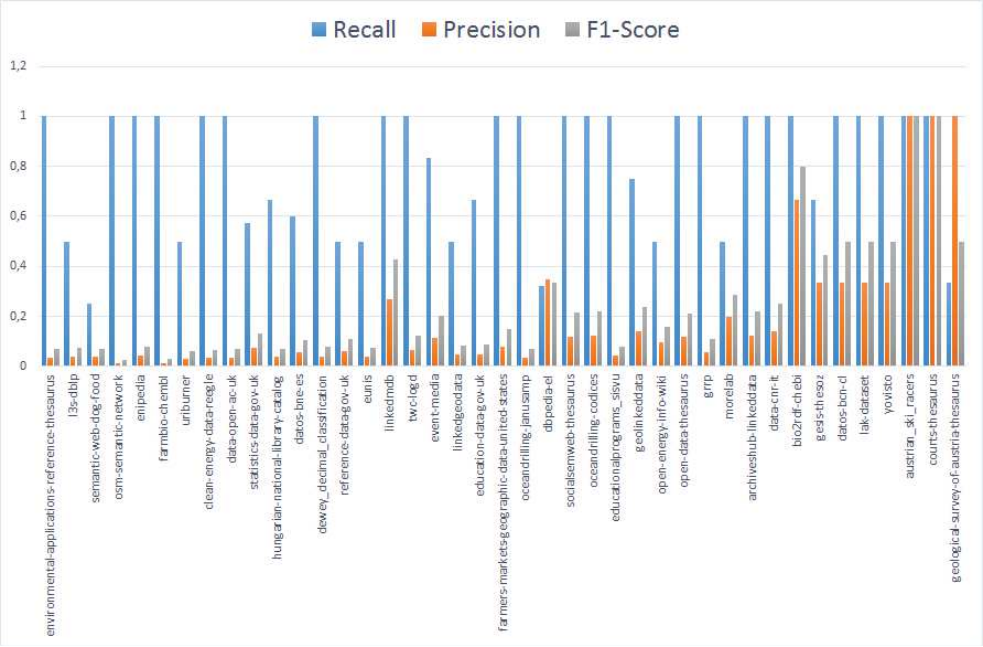


Figure 4.6: Recall/Precision/F1-Score over all recommendation lists for all source datasets in $\mathcal{D}'$ and all target datasets in $\mathcal{D}$.

### 4.5.2 Results and Analysis

We ran our recommendation workflow as described in step 4 in Fig. 4.4. Using 5-fold cross-validation, for each dataset in $\mathcal{D}'$, we recommended an ordered list of datasets from $\mathcal{D}$. The results are given in Fig.4.6.

The results show a high average recall of up to 81%. Note that the recommendation results for 59% of the source datasets have a recall of 100% and two of them have an F1-score of 100%. As mentioned in Section 4.5.1, we considered only the binary information of the existence of a link in the LOD as evaluation data in the testing set. This simplification has been adopted due to the difficulty of retrieving all actual links in the LOD graph (implying the application of heavy instance matching or data linking algorithms on a very large scale). Certainly, the explicit currently existing links are only a useful measure for recall, but not for precision. In our experiments, we measured an average precision of 19%. We explain that by the fact that the amount of explicitly declared links in the LOD cloud as ED **is certain but far from being complete to be considered as ground truth**. Subsequently, we are forced to assume that the false positive items would have not been used even if they had been recommended, i.e., that they are uninteresting or useless to the user. For this reason, based on our evaluation data, a large amount of false positives occur, which in reality are likely to be relevant recommendations. In order to rate this error, we calculated the false positive rate over all results, shown in the Fig.

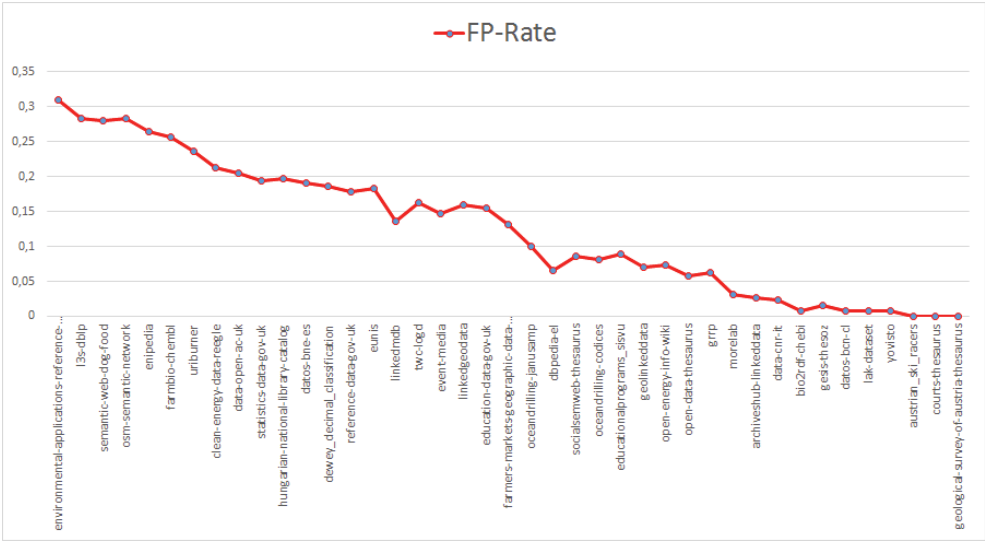


Figure 4.7: False Positive Rates over all recommendation lists over all $\mathcal{D}'$ datasets.

4.7. The small values of this rate indicate that every time you call a positive, you have a probability of being right, which provide support to our hypothesis with an average FP-Rate of 13%.

To further illustrate the effect of false positives overestimation, we included in the ED new dataset links based on the shared keywords of the datasets. Precisely, if two datasets share more then 80% of their VoID tags, they are considered as linked, and are added to the ED. For example, *linkedgeodata* is connected to 4 datasets in the main ED: *osm-semantic-network, dbpedia, tip et dbpedia-el.* However, we found that *linkedgeodata* shared more than 80% of its tags with *fu-berlin-eurostat* and *twarql*[8]. By adding both links to the original ED, we noted a gain in precision of 5% for the *linkedgeodata* dataset with no impact on recall. Thus, we believe that our system can perform much better on more complete ED.

The main goal of a recommender system is to reduce the cost of identifying candidate datasets for the interlinking task. Some systems may provide recommendations with high quality in terms of both precision and recall, but only over a small portion of datasets (as is the case in [108]). We obtain high recall values for the majority of datasets over *the entire set of LOD datasets* with a price to pay of having relatively low precision. Here, low precision/high recall systems still offer significant contributions by lowering the search space. Therefore, we highlight the efficiency of our system in reducing the search space size. Figure. 4.8 depicts the reduction of the original search space size (258 datasets) in percentage over all source datasets to be linked. The average space size reduction is of up to 86%.

[8] Example: *linkedgeodata* has 11 tags and *twarql* has 9 tags. We considered as connected since they shared 8 tags which is higher than the 80% of the average amount, i.e., $8 < (0.8 * (11 + 9)/2)$.

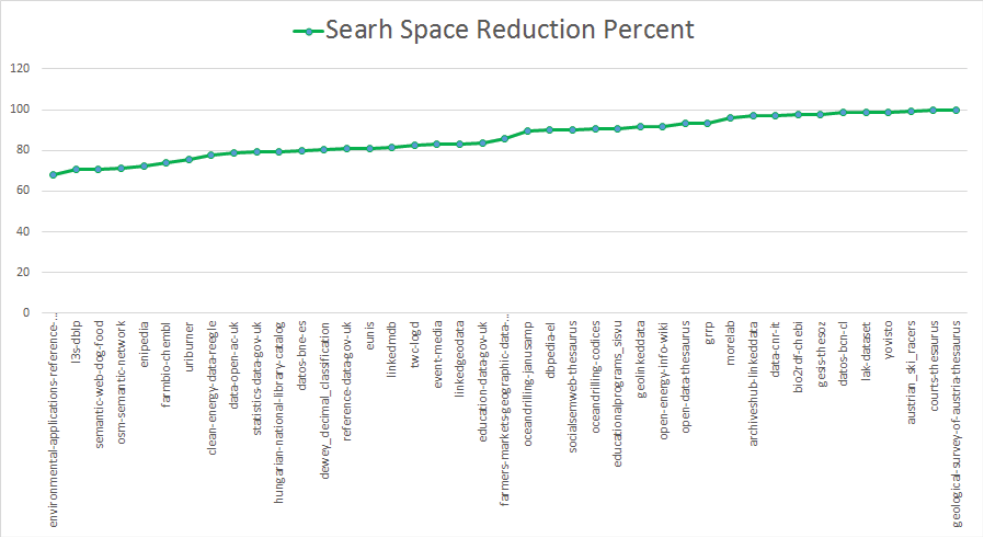


Figure 4.8: Search space reduction in percent over all recommended sets and over all $\mathcal{D}'$ datasets.

As mentioned previously, our system can cover 100% of the available linked datasets, since the topics-datasets profiling approach [1] as well as our profile expansion approach presented in Section 4.2 are able to profile any arbitrary dataset. Our system is also capable of dealing with the well-known cold-start problem (handling correctly newly published and unlinked datasets), since any new dataset can be indexed by the topic profiles graph $\mathcal{P}$ or by the LOD profiles graph $\mathcal{P}_l$ and fed into the recommendation pipeline.

### 4.5.3 Baselines and Comparison

To the best of our knowledge, there does not exist a common benchmark for dataset interlinking recommendation. One of the contributions of this dissertation is the provision of three simple baseline approaches for this problem. Given two datasets, $D_0$ and $D_j$, we define the following baseline recommendation methods.

**Shared Keywords Recommendation:** if $D_0$ and $D_j$ share $N_{tags}$ of VoID:Tags extracted from `http://www.lirmm.fr/benellefi/void.ttl` with $N_{tags} > 0$, then we recommend $(D_j, N_{tags})$ to $D_i$, where $N_{tags}$ acts as a rank score.

**Shared Links Recommendation:** if $D_0$ and $D_j$ have $N_{links}$ connected datasets in common from `http://www.lirmm.fr/benellefi/void.ttl` with $N_{linksets} > 0$, then we recommend $(D_j, N_{links})$ to $D_0$, where $N_{links}$ acts as a rank score.

**Shared Topics Recommendation:** if $D_0$ and $D_j$ share $N_{topics}$ topics extracted from `http://data-observatory.org/lod-profiles/sparql` with $N_{topics} >$

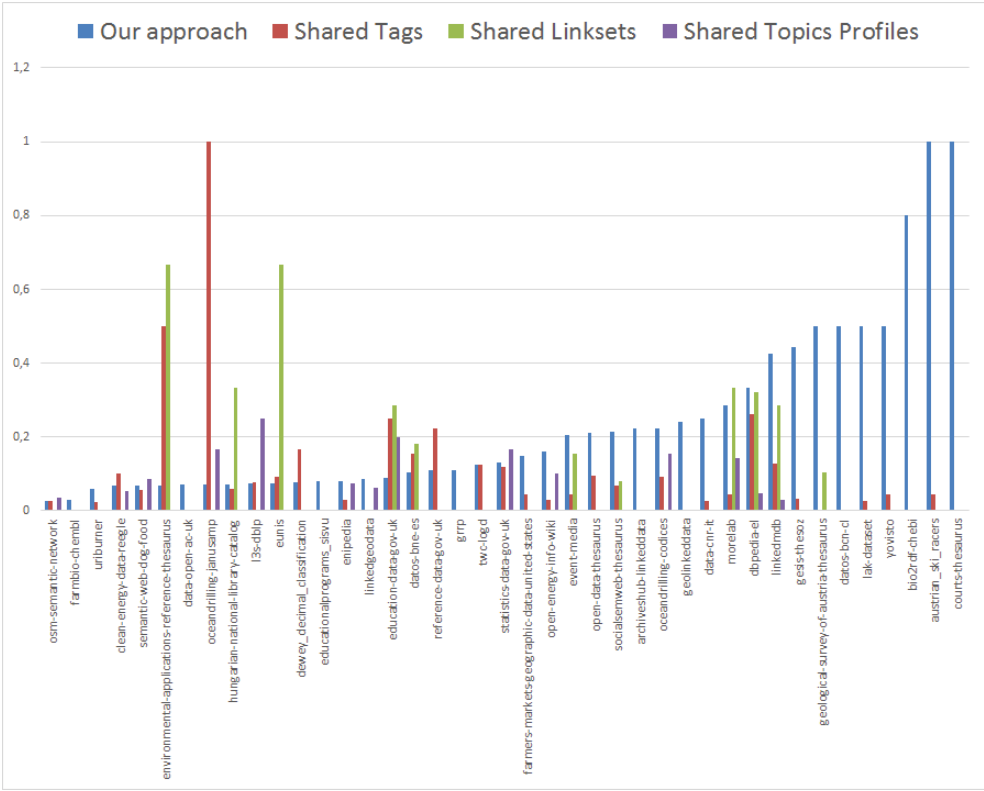


Figure 4.9: F1-Score values of our approach versus the baselines overall $\mathcal{D}'$ datasets

0, then we recommend $(D_j, N_{topics})$ to $D_0$, where $N_{topics}$ acts as a rank score.

The recommendation results for all LOD datasets ($D_0$ covering $D$) of the three baseline approaches are made available on `http://www.lirmm.fr/benellefi/Baselines.rar`.

Figure. 4.9 and Figure. 4.10, respectively, depict detailed comparisons of the F1-Score and the Recall values between our approach and the baselines over all $\mathcal{D}'$ datasets taken as source datasets. From these figures, it can be seen that our method largely outperforms the baseline approaches, which even fail to provide any results at all for some datasets. The baseline approaches have produced better results than our system in a limited number of cases, especially for source and target datasets having the same publisher. For example, the shared keywords baseline generated an F-Score of 100% on *oceandrilling-janusamp*, which is connected to *oceandrilling-codices*, due to the fact that these two datasets are tagged by the same provenance (`data.oceandrilling.org`).

Table 5.1 compares the performance of our approach to the three baseline methods in terms of average precision, recall and F1-score.

As a general conclusion, these observations indicate that the collaborative filtering-

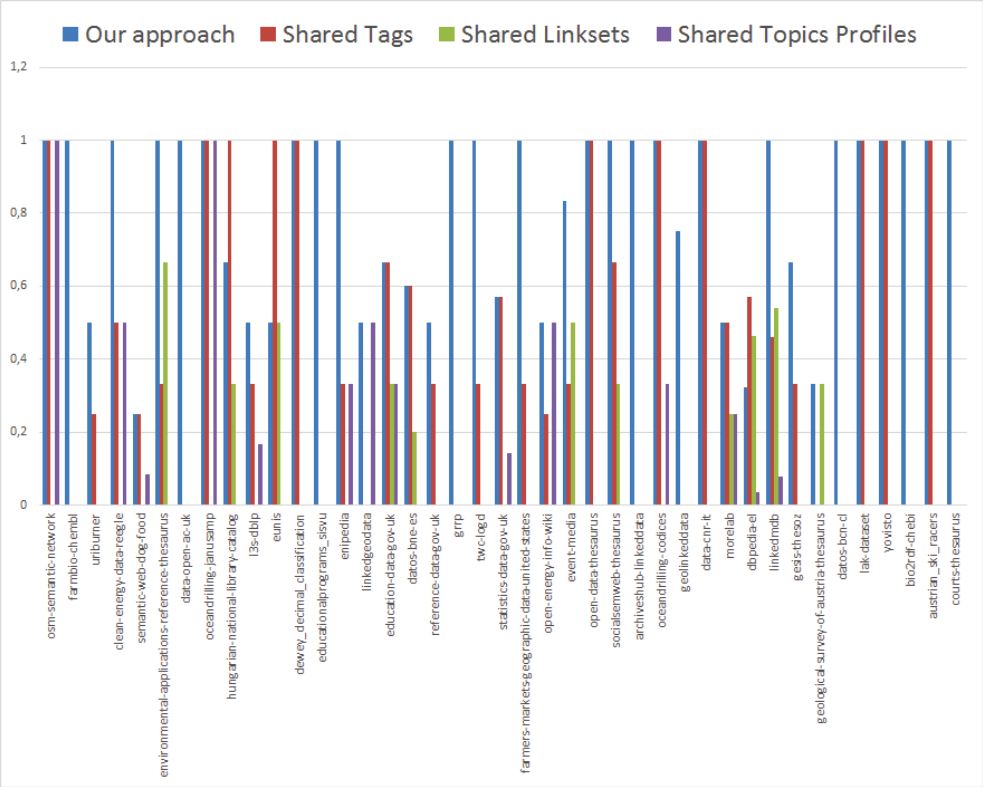


Figure 4.10: Recall values of our approach versus the baselines overall $\mathcal{D}'$ datasets

| | Our approach | Shared Keywords | Shared linksets | Shared Topics Profiles |
|---|---|---|---|---|
| AVG Precision | **19%** | 9% | 9% | 3% |
| AVG Recall | **81%** | 47% | 11% | 13% |
| AVG F1-Score | **24%** | 10% | 8% | 4% |

Table 4.1: Average precision, recall and F1-score of our system versus the baselines over all $\mathcal{D}'$ datasets based on the ED.

like recommendation approach, which exploits both existing dataset profiles as well as traditional dataset connectivity measures, show a high performance on identifying candidate datasets for the interlinking task.

## 4.6 Conclusion

In this chapter, we have presented an interlinking candidate dataset recommendation approach based on topic-profiles. We demonstrated the effectiveness of our approach in term of common evaluation measures for recommender systems and specially on

the search space reduction metric. Furthermore, we have shown that this approach outperforms the three baseline which have been developed for the purposes of this study and made available to the community. An additional contribution through this chapter was by providing a new technique for dataset profiles propagation in order to index the entire LOD datasets, starting off with a limited number of profiled datasets.

This collaborative filtering-like recommendation approach gets its full performance from learning the connectivity behaviour of the existing linksets between LOD datasets. However, the amount of explicitly declared links in the LOD cloud, has led to a weak considered learning data, and consequently a precision of 19% that needs improvements. To this end, the next chapter will introduce our new dataset recommendation approach that adopts a new profile features and breaks away from the learning step.

# Chapter 5

# Dataset Recommendation for Data Linking based on Intensional Profiling

## Contents

# Introduction

We remind that the overarching aim of this dissertation is to provide an efficient candidate dataset recommendation approach in order to change the fact of only few LOD datasets are reused and linked while a large amount of datasets is ignored.

For this purpose, as argued in the previous chapter, we have introduced our topic-based dataset recommendation approach that learns the connectivity behaviour of the LOD datasets based on their topic profiles. Nevertheless, we demonstrate that the incompleteness of the LOD –in its current version of links– made it "to poor" from providing a high quality learning data for our recommendation system. To deal with the latter issue, we introduce in this chapter a new candidate dataset recommendation approach which skips the learning step and adopts the notion of *intensional profile* – a set of schema concept descriptions representing the dataset. Furthermore, each concept is mapped to larger text document which provides richer contextual and semantic information for better intensional representation. In line with our datasets profiles features taxonomy (*cf.* Fig.2.1), the adopted intensional profile can be fitted in the semantic characteristics part, as described in Section 2.1.1.

For better understanding of the intuition behind the proposed approach, we present our working hypothesis as follows: datasets that share at least one pair of semantically similar concepts, are likely to contain at least one pair of instances to be linked by a "owl:sameAs" statement. We base our recommendation procedure on this hypothesis and propose an approach in two steps: (1) for every source dataset $D_S$, we identify a cluster [1] of datasets that share schema concepts with $D_S$ and (2) we rank the datasets in each cluster with respect to their relevance to $D_S$.

In step (1), we identify concept labels that are semantically similar by using a similarity measure based on the frequency of term co-occurence in a large corpus (the web) combined with a semantic distance based on WordNet without relying on string matching techniques [119]. For example, this allows to recommend to a dataset annotated by "school" one annotated by "college". In this way, we form clusters of "comparable datasets" for each source dataset. The intuition is that for a given source dataset, any of the datasets in its cluster is a potential target dataset for interlinking.

Step (2) focuses on ranking the datasets in a $D_S$-cluster with respect to their importance to $D_S$. This allows to evaluate the results in a more meaningful way and of course to provide quality results to the user. The ranking criterium should not be based on the amount of schema overlap, because potential to-link instances can be found in datasets sharing 1 class or sharing 100. Therefore, we need a similarity

[1] We note that we use the term "cluster" in its general meaning, referring to a set of datasets grouped together by their similarity and not in a machine learning sense.

measure on the profiles of the comparable datasets. We have proceeded by building a vector model for the document representations of the profiles and computing cosine similarities.

To evaluate the approach, we have used the current topology of the LOD as evaluation data (ED). As mentioned in the beginning, the LOD link graph is far from being complete, which complicates the interpretation of the obtained results—many false positives are in fact missing positives (missing links) from the evaluation data—a problem that we discussed in Section 4.4 and we will confirm in the sequel. Note that as a result of the recommendation process, the user is not only given candidate datasets for linking, but also pairs of classes where to look for identical instances. This is an important advantage allowing to run more easily linking systems like SILK [103] in order to verify the quality of the recommendation and perform the actual linking.

To sum up, this chapter contains the following contributions: (1) a new definition of dataset profile based on schema concepts, (2) a recommendation framework allowing to identify the datasets sharing schema with a given source dataset, (3) an efficient ranking criterium for these datasets, (4) an output of additional metadata such as pairs of similar concepts across source and target datasets, (5) a large range of reproducible experiments and in depth analysis with all of our results made available.

We start by describing in Section 5.1 the profiles features that we adopt in our approach. Then, we proceed to present the theoretical grounds of our technique in Section 5.2. Section 5.3 defines the evaluation framework that has been established and reports on our experimental results. A discussion about the results is presented in Section 5.4. Then we discuss the positioning of our approach with respect to related works in Section 5.5. Finally, we conclude in Section 5.6. This chapter is based on [120]

## 5.1 Dataset Intensional Profiling

Recall that dataset profiles can be seen as a set of dataset characteristics that allow to describe in the best possible way a dataset and that separate it maximally from other datasets. A feature-based representation of this kind allows to compute distances or measure similarities between datasets (or for that matter profiles), which unlocks the dataset recommendation procedure. These descriptive characteristics, or features, can be of various kinds (statistical, semantic, extensional, etc.). Recall that, as introduced in Section 2.1.1, a dataset profile can be defined by a set of types (schema concepts) names that represent the topic of the data and the covered domain. In line with that definition, we are interested here in intensional dataset characteristics in the form of a set of keywords together with their definitions that best describe a dataset.

**Definition 9 (Dataset Label Profile)** *The* **label profile** *of a dataset $D$, denoted by $\mathcal{P}_l(D)$, is defined as the set of $n$ schema concept labels corresponding to $D$: $\mathcal{P}_l(D) = \{L_i\}_{i=1}^{n}$.*

Note that the representativity of the labels in $\mathcal{P}_l(D)$ with respect to $D$ can be improved by filtering out certain types. We rely on two main heuristics: (1) remove too popular types (such as foaf:Person), (2) remove types with too few instances in a dataset. These two heuristics are based on the intuition that the probability of finding identical instances of very popular or underpopulated classes is low. We support (1) experimentally in Section 5.3 while we leave (2) for future work.

Each of the concept labels in $\mathcal{P}_l(D)$ can be mapped to a text document consisting of the label itself and a textual description of this label. This textual description can be the definition of the concept in its ontology, or any other external textual description of the terms composing the concept label. We define a document profile of a dataset in the following way.

**Definition 10 (Dataset Document Profile)** *The* **document profile** *of a dataset $D$, $\mathcal{P}_d(D)$, is defined as a text document constructed by the concatenation of the labels in $\mathcal{P}_l(D)$ and the textual descriptions of the labels in $\mathcal{P}_l(D)$.*

Note that there is no substantial difference between the two definitions given above. The document profile is an extended label profile, where more terms, coming from the label descriptions, are included. This allows to project the profile similarity problem onto a vector space by indexing the documents and using a term weighting scheme of some kind (e.g., *tf*idf*).

By the help of these two definitions, a profile can be constructed for any given dataset in a simple and inexpensive way, independent on its connectivity properties on the LOD. In other words, a profile can be easily computed for datasets that are already published and linked, just as for datasets that are to be published and linked, allowing to use the same representation for both kinds of datasets and thus allowing for their comparison by the help of feature-based similarity measures.

As stated in the introduction of this chapter, we rely on the simple intuition that datasets with similar intension have extensional overlap. Therefore, it suffices to identify at least one pair of semantically similar types in the schema of two datasets in order to select these datasets as potential linking candidates. We are interested in the semantic similarity of concept labels in the dataset label profiles. There are many off-the-shelf similarity measures that can be applied, known from the ontology matching literature. We have focused on the well known semantic measures Wu Palmer [121] and Lin's [122], as well as the UMBC [119] measure that combines semantic distance in WordNet with frequency of occurrence and co-occurrence of terms in a large external corpus (the web). In the following, we depict the definitions of the adopted measures with respect to their definitions in their respective references papers.

For two labels, $x$ and $y$, we have:

$$sim_{\text{WUP}}(x, y) = \frac{2 \times \text{depth(LCS)}}{\text{depth}(x) + \text{depth}(y)} \tag{5.1}$$

$$sim_{\text{LIN}}(x, y) = \frac{2 \times \text{IC(LCS)}}{\text{IC}(x) + \text{IC}(y)} \tag{5.2}$$

where "LCS" is the least common subsumer (most informative subsumer) and "IC" is the information content.

$$sim_{\text{UMBC}}(x, y) = sim_{\text{LSA}}(x, y) + 0.5e^{-\alpha D(x,y)}, \tag{5.3}$$

where $sim_{\text{LSA}}(x, y)$ is the Latent Semantic Analysis (LSA) [114] word similarity, which relies on the words co-occurrence in the same contexts computed in a three billion words corpus[2] of good quality English. $D(x, y)$ is the minimal WordNet [123] path length between $x$ and $y$. According to [119], using $e^{-\alpha D(x,y)}$ to transform simple shortest path length has shown to be very efficient when the parameter $\alpha$ is set to 0.25.

With a concept label similarity measure at hand, we introduce the notion of dataset comparability, based on the existence of shared intension.

**Definition 11 (Comparable Datasets)** *Two datasets $D'$ and $D"$ are comparable if there exists $L_i$ and $L_j$ such that $L_i \in \mathcal{P}_l(D')$, $L_j \in \mathcal{P}_l(D")$ and $sim_{UMBC}(L_i, L_j) \geq \theta$, where $\theta \in [0, 1]$.*

## 5.2 Recommendation Process: the CCD-CosineRank Approach

A dataset recommendation procedure for the linking task returns, for a given source dataset, a set of target datasets ordered by their likelihood to contain instances identical to those in the source dataset. In the following we will detail each step in the CCD-CosineRank workflow as depicted in Figure 5.1.

**Preprocessing step** – This phase consists in representing each dataset by its descriptive profile: its label profile (*cf.* Definition 9) and its document profile (*cf.* Definition 10).

**Target Datasets Filtering step** – Let $D_S$ be a source dataset. We introduce the notion of a *cluster of comparable datasets* related to $D_S$, or $CCD(D_S)$ for short, defined as the set of target datasets, denoted by $D_T$, that are comparable to

[2] `http://ebiquity.umbc.edu/resource/html/id/351`

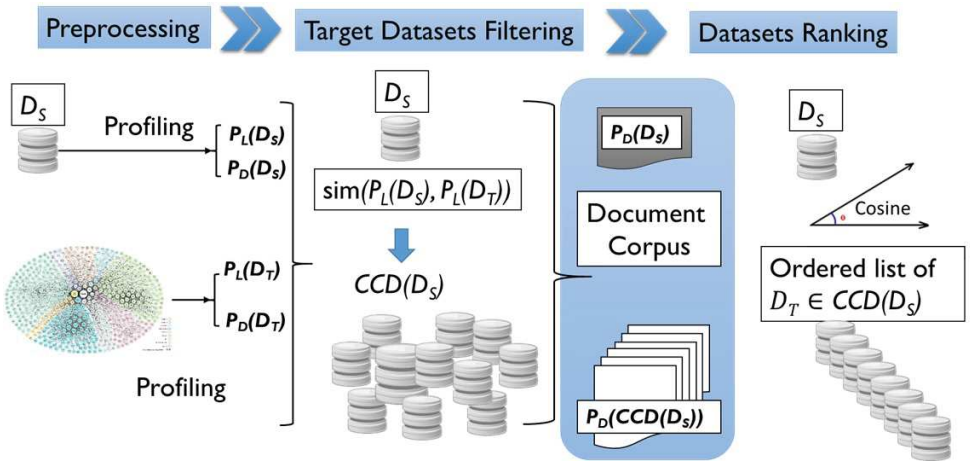


Figure 5.1: Recommendation Process: the CCD-CosineRank Workflow

$D_S$ according to Def. 11. Thus, $D_S$ is identified by its $CCD$ and all the linking candidates $D_T$ for this dataset are found in its cluster, following our working hypothesis. Hence, this step consists of limiting the search space of potential target datasets to $CCD$.

**Datasets ranking step** – This step involves a ranking of the filtered datasets in $CCD(D_S)$ with respect to $D_S$. The ranking score should express the likelihood of a dataset in $CCD(D_S)$ to contain identical instances with those of $D_S$. To this end, we need a similarity measure on the dataset profiles. Since datasets are represented as text documents profile (*cf.* Def. 10), we can easily build a vector model by indexing the documents in the corpus formed by all datasets of interest – the ones contained in one single $CCD$. We use a *tf*idf* weighting scheme, which allows to compute the cosine similarity between the document vectors and thus assign a ranking score to the datasets in a $CCD$ with respect to a given dataset from the same $CCD$. Note that this approach allows to consider the information of the intensional overlap between datasets prior to ranking and indexing – we are certain to work only with potential linking candidates when we rank, which improves the quality of the ranks. For a given dataset $D_S$, the procedure returns datasets from $CCD(D_S)$, ordered by their cosine similarity to $D_S$.

Finally, an important outcome of the recommendation procedure is the fact that, along with an ordered list of linking candidates, the user is provided the pairs of schema types of two datasets—a source and a target—where to look for identical instances. This information facilitates considerably the linking process, to be performed by an instance matching tool, such as SILK.

#### 5.2.0.1 Application of the Approach: an Example

We illustrate our approach by an example. We consider *education-data-gov-uk*[3] as a source dataset ($D_S$). The first step consists in retrieving the schema concepts from this dataset and constructing a clean label profile (we filter out noisy labels, as discussed above), as well as its corresponding document profile (Def. 9 and Def. 10, respectively). We have $\mathcal{P}_l$(*education-data-gov-uk*) = {London Borough Ward, School, Local Learning Skills Council, Adress}. We perform a semantic comparison between the labels in $\mathcal{P}_l$(*education-data-gov-uk*) and all labels in the profiles of the accessible LOD datasets. By fixing $\theta = 0.7$, we generate $CCD$(*education-data-gov-uk*) containing the set of comparable datasets $D_T$, as described in Def. 11. The second step consists of ranking the $D_T$ datasets in $CCD$(*education-data-gov-uk*) by computing the cosine similarity between their document profiles and $\mathcal{P}_d$(*education-data-gov-uk*). The top 5 ranked candidate datasets to be linked with *education-data-gov-uk* are (1) *rkb-explorer-courseware*[4], (2) *rkb-explorer- courseware*[5], (3) *rkb-explorer-southampton*[6], (4) *rkb-explorer-darmstadt*[7], and (5) *oxpoints*[8].

Finally, for each of these datasets, we retrieve the pairs of shared (similar) schema concepts extracted in the comparison part:

– *education-data-gov-uk* and *statistics-data-gov-uk* share two labels "London Borough Ward" and "LocalLearningSkillsCouncil".

– *education-data-gov-uk* and *oxpoints* contain similar labels which are, respectively, "School" and "College", as described in the SILK results (*cf.* Section **??**).

## 5.3 Experiments and Results

This section will present different experiments conducted for the CCD-CosineRank approach evaluation. We start by describing the evaluation framework followed by a discussion on the setting of our experiments. Then, we proceed to report on the experiments conducted in support of the proposed recommendation approach.

[3] `http://education.data.gov.uk/`
[4] `http://courseware.rkbexplorer.com/`
[5] `http://courseware.rkbexplorer.com/`
[6] `http://southampton.rkbexplorer.com/`
[7] `http://darmstadt.rkbexplorer.com/`
[8] `https://data.ox.ac.uk/sparql/`

### 5.3.1 Evaluation Framework

In line with Section 4.4, to the best of our knowledge, there does not exist a common evaluation framework for candidate dataset recommendation, hence, we evaluate our approach with an offline experiment with respect to the current state of LOD considered as evaluation data, since it is the only connected available graph in the time of our experiments.

We recall that in the proposed recommendation approach (*cf.* Section 5.2), for a given source dataset $D_S$, we identify a cluster of target datasets $D_T$, ranked with respect to $D_S$. Henceforth, to evaluate the quality of the recommendation results with respect to the ED, we adopt the most common evaluation measures for recommender systems, namely precision and recall as described in Section 4.4.

Further evaluation for recommender system effectiveness can be done by rating potentially relevant results with respect to their ranks. In other words, we evaluate the precision of our system at given rank $k$ denoted by $P@k$ – the number of relevant results in a result set of size $k$. Additionally, we evaluate the precision of our recommendation when the level of recall is 100% by using the mean average precision at $Recall = 1$, MAP@R, given as:

$$MAP@R = \frac{\sum_{q=1}^{\text{Total}_{D_S}} Pr@R(q)}{\text{Total}_{D_S}}, \tag{5.4}$$

where $R(q)$ corresponds to the rank, at which recall reaches 1 for the $q$th dataset and $\text{Total}_{D_S}$ is the entire number of source datasets in the evaluation.

It should be noted that this evaluation framework differs from the evaluation described in the Section 4.4, notably with the adoption of the *5-fold cross-validation* for the topic profiles-based recommendation approach in order to use the LOD as evaluation data and learning data.

### 5.3.2 Experimental Setup

We started by crawling all available datasets in the LOD cloud group on the Data Hub[9] in order to extract their profiles. In this crawl, only 90 datasets were accessible via endpoints or via dump files. In the first place, for each accessible dataset, we extracted its implicit and explicit schema concepts and their labels, as described in Def. 9. The explicit schema concepts are provided by resource types, while the implicit schema concepts are provided by the definitions of a resource properties [124]. As noted in Section 5.2, some labels such as “Group”, “Thing”, “Agent”, “Person” are very generic, so they are considered as noisy labels. To address this

[9] `http://datahub.io/group/lodcloud`

problem, we filter out schema concepts described by generic vocabularies such as VoID[10], FOAF[11] and SKOS[12]. The dataset document profiles, as defined in Def. 10, are constructed by extracting the textual descriptions of labels by querying the Linked Open Vocabularies[13] (LOV) with each of the concept labels per dataset.

To form the clusters of comparable datasets from Def. 11, we compute the semantico-frequential similarity between labels (given in eq. (5.3)). We apply this measure via its available web API service[14]. In addition, we tested our system with two more semantic similarity measures based on WordNet: Wu Palmer and Lin's. For this purpose, we used the 2013 version of the $WS4J$[15] java API.

In the same manner as the setups in Section 4.5.1, the ED corresponds to the outgoing and incoming links extracted from the generated VoID file using the *datahub2void* tool[16]. It is made available on `http://www.lirmm.fr/benellefi/void.ttl`. We note that out of 90 accessible datasets, only those that are linked to at least one accessible dataset in the ED are evaluated in the experiments. The experiments have been executed on a single machine with 32GB RAM and processor $CPU@2.8Ghz$, Intel(R) core(TM) i7-4900MQ.

### 5.3.3 Evaluation Results

We started by considering each dataset in the ED as an **unlinked source** or newly published dataset $D_S$. Then, we ran the *CCD-CosineRank* workflow, as described in Section 5.2. The first step is to form a $CCD(D_S)$ for each $D_S$. The $CCD$ construction process depends on the similarity measure on dataset profiles. Thus, we evaluated the $CCD$ clusters in terms of recall for different levels of the threshold $\theta$ (*cf.* Def. 11) for the three similarity measures that we apply. We observed that the recall value remains 100% in the following threshold intervals per similarity measure: **Wu Palmer**: $\theta \in [0, 0.9]$; **Lin**: $\theta \in [0, 0.8]$; **UMBC**: $\theta \in [0, 0.7]$.

The $CCD$ construction step ensures a recall of 100% for various threshold values, which will be used to evaluate the ranking step of our recommendation process by the Mean Average Precision (MAP@R) at the maximal recall level, as defined in Def. 5.4. The results in Fig. 5.2 show highest performance of the UMBC's measure with a $MAP@R \cong 53\%$ for $\theta = 0.7$, while the best MAP@R values for Wu Palmer and Lin's measures are, respectively, 50% for $\theta = 0.9$ and 51% for $\theta = 0.8$. Guided by these observations, we evaluated our ranking in terms of precision at

[10] `http://rdfs.org/ns/void`
[11] `http://xmlns.com/foaf/0.1/`
[12] `http://www.w3.org/2004/02/skos/core`
[13] `http://lov.okfn.org/dataset/lov/`
[14] `http://swoogle.umbc.edu/SimService/`
[15] `https://code.google.com/p/ws4j/`
[16] `https://github.com/lod-cloud/datahub2void`

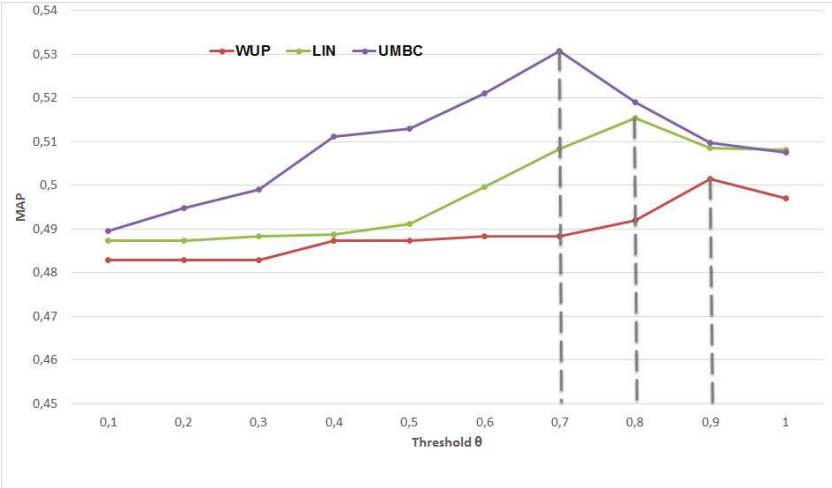


Figure 5.2: The MAP@R of our recommender system by using three different similarity measures for different similarity threshold values

ranks $k = \{5, 10, 15, 20\}$, as shown in Table 5.1. Based on these results, we choose UMBC at a threshold $\theta = 0.7$ as a default setting for *CCD-CosineRank*, since it performs best for three out of four $k$-values and it is more stable than the two others especially with MAP@R.

### 5.3.4 Baselines and Comparison

To the best of our knowledge, there does not exist a common benchmark for dataset interlinking recommendation. Hence, in this section, we implemented two recommendation approaches that are considered more advanced baselines than the ones proposed in section 4.5.3 in term of performance. In details, since our method uses both label profiles and document profiles, the tow proposed baselines are respectively – one using document profiles only, and another one using label profiles:

***Doc-CosineRank*:** All datasets are represented by their *document profiles*, as given in Def. 10. We build a vector model by indexing the documents in the corpus formed by all available LOD datasets (no $CCD$ clusters). We use a *tf*idf* weighting scheme, which allows us to compute the cosine similarity between the document vectors and thus assign a ranking score to each dataset in the entire corpus with respect to a given dataset $D_S$.

***UMBCLabelRank:*** All datasets are represented by their *label profiles*, as given in Def. 9. For a source dataset $D_S$, we construct its $CCD(D_S)$ according to Def. 11 using UMBC with $\theta = 0.7$. Thus, $D_S$ is identified by its $CCD$ and all target datasets $D_T$ are found in its cluster. Let $Avg$UMBC be a ranking

| Measure \ P@k | P@5 | P@10 | P@15 | P@20 |
|---|---|---|---|---|
| WU Palmer ($\theta = 0.9$) | 0, 56 | 0, 52 | 0.53 | 0.51 |
| Lin ($\theta = 0.8$) | 0.57 | **0.54** | **0.55** | 0.51 |
| UMBC ($\theta = 0.7$) | **0.58** | **0.54** | 0.53 | **0.53** |

Table 5.1: Precision at 5, 10, 15 and 20 of the *CCD-CosineRank* approach using three different similarity measures over their best threshold values based on Fig.5.2

function that assigns scores to each $D_T$ in $CCD(DS)$, defined by:

$$Avg\text{UMBC}(D', D\text{“}) = \frac{\sum_{i=1}^{|\mathcal{P}_l(D')|} \sum_{j=1}^{|\mathcal{P}_l(D\text{“})|} \max sim_{UMBC}(L_i, L_j)}{\max(|\mathcal{P}_l(D')|, |\mathcal{P}_l(D\text{“})|)}, \quad (5.5)$$

where $L_i$ in $\mathcal{P}_l(D')$ and $L_j$ in $\mathcal{P}_l(D\text{“})$.

Fig. 5.3 depicts a detailed comparison of the precisions at recall 1 obtained by the three approaches for each $D_S$ taken as source dataset. It can be seen that the *CCD-CosineRank* approach is more stable and largely outperforms the two other approaches by an MAP@R of up to 53% as compared to 39% for *UMBCLabelRank* and 49% for *CCD-CosineRank*. However, the *UMBCLabelRank* approach produces better results than the other ones for a limited number of source datasets, especially in the case when $D_S$ and $D_T$ share a high number of identic labels in their profiles.

The performance of the *CCD-CosineRank* approach demonstrates the efficiency and the complementarity of combining in the same pipeline (i) the *semantic* similarity on labels for identifying recommendation candidates ($CCD$ construction process) and (ii) the *frequential* document cosine similarity to rank the candidate datasets. We make all of the ranking results of the *CCD-CosineRank* approach available to the community on `http://www.lirmm.fr/benellefi/CCD-CosineRank_Result.csv`.

## 5.4 Results Discussion

The aim of this section is to discuss the different results depicted in the previous section. We begin by a note on the vocabulary filtering that we perform (Section. 5.3.2). We underline that we have identified the types which improve/decrease the performance empirically. As expected, vocabularies, which are very generic and wide-spread have a negative impact, acting like hub nodes, which dilute the results. The results of the recommendation before removal are made available on: `http://www.lirmm.fr/benellefi/RankNoFilter.csv`.

The different experiments described above show a high performance of the introduced recommendation approach with an average precision of 53% for a recall of

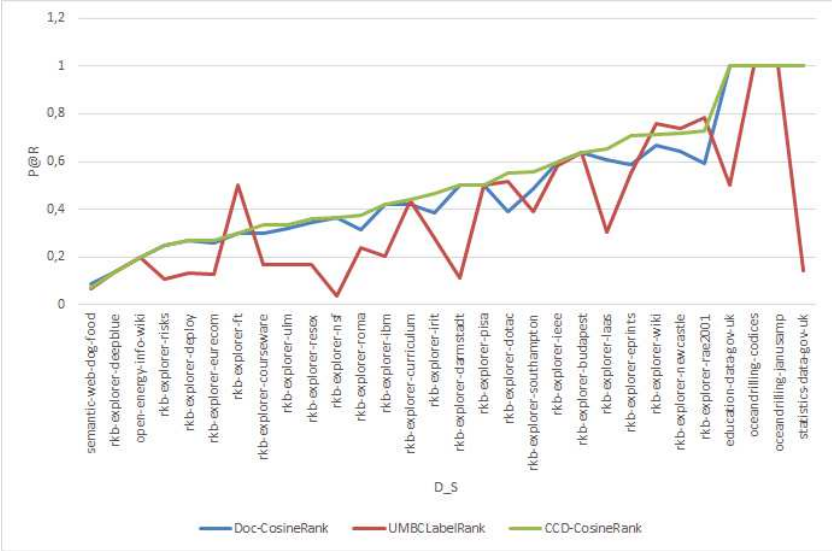


Figure 5.3: Precisions at recall=1 of the *CCD-CosineRank* approach as compared to *Doc-CosineRank* and *UMBCLabelRank*

100%. Likewise, it may be observed that this performance is completely independent of the dataset size (number of triples) or the schema cardinality (number of schema concepts by datasets). However, we note that better performance was obtained for datasets from the *geographic* and *governmental* domains with precision and recall of 100%. Naturally, this is due to the fact that a recommender system in general and particularly our system performs better with datasets having high quality schema description and datasets reusing existing vocabularies (the case for the two domains cited above), which is considered as linked data modeling best practice. Hence, an effort has to be made for improving the quality of the published dataset, as will be further explored in the following chapter.

Furthermore, we believe that our method can be given a more fair evaluation if better evaluation data in the form of ground truth are used. Indeed, our results are impacted by the problem of false positives overestimation. Since data are not collected using the recommender system under evaluation, we are forced to assume that the false positive items would have not been used even if they had been recommended, i.e., that they are uninteresting or useless to the user. This assumption is, however, generally false, for example when the set of unused items contains some interesting items that the user did not select. In our case, we are using declared links in the LOD cloud as ED, which is certain but far from being complete for it to be considered as ground truth. Thus, in the recommendation process the number of false positives tends to be overestimated, or in other words an important number of missing positives in the ED translates into false positives in the recommendation process.

In line with Section 4.5.2 and for further illustration of the effect of false positives overestimation, we ran SILK as an instance matching tool to discover links between $D_S$ and their corresponding $D_T$s that have been considered as false positives in our ED. SILK takes as an input a *Link Specification Language* file, which contains the instance matching configuration. Listing 6.2 depicts an example of LSL file aiming to find equivalent instances corresponding to "dbpedia:Country" and "akt:Country". We recall that our recommendation procedure provides pairs of shared or similar types between $D_S$ and every $D_T$ in its corresponding $CCD$, which are particularly useful to configure SILK. However, all additional information, such as the datatype properties of interest, has to be given manually. This makes the process very time consuming and tedious to perform over the entire LOD. Therefore, as an illustration, we ran the instance matching tool on two flagship examples of false positive $D_T$s:

**Semantically Similar Labels:** We choose *education-data-gov-uk*[17] as a $D_S$ and its corresponding false positive $D_T$ *oxpoints*[18]. The two datasets contain in their profiles, respectively, the labels "School" and "College", detected as highly similar labels by the UMBC measure, with a score of 0.91. The instance matching gave as a result 10 accepted "owl:sameAs" links between the two datasets.

**Identical Labels:** We choose *rkb-explorer-unlocode*[19] as a $D_S$ and its corresponding $D_T$s, which are considered as FP: *yovisto*[20] *datos-bcn-uk*[21] *datos-bcn-cl*[22]. All 4 datasets share the label "Country" in their corresponding profiles. The instance matching process gave as a result a set of accepted "owl:sameAs" links between *rkb-explorer-unlocode* and each of the three $D_T$.

We provide the set of newly discovered linksets to be added to the LOD topology and we made the generated linksets and the corresponding SILK configurations available on `http://www.lirmm.fr/benellefi/Silk_Matching`.

Listing 5.1: An example of SILK LSL file aiming to find equivalent instances between yovisto (`http://sparql.yovisto.com/`) and unlocode (`http://unlocode.rkbexplorer.com/sparql/`) corresponding respectively to the concepts "dbpedia:Country" and "akt:Country"

```
1 <Silk>
2   <Prefixes>
3     <Prefix id="rdf" namespace="http://www.w3.org/1999/02/22-rdf-
          syntax-ns#" />
4     <Prefix id="rdfs" namespace="http://www.w3.org/2000/01/rdf-
          schema#" />
```

[17] `http://education.data.gov.uk/`
[18] `https://data.ox.ac.uk/sparql`
[19] `http://unlocode.rkbexplorer.com/sparql/`
[20] `http://sparql.yovisto.com/`
[21] `http://data.open.ac.uk/query`
[22] `http://data.open.ac.cl/query`

```
5     <Prefix id="dbpedia" namespace="http://dbpedia.org/ontology/" />
6    <Prefix id="os" namespace="http://www.ordnancesurvey.co.uk/ontology/AdministrativeGeography/v2.0/AdministrativeGeography.rdf#"/>
7     <Prefix id="akt" namespace="http://www.aktors.org/ontology/portal#" />
8     <Prefix id="akts" namespace="http://www.aktors.org/ontology/support#" />
9   </Prefixes>
10 <DataSources>
11     <DataSource type="sparqlEndpoint" id="yovisto">
12       <Param name="endpointURI" value="http://sparql.yovisto.com/" />
13     </DataSource>
14     <DataSource type="sparqlEndpoint" id="rkb-explorer-unlocode">
15       <Param name="endpointURI" value="http://unlocode.rkbexplorer.com/sparql/" />
16     </DataSource>
17   </DataSources>
18  <Interlinks>
19     <Interlink id="Country">
20       <LinkType>owl:sameAs</LinkType>
21       <SourceDataset dataSource="yovisto" var="a">
22         <RestrictTo>
23           ?a rdf:type dbpedia:Country
24         </RestrictTo>
25       </SourceDataset>
26       <TargetDataset dataSource="rkb-explorer-unlocode" var="b">
27         <RestrictTo>
28           ?b rdf:type akt:Country
29         </RestrictTo>
30       </TargetDataset>
31       <LinkageRule>
32         <Compare metric="levenshteinDistance" threshold="2" required="true">
33         <TransformInput function="lowerCase">
34             <Input path="?a/rdfs:label" />
35          </TransformInput>
36           <TransformInput function="lowerCase">
37             <Input path="?b/akts:has-pretty-name" />
38          </TransformInput>
39           </Compare>
40         </LinkageRule>
41      <Outputs>
42         <Output type="file" minConfidence="0.9">
43           <Param name="file" value="accepted_links.nt" />
44           <Param name="format" value="ntriples" />
45         </Output>
46         <Output type="file" maxConfidence="0.9">
47           <Param name="file" value="verify_links.nt" />
48           <Param name="format" value="alignment" />
```

```
49            </Output>
50          </Outputs>
51        </Interlink>
52    </Interlinks>
53 </Silk>
```

It should be noted that the recommendation results provided by our approach may contain some broader candidate datasets with respect to the source dataset. For example, two datasets that share schema labels such as books and authors are considered as candidates even when they are from different domains like science vs. literature. This outcome can be useful for predicting links such as "rdfs:seeAlso" (rather than "owl:sameAs"). We have chosen to avoid the inclusion of instance-related information in order to keep the complexity of the system as low as possible and still provide reasonable precision by guaranteeing a 100% recall.

As a conclusion, we outline three directions of work in terms of dataset quality that can considerably facilitate the evaluation of any recommender system in that field: (1) improving descriptions and metadata; (2) improving accessibility; (3) providing a reliable ground truth and benchmark data for evaluation.

## 5.5 Related Works Positioning

The aim of this section is to provide a positioning of our approach with respect to related works and the topic profiles-based recommendation approach.

In line with the studies outlined in Section 3.2.2, none of these works have evaluated the ranking measure in terms of Precision/Recall, except for [109] which, according to authors, achieves a mean average precision of around 60% and an excepted recall of 100%. However, a direct comparison to our approach seems unfair since authors did not provide the set of considered datasets as sources and the corresponding ranking scores or the corresponding target list.

Furthermore, in line with the considered state-of-the-art approaches, we highlight the efficiency of our method in overcoming a series of complexity related problems, precisely, considering the complexity to generate the matching in [107], to produce the set of domain-specific keywords as input in [110] and to explore the set of features of all the network datasets in [109]. Our recommendation results are much easier to obtain since we only manipulate the schema part of the dataset. In other worlds, we wanted to avoid the inclusion of instance-related information in order to keep the complexity of the system as low as possible and still provide reasonable precision by guaranteeing a 100% recall. In details, during the experiments, the execution times of our recommendation per source dataset are around:

- Label profile extraction: $2,09200$ seconds.

- Document profile extraction: $86,17900$ seconds.
- Ranking process: $1,66600$ seconds.

It should be noted that the computation didn't require all resources that are mentioned in Section 5.3.2, but, the same performance could be achieved for a lower resources, i.e., with an $i7$ CPU ($6GB$ RAM).

In line with topic-based recommendation approach (*cf.* Section 4.3), experiment results depict that the CCD-CosineRank obviously outperforms the latter approach in term of the considered evaluation metrics, i.e., a recall up to 100% *vs.* only 81% and respectively for an average precision of up to 53% *vs.* 19%. However, we believe that the low performance of the first approach, compared to the new approach, is due to the weakness of its learning data which confirms and validate our hypothesis. To be fair with our first approach, we believe that a better ground truth may lead to a much richer learning data for the topic-based approach and thus will significantly improve its ranking performance.

## 5.6 Conclusion

Following the linked data best practices, metadata designers reuse and build on, instead of replicating, existing RDF schema and vocabularies. Motivated by this observation, this chapter presented the *CCD-CosineRank* candidate dataset recommendation approach, based on concept label profiles and schema overlap across datasets. Our approach consists of identifying clusters of comparable datasets, then, ranking the datasets in each cluster with respect to a given dataset. We discuss three different similarity measures, by which the relevance of our recommendation can be achieved. We evaluate our approach on real data coming from the LOD cloud and compare it two baseline methods. The results show that our method achieves a mean average precision of around 53% for recall of 100%, which reduces considerably the cost of dataset interlinking. In addition, as a post-processing step, our system returns sets of schema concept mappings between source and target datasets, which decreases considerably the interlinking effort and allows to verify explicitly the quality of the recommendation.

One of the conclusions of our study shows that the recommendation approach is limited by the lack of accessibility, explicit metadata and quality descriptions of the datasets. Hence, in the next chapter, we will deal with linked data modeling task by introducing our new tool that assists metadata designers to ensure a high quality representative datasets schema profiles.

# Chapter 6

# Vocabulary Recommendation for Linked Data Modeling

## Contents

# Introduction

During the last years, the increasing adoption of LOD principles by Web practitioners around the world has lead to a growing interconnected web-scale data network. Behind this growth, there is a huge effort of data providers not only to publish their data but also to model and describe them following the LOD best practices. However, to ensure the interoperability of this large scale web of data, we would like to point out to the recommendation of building on, instead of replicating, existing *RDF schema* and vocabularies.

For more detailed explanation of the proposed recommendation, we start by the definition of the term *schema*, which is understood in the Linked Data context as "*the mixture of distinct terms from different RDF vocabularies that are used by a data source to publish data on the Web. This mixture may include terms from widely used vocabularies as well as proprietary terms*" (*cf.* the linked data guidelines [125]). Hereby, an important step towards linked data modeling is the discovery of all relevant vocabularies to reuse. This suggests that the metadata designer should reuse classes and properties from existing vocabularies rather than re-inventing them, and also combine several vocabularies when and where appropriate. For this purpose, existing ontology catalogues like Swoogle[1] provide the possibility of manual search for vocabularies terms. However, with the rapid growth of the LOD, the access[2] and the identification of suitable vocabularies to reuse, are becoming more and more difficult to perform manually.

The focus of this chapter falls on increase metadata designers and data provides awareness regarding vocabulary reuse in the context of LOD datasets. For this purpose, we introduce our vocabulary recommendation tool *Datavore* – *Da*ta *vo*cabulary *re*commender [126]. The tool is oriented towards metadata designers providing ranked lists of vocabulary terms to reuse in the web of data modeling process, together with additional metadata and cross-terms relations. *Datavore* relies on the Linked Open Vocabulary (LOV)[3] ecosystem for acquiring vocabularies and metadata. The system is able to deal with noisy and multilingual input data. The remainder of this chapter is structured as follows: Section 6.1 provides a detailed discussion of the various types of support features for modeling LOD datasets. Section 6.2 presents our tool *Datavore* and illustrates the general concept of our approach. Finally, we conclude in Section 6.3.

[1] `http://swoogle.umbc.edu/`
[2] vocabulary websites may be down or not up to date
[3] `lov.okfn.org`

## 6.1 Linked Data Modeling

The aim of section is to provide an overview of common practices and methods for LOD modeling, that is based on efficient selection and reuse of already existing vocabularies.

In the light of Linked Data publishing, the cookbook for open government Linked Data [8] defined the transformation of a domain model into an RDF schema into six steps:

1. Start with a robust Domain Model developed following a structured process and methodology.
2. Research existing terms and their usage and maximise reuse of those terms.
3. Where new terms can be seen as specialisations of existing terms, create sub class and sub properties as appropriate.
4. Where new terms are required, create them following commonly agreed best practice in terms of naming conventions etc
5. Publish within a highly stable environment designed to be persistent.
6. Publicise the RDF schema by registering it with relevant services.

In contrast, we summarize this process by three steps used to model Linked Data. The first step is data transformation, where data is converted from its original format (relational databases, CSV, HTML, ...) into a structured connected RDF representation using transformation tools (see Section 6.1.1). A further step is the vocabulary exploration where the search for existing ontology terms using vocabulary search engines (see Section 6.1.2). Finally, the last step is to develop the model using an ontology development tool, i.e., Protégé[4] or the vocabulary editor tool Neologism[5].

### 6.1.1 Data Transformation

We start by presenting the basic vision of [127], where an automatic mapping generation between relational databases and RDF models is introduced. To that end, data are presented as tables, which consists of rows, or records, where each record consists of a set of fields. The record is nothing but the content of its fields, just as an RDF node is nothing but the connections: the property values. In contrast, the mapping is very direct: (i) a record is an RDF node; (ii) the field (column) name is RDF propertyType; and (iii) the record field (table cell) is a value.

[4] `http://protege.stanford.edu/`
[5] `http://neologism.deri.ie`

In other words, the conversion is the result of a naive transformation where rows become subjects, column headers become predicates, and cells assert a single triple with an untyped string literal. Inspired by this representation, many tools are developed and can be used to deal for the tabular to RDF transformation. For instances the CSV and HTML tables can be represented in RDF using tools such as CSV2RDF [128] and DRETa [129]. For more complexes tabular formats, like Microsoft Excel, Google Sheets, and tables encoded in JSON or XML, a sophisticated framework like Opencube [130] and OpenRefine [131] can be used.

Equally important to note, mapping languages such as R2RML [132] or D2RQ [133] are used to express customized mappings from relational databases to RDF datasets. Such mappings provide the ability to view existing relational data in the RDF data model, expressed in a structure and target vocabulary of the mapping author's choice. Moreover, vocabularies such as CSVW Namespace Vocabulary Terms[6], enable the access of the of CSV metadata on the web by describing the content metadata in a separate JSON file that makes use of RDF vocabulary.

At the same time, this conversion process uses the structure of the data as it is organized in the database, which may not be the most useful structure of the information in RDF. But, either way, tabular data lacks features for expressing the semantics associated with the data contained in it, so it is challenging to know, in an automated and interoperable way, the meaning of the data enclosed in the data file.

On the other hand, semi automatic approaches are based on another common tabular data modeling strategy where each column name can be mapped to a class name or a pair of a property name and its domain, moreover to the class name or a subclass of it.

### 6.1.2 Vocabulary Exploration Methods

After transforming the data into well structured RDF, metadata designers are requested to reuse existing vocabulary terms to describe the well-structured RDF data. However, a subsequent challenge arise– "*how to find the appropriate set of terms to reuse from the wide variety of existing vocabularies?*".

Vocabulary search engines can provide helps in dealing with this challenge, from which we can cite Swoogle[7] which contains over 10,000 ontologies from RDF documents on the Web and the LOV which comprises more than 500 vocabulary spaces in the Linked Open Data cloud. An other directory of vocabularies on the LOD cloud is vocab.cc[8] which provides a lists of the top 100 classes and properties in the

[6] `https://www.w3.org/ns/csvw`
[7] `http://swoogle.umbc.edu/`
[8] `http://vocab.cc/`

Billion Triple Challenge data set[9]. In addition to generic vocabularies catalogues, we cite some catalogues of ontologies for a specific domain such as biomedicine with the BioPortal [134], the marine science domain with MMI[10] and the geospatial ontologies with SOCoP+OOR[11].

While vocab.cc only provides a link to the source representation of the vocabulary, LOV and Swoogle display the metadata provided by the vocabulary provider, which includes human readable text on how to use the terms semantically correct. In contrast, the metadata designer requires description on terms to reuse a specific class or property. Such metadata would include a ranking metric of a specific term. The most common type of ranking is the popularity indicating whether it is used by many or no data providers. In other words, this metric provides an indication on how widely a term is already used (in frequency and in the number of datasets using it).

The LOV search results rank each vocabulary terms based on its popularity in the LOD combined with an adaptation of the term frequency inverse document frequency (tf-idf) to the graph-structure of LOV dataset. In contrast, since the basic unit is not a word, but rather a vocabulary term in a vocabulary, LOV reuse the augmented frequency variation of term frequency formula to prevent a bias towards longer vocabularies. Further, LOV provide a sophisticated search interface where users can narrow a search by filtering on term type (class, property, datatype, instance), language, vocabulary domain and vocabulary. In other words, For every vocabulary in LOV, terms (classes, properties, datatypes, instances) are indexed and a full text search feature is offered[12]

In line with vocabulary ranking approaches, we cite [135] that examine the problem of vocabulary recommendation based on a ranking metric that has been developed by introducing the concept of Information Content (IC) to the context of LOV.

The *DWRank* approach [136] that address the ontology ranking problem by introducing two main scores: the *Hub score* and the *Authority score*, which aim to measure the centrality of the concepts within the ontology and the ontology authoritativeness (i.e. the importance of the ontology in the ontologies space), respectively.

Fernandez *et al.* [137] developed a collaborative ontology reuse and evaluation tool *CORE*. The tool receives as input a set of terms which expand using WordNet, than, performs a keyword-based searches to return a a ranked list of its indexed ontologies for reuse. An additionally step of CORE is the evaluation of the returned result based on criteria like semantic correctness, popularity. Romero *et al.* [138], developed a similar approach which make use of wikipedia and del.icio.us[13] to compute

[9] `http://km.aifb.kit.edu/projects/btc-2012/`
[10] `https://marinemetadata.org/aboutmmi`
[11] `https://ontohub.org/socop`
[12] `http://lov.okfn.org/dataset/lov/terms`
[13] http://del.icio.us

the popularity of terms.

Schaible *et al.* [139] provided a "Learning To Rank" approach called *TermPicker* which is based on features that combine the term's popularity combined with the schema-level pattern (SLP) feature. The SLP feature learn from the well-established LOD cloud, which terms other metadata designers used to describe their data. In contrast, the learning is based on how data providers on the LOD cloud actually combine the RDF types and properties from the different vocabularies to model their data. In other words, the user have to pick the subject $S$ of its triple and the system will perform the recommendation of a ranked list of ¡predicate, object¿ corresponding $S$.

However, reusing recommended vocabularies terms still requires additional efforts which consist in measuring the correctness of the mapping and incorporating the terms into the transformation system. The measurement of the semantic correctness of a mapping needs to be performed semi-automatically as it requires a human verification, preferably from a collaboration between a metadata designer and a domain expert. Metadata deseigner verifies the model quality with respect to the LOD best practices and guideline. On the other hand, the domain expert verifies the correctness of the model with respect to its representativeness of the data.

To the best of our knowledge, Karma [140] is the only tool that fills this gap between transformation tools and the recommendation system. Karma is an interactive tool assisting metadata designers (+domain experts) in modeling data sources as linked data by incrementally selecting recommended vocabulary **terms** although with a mapping for the data source. However, the main drawback of Karma is the fact that this tool does not handle the identification of suitable ontology (ies) to reuse. This process have to be done separately by the human user, which have to introduce selected ontology as input to the tool. From this input ontology, Karma recommends terms in order to be mapped to the data source.

Hence, in the next section, we will introduce our tool *Datavore* which handles this drawback by assisting metadata designer for a direct mapping between raw data source and the entire LOV ecosystem for a better vocabularies reuse.

## 6.2 Datavore: A Vocabulary Recommender Tool Assisting Linked Data Modeling

This section will introduce our user-interactive tool *Datavore* for linked data modeling assist. Our tool takes as input a tabular data and produce a ranked lists of vocabularies from the LOV. We start by arguing our choice of the LOV as the used vocabulary ecosystem, in other words "Why LOV ?".

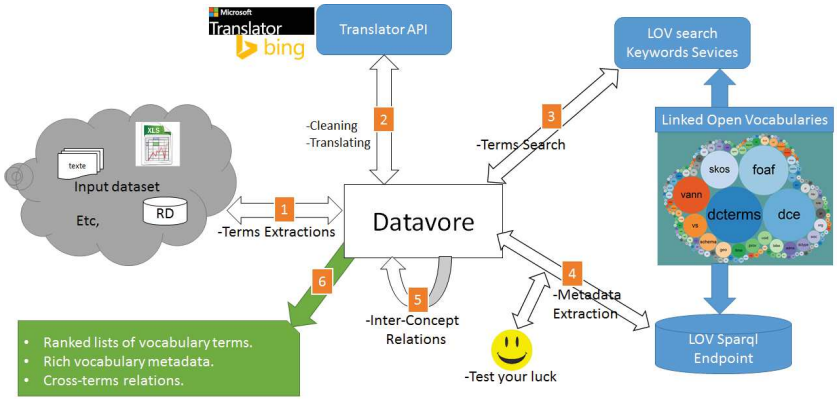


Figure 6.1: Datavore Workflow

As described in Section 6.1.2 from the existing vocabulary search engines, the search can be performed by utilizing a keyword-based or via a web services API. However, only LOV offers an access via a SPARQL endpoint and a dump file[14]. Furthermore, to ensure the high quality of LOV data, the insertion of new vocabularies in the LOV dataset passes through a manual inspection by LOV curators. These stringent requirements for vocabularies insertion are explained in the LOV handbook[141]. Moreover, it should be noted that since 2015, LOV has stopped removing offline vocabularies and decided to keep them with a special flag, making LOV the only source of continuity for datasets referencing unreachable vocabularies. Subsequently, and the most importantly all referenced vocabulary in LOV are accessible and having a high quality metadata descriptions.

In the following, we proceed to describe the workflow of our system which is depicted in Figure 6.1.

**Source Terms Extraction (1).** The input of *Datavore* is a list of terms extracted from the data source. In the current version of the prototype, we parse a loaded CSV file as input and extract the list of column names. We note that considering the row values from the CSV misled the recommendation in most cases that is why we limited ourselves to the column names. The use of other kinds of structured or semi-structured input data is envisageable.

**Cleaning and Translating (2).** Since the input is an untransformed raw data, in most cases, the extracted string of characters needs to be cleaned-up by removing or modifying the unwanted characters. We use the *Microsoft Translator* java API[15] in order to clean up the initial string and render it in a linguistically correct form.

[14] `http://lov.okfn.org/dataset/lov/sparql`
[15] `https://code.google.com/p/microsoft-translator-java-api/`

For example, translate(Cat$\tilde{A}$©gorie, fr, fr) returns Catégorie. In case there are no sufficient or satisfactory results by using the source language, the system uses the same service to translate the source item into English, the most common language on the LOD.

**Terms Search (3).** As we explained in the beginning of this section, we opted for the LOV as a vocabulary search engine, which, to the best of our knowledge, is the only purpose-built vocabulary search engine with an up-to-date index. As a design decision, *Datavore* queries the LOV search service with the extracted cleaned or/and translated terms. The result is a list of concepts for each source term ranked by the LOV metric, which is based on the popularity of the vocabulary terms in both the LOD datasets and the LOV ecosystem [142].

**Metadata Extraction (4).** Metadata designers are recommended to select popular vocabularies found in the search phase but it is not straightforward to judge which vocabulary is better suited to the application. *Datavore* queries the LOV endpoint (/dump file) to extract the needed metadata to help designers to choose the appropriate vocabularies. As a result, for each concept $c$, extract:

- The set of object properties having $c$ as domain that includes labels and hierarchical relations.
- The set of datatype properties that can be used with $c$ as domain. Example: Using the SPARQL endpoint of LOV to run the query shown in Figure 6.2, we were able to extract the datatype properties that can be used with the concept `http://rdf.insee.fr/def/geo#Commune`
- A link to the vocabulary web site.

In addition, we provide a "*test-your-luck*" option, which recommends to the user only one, the top ranked, datatype property. This "lucky" property has the highest Levenshtein string similarity [143] with the source term. Figure. 6.3 depicts the interface of *Datavore* in terms of the different cited options.

**Inter-Concept Relations (5).** From the extracted lists of recommended concepts, *Datavore* queries the LOV endpoint (/dump file) to retrieve cross-lists relations, i.e., relations between concepts from different lists. These metadata are crucial for selecting the best combination of predicate names to reuse for the input dataset.

```
PREFIX insee:<http://rdf.insee.fr/def/geo#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf
-syntax-ns#>
SELECT DISTINCT
?datatypePropURI ?propertyLabel ?Datatype
where {
?datatypePropURI rdf:type owl:DatatypeProperty.
?datatypePropURI rdfs:range ?Datatype.
OPTIONAL
    {?datatypePropURI rdfs:label ?propertyLabel.}
    {?datatypePropURI rdfs:domain insee:Commune.}
UNION{
    ?datatypePropURI rdfs:domain ?super.
    insee:Commune rdfs:subClassOf* ?super.
    }
}
```

Figure 6.2: SPARQL to retrieve the datatype properties having the concept "insee:Commune" as domain

| **Full Name** | **Profession** | **Lab** | **City** | **PostC** | **Country** |
|---|---|---|---|---|---|
| M. Ben Ellefi | PhD Student | LIRMM | Montpellier | 34090 | France |
| K. Todorov | Assoc. Pr. | LIRMM | Montpellier | 34095 | France |
| Z. Bellahsene | Pr. | LIRMM | Montpellier | 34095 | France |

Table 6.1: LIRMM Open Data Team Example.

### 6.2.1 Example Scenario

Imagine a designer who wants to model the data in Table 6.1 using an ontology editor. *Datavore* will guide him/her to find vocabularies to reuse, returning a sorted list of concepts for each column name. For the column "City Name" *Datavore* queries the LOV using the keyword "City" and returns the sorted list of concepts {"akt:City", "place:City", "lgdo:City", etc.}. When the designer selects the concept "akt:City", *Datavore* presents to him/her the following metadata: (i) literals (like rdfs:label, rdfs:comment, etc.) and the hierarchical relations of "akt:City", (ii) a set of datatype properties like "foaf:name" that have "akt:City" as *rdf:domain* to represent the column "City Name". After the concepts extraction, *Datavore* queries the LOV again to extract inter-columns triples and recommends to the user a set of relations between column names. In our example, the recommended relation between the two columns "Person Name" and "PostalAdress" is the triple: $< foaf : Person >< akt : hasAddress >< akt : PostalAddress >$.

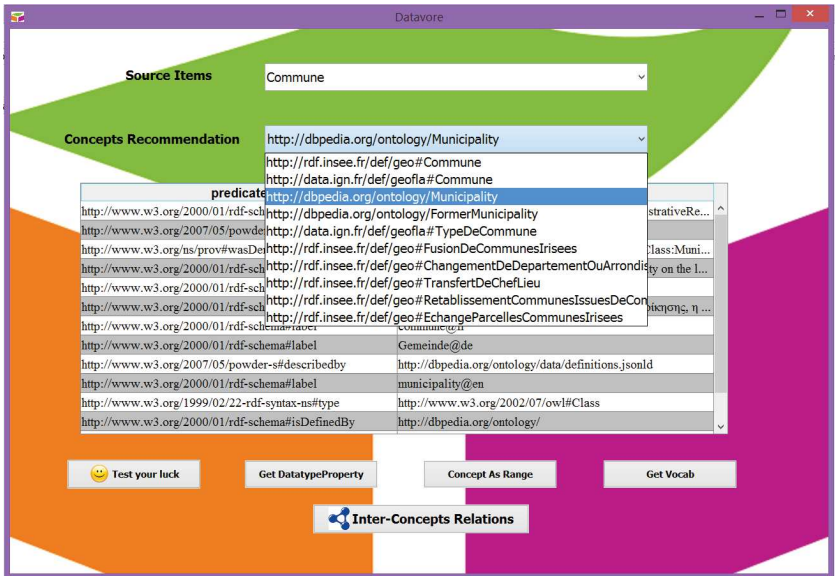


Figure 6.3: *Datavore* User Interface

### 6.2.2 Discussion

For a proper use of *Datavore*, we take note of the following. First, it should be noted that *Datavore* is not an RDF transformation tool neither an ontology development tools, but, a support and assist tool for metadata designer (+ domain expert) in the identification of suitable vocabulary terms for reuse. Hence, the main contribution of *Datavore* is the benefit of the whole LOV up-to-date catalogue in service of mapping the untransformed data to the recommended vocabularies terms. An additional contributions is the recommendation of suitable triples models (*inter-concept relations*) corresponding to the selected terms.

To our knowledge, there is only one comparable tool *Karma* [140] (see section 6.1.2). As there is no gold standard for such methods, we believe that *Datavore* can be evaluated in a user study manner. In other words, by asking the participants to model a number of datasets as LOD using three different methods:

- {*Datavore* + *protégé*}.
- {*Karma*}.
- {the LOV keywords search + *protégé*}.

Hence, *Datavore* will be evaluated against the two mentioned baselines in term of the following evaluation metrics: (i) the processing time – the total working time, (ii) the recommendation acceptance rate – the number of times that the participants chose the recommended term vs. the manual searching process), and (iii) the model quality – we can assess the quality by comparing the participants' models to models

generated by domains experts independently, i.e., did the participants choose the same vocabulary terms as the experts. A similar evaluation is done in the work of [144].

Considering the complexity, the tool is interactive, only the *inter-concept relations* task which is calculated automatically in term of the selected terms. The complexity of this task is of order $O(N^2 M_1 M_2)$, where, $N$ is the cardinality of the source terms, $M_1$ and $M_2$ are respectively the cardinalities of the two compared lists of recommended concepts.

Finally, *Datavore* has been implemented in Java and it is available as a $GUI$ desktop application[16] together with a demonstration video[17].

## 6.3 Conclusion

This chapter contains literature overview on Linked Data modeling process. We start by highlighting the different steps for transforming of domain model, in particular row data, into a Linked Data model. After transforming the data into well structured RDF, we discussed current strategies that explore existing vocabulary terms to be reused by metadata designers.

Next, we introduced our tool, *Datavore* - a vocabulary recommender system based on LOV assisting linked data modeling. Among the original features of the tool is the fact of providing metadata of recommended items as well as the recommendation of potential RDF relations which is particularly useful in the modeling step.

[16]Download and unzip this file: `http://www.lirmm.fr/benellefi/Datavore_ExeFile`
[17]`http://www.lirmm.fr/benellefi/Datavore_VideoDemo`

# Chapter 7

# Conclusion and Perspectives

## Contents

This chapter aims to summarize the main contributions of this thesis and to outline a number of research directions for future work. Section 7.1 highlights our contributions that deal with different challenges addressed in this thesis. Next, in Section 7.2, we discuss the future directions in order to extend and to broaden the research conducted in this dissertation.

## 7.1 Summary of Contributions

As stated in the introduction, challenges that we are dealing with, can be like "looking for a needle in a haystack", whether when looking for candidate RDF datasets for the interlinking task or looking for candidate vocabulary terms for the linked data modeling task. Furthermore, to this end, we need to extract the most representative set of features that better describe an RDF dataset for the given task.

Our **first contribution** consisted on the introduction of a new notion of features-based RDF dataset profiling which aim to be a guide for features extraction methods with respect to a given application scenario (see chapters 2 and 3). We provided a broad definition of dataset profiles where a profile can be seen as a comprehensive set of commonly investigated features that describe an RDF dataset. Furthermore, this definition has led to the creation of a new taxonomic classification of existing approaches in the field of profiles extraction systems. The proposal is structured around four main areas: semantic, qualitative, statistical and temporal. We stress on the fact that the organization of the taxonomy is in a non-strict hierarchy and the profile features are essentially selected with respect to a given application scenario.

Next, our work concerns the challenging problem of candidate dataset identification for a data linking scenario. To this end, **the second contribution** is the development of a Collaborative Filtering-like recommendation approach aiming to identify target datasets, potentially candidates to be linked with a source dataset (see Chapter 4). In line with the profiles features taxonomy, the proposed approach adopts an existing topic profiling method to represent different datasets. Furthermore, the approach uses the current topology of the LOD as learning data in order to learn the *connectivity behavior* of datasets using their topics profiles. Over the learning step, we have also discovered a new method that allows the propagation of topic profiles to the unprofiled LOD datasets in inexpensive manner. Extensive experiments aiming to evaluate the approach have been conducted using real world LOD datasets. Our approach showed a superior performance compared to the considered baselines. Moreover, our technique showed a generally good performance with respect to the evaluation data where it achieves a reduction of the original (LOD) search space of up to 86% on average. The main weak spot of this strategy is its dependence from the current topology of the LOD considered as weak learning data, which led to a precision of 19%.

In order to break away from this learning data, our **further contribution** consisted on providing another new candidate dataset recommendation approach that adopts different profile features, defined as *intensional profile* (see Chapter 5). To this extent, each dataset is profiled by a set of schema concept labels enriched by their corresponding textual descriptions. This intensional profiling method can be easily applied for any given dataset in a simple and inexpensive way. Based on this profile representation, the proposed approach recommends candidate datasets using a semantico-frequential concept similarity measure and a ranking criterion based on the tf*idf cosine similarity model. In line with the first approach, experiments were conducted using current LOD topology as evaluation data. This intensional approach showed a high recommendation efficiency by achieving an average precision of up to 53% for a recall of 100%. Hence, this learning break-off has led to a considerable improvement of our ranking efficiency, which confirms our hypothesis. Further contribution of this method is the ability to return mappings between the schema concepts across datasets, which is particularly useful in the linking configurations.

While the quality of our first recommendation approach is dependant on the learning step, we demonstrated that intension-based recommendation approach is totally dependant on the datasets schema description. To this extent, our **final contribution** aims to ensure high quality datasets intensional profiles by providing assistance in the Linked Data modeling process. To this end, we introduced *Datavore* - a vocabulary recommendation tool that assists metadata designers in the Linked Data modeling process. Our tool is designed to provide a ranked lists of vocabulary terms to reuse together with the descriptive metadata. Furthermore, *Datavore* provides a list of existing cross-terms relations in the form of triples. The current version of our prototype relies on the LOV ecosystem for acquiring existing vocabularies and metadata. Finally, *Datavore* is made available for download as a GUI prototype.

## 7.2 Open Issues

In the following, we outline various avenues for future work and research directions in order to improve or extend the main focuses that we are concerned with.

### Dataset Profiling

- We introduced the notion of profiles features which, to the best of our knowledge, has been the first study of its kind. The introduced taxonomy is in a non-strict hierarchy, in particular, we can envisage the possibility of expanding and collapsing hierarchical levels of a node. Furthermore, since it will be always new approaches for profiles features extraction, we intend to keep the

introduced taxonomy up-to-date by integrating new approaches or deleting outdated ones as possible.

- We provided a structured RDF vocabulary -VoDP- describing dataset profiling features which can be used to identify features as part of a dataset description. However, explicit mappings with available vocabularies have not been provided yet. In order to simplify dataset representation, we are currently working on providing explicit mappings between our dataset profiling features vocabulary and existing vocabularies meant for describing individual features. For instance, a range of potential existing vocabularies can be mapped to VoDP from which we can cite:

  **VoID -** the Vocabulary of Interlinked Datasets [145] which provides a core vocabulary for describing datasets and their links.

  **DCAT -** the Data Catalog Vocabulary [1] which follows a similar rationale as VoID and has been created based on a survey of government data catalogues[146].

  **WIQA-PL -** the Information Quality Assessment Policy Language [2] which is a vocabulary for modeling content access policies.

  **SCOVO -** the Statistical Core Vocabulary[3] can be used for statistical information.

## Identifying Candidate Datasets for Data Linking

- Our collaborative filtering-like recommendation approach gets its full performance from learning the connectivity behaviour of the existing linksets between LOD datasets. However, we demonstrated that the current topology of the LOD cloud group is far from being complete to be considered as a ground truth. Henceforth, in the future, we plan to improve the learning data quality by developing a more reliable and complete ground truth for dataset recommendation. To this extent, one future direction can be the use of crowdsourcing methods that can be implemented as follows:
  (1) The identification of new possible relations among Web datasets, i.e., the recommendation results of the CCD-CosineRank approach: `http://www.lirmm.fr/benellefi/CCD-CosineRank_Result.csv`.
  (2) The implementation of a questionnaire asking for the validity of these new relations. Each dataset will be represented by its descriptive profile, e.g., the title, the set of schema concepts, the accessibility point, the publishers and

[1] `http://www.w3.org/TR/vocab-dcat/`
[2] `http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/#wiqapl`
[3] `http://purl.org/NET/scovo`

maintainers, etc.
(3) A rigourous validation of these relations can be done by Semantic Web researchers who are trusted by the community.

Moreover, this ground truth can be used as an evaluation data. Hence, a further future work can be to improve the evaluation framework in order to deal with the false positives overestimation problem.

- Regarding the intension-based dataset recommendation approach, further work can go into obtaining higher quality intensional profiles, in particular by:
(1) Assessing the population of the schema elements which can lead to assigning a weight to each label profile or even filtering out certain concepts schema based on a given threshold.
(2) Including the dataset context in its profile which can lead to avoid some ambiguities, such as special cases when datasets with similar intension may have extensional overlap, i.e., two datasets talking about books and authors, but from different domains science *vs.* literature.
(3) Considering different natural languages, in particular by comparing multilingual label profiles, i.e., "country" *vs.* "land" *vs.* "pays".

## Vocabulary Selection for Linked Data Modeling

- Our prototype *Datavore* is based on the trustworthy LOV search engine. However, it should be noted that this vocabulary catalogue, in its current version, contains only less than 550 vocabularies. Hence, one future direction can be to look for the possibilities of extending *Datavore* recommendation via larger vocabularies catalogues like Swoogle and even providing a new cross ranking approach for the combined search results via multiple catalogues.

  In addition, *Datavore* can be integrated with an ontology development tool, such as Protégé, in order to reduce the user effort by having one GUI tool for Linked Data modeling.

  Moreover, another future direction can be the provision of a new evaluation framework for the linked data modeling community notably in a user study manner and crowdsourcing-based in order to ensure a fair evaluation for LD-modeling support systems.

- Many works have studied the strategies of vocabulary reuse and the rank factors for linked data vocabulary term recommendations. To the best of our knowledge, there is no study which empirically examines the vocabulary terms ranking strategies based on the data structure factors. More insights about the difference between tabular sources modeling and web pages annotation which

influence the metadata designers in their decision to select reusable classes and properties. We observed that there is a need of an empirical study of different ranking approaches and recommender systems to identify these two strategies and factors requirements. We plan to provide a guideline survey about the modeling process in order to assist metadata designers in deciding: *which ranking strategy is the best for their needs*?

The ultimate goal of our work is to contribute to the improvement of the quality of the current "global search space", which, in the long run, will benefit the society as a whole by providing better means to access and interpret information.

# Bibliography


[1] B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl, "A scalable approach for efficiently generating structured dataset topic profiles," in *Proc. of the 11th ESWC*, Springer, 2014.

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2009.

[3] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[4] p. World Wide Web Consortium, year=2009, *W3C Semantic Web Activity, http://www.w3.org/2001/sw/.*

[5] p. World Wide Web Consortium, year=2004, *W3C Resource Description Framework (RDF), http://www.w3.org/RDF/.*

[6] L. Masinter, T. Berners-Lee, and R. T. Fielding, "Uniform resource identifier (uri): Generic syntax," 2005.

[7] C. Basca, S. Corlosquet, R. Cyganiak, S. Fernández, and T. Schandl, "Neologism: Easy vocabulary publishing," in *4th Workshop on Scripting for the Semantic Web*, 2008.

[8] B. Hyland, B. Terrazas, and S. Capadisli, "Cookbook for open government linked data," *W3C, W3C Task Force-Government Linked Data Group*, 2011.

[9] M. H. R. Cygankiak, "Linked data life cycles," *formerly at http://linked-data-life-cycles.info/.*

[10] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, "Methodological guidelines for publishing government linked data," in *Linking government data*, pp. 27–49, Springer, 2011.

[11] F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Képéklian, F. Cotton, *et al.*, "Enabling linked-data publication with the datalift platform," in *Proc. AAAI workshop on semantic cities*, pp. No–pagination, 2012.

[12] S. Auer and J. Lehmann, "Creating knowledge out of interlinked data," *Semantic Web*, vol. 1, no. 1, 2, pp. 97–104, 2010.

[13] F. Ricci, L. Rokach, and B. Shapira, *Introduction to recommender systems handbook*. Springer, 2011.

[14] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[15] S. Lalithsena, P. Hitzler, A. P. Sheth, and P. Jain, "Automatic domain identification for linked open data," in *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17-20, 2013*, pp. 205–212, 2013.

[16] B. Fetahu, S. Dietze, B. P. Nunes, M. A. Casanova, D. Taibi, and W. Nejdl, "A scalable approach for efficiently generating structured dataset topic profiles," in *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pp. 519–534, 2014.

[17] A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Entity-based data source contextualization for searching the web of data," in *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, Crete, Greece, May 26, 2014.*, 2014.

[18] A. Wagner, P. Haase, A. Rettinger, and H. Lamm, "Discovering related data sources in data-portals," in *Proceedings of the First International Workshop on Semantic Statistics, co-located with the the International Semantic Web Conference*, 2013.

[19] M. Konrath, T. Gottron, S. Staab, and A. Scherp, "Schemex - efficient construction of a data catalogue by stream-based indexing of linked data," *J. Web Sem.*, vol. 16, pp. 52–58, 2012.

[20] M. B. Ellefi, Z. Bellahsene, F. Scharffe, and K. Todorov, "Towards semantic dataset profiling," in *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, Crete, Greece, May 26, 2014.*, 2014.

[21] M. Atencia, J. David, and F. Scharffe, "Keys and pseudo-keys detection for web datasets cleansing and interlinking," in *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pp. 144–153, 2012.

[22] F. Naumann, "Data profiling revisited," *SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2013.

[23] P. N. Mendes, H. Mühleisen, and C. Bizer, "Sieve: linked data quality assessment and fusion," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pp. 116–123, 2012.

[24] C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," *J. Web Sem.*, vol. 7, no. 1, pp. 1–10, 2009.

[25] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, "An empirical survey of linked data conformance," *J. Web Sem.*, vol. 14, pp. 14–44, 2012.

[26] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment methodologies for linked open data," in *Under Review*, 2014.

[27] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[28] C. BIZER, *Quality-Driven Information Filtering in the Context of Web-Based Information Systems.* PhD thesis, Freie Universitat, Berlin, March 2007.

[29] L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[30] S. Auer, J. Demter, M. Martin, and J. Lehmann, "Lodstats - an extensible framework for high-performance dataset analytics," in *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pp. 353–362, 2012.

[31] B. Forchhammer, A. Jentzsch, and F. Naumann, "LODOP - multi-query optimization for linked data profiling queries," in *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, May 26, 2014.*, 2014.

[32] T. Käfer, J. Umbrich, A. Hogan, and A. Polleres, "Dyldo: Towards a dynamic linked data observatory," in *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, 2012.

[33] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan, "Observing linked data dynamics," in *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, pp. 213–227, 2013.

[34] N. Popitsch and B. Haslhofer, "Dsnotify: handling broken links in the web of data," in *Proceedings of the 19th International Conference on World Wide*

*Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pp. 761–770, 2010.

[35] T. Gottron and C. Gottron, "Perplexity of index models over evolving linked data," in *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pp. 161–175, 2014.

[36] R. Q. Dividino, A. Scherp, G. Gröner, and T. Grotton, "Change-a-lod: Does the schema on the linked data cloud change or not?," in *Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, 2013*, 2013.

[37] C. Böhm, J. Lorey, and F. Naumann, "Creating void descriptions for web-scale data," *J. Web Sem.*, vol. 9, no. 3, pp. 339–345, 2011.

[38] D. Symeonidou, V. Armant, N. Pernelle, and F. Saïs, "Sakey: Scalable almost key discovery in RDF data," in *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pp. 33–49, 2014.

[39] D. Symeonidou, N. Pernelle, and F. Saïs, "KD2R: A key discovery method for semantic reference reconciliation," in *On the Move to Meaningful Internet Systems: OTM 2011 Workshops - Confederated International Workshops and Posters: EI2N+NSF ICE, ICSP+INBAST, ISDE, ORM, OTMA, SWWS+MONET+SeDeS, and VADER 2011, Hersonissos, Crete, Greece, October 17-21, 2011. Proceedings*, pp. 392–401, 2011.

[40] Y. Gil and V. Ratnakar, "TRELLIS: an interactive tool for capturing information analysis and decision making," in *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Proceedings*, pp. 37–42, 2002.

[41] J. Blythe and Y. Gil, "Incremental formalization of document annotations through ontology-based paraphrasing," in *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pp. 455–461, 2004.

[42] J. Golbeck, B. Parsia, and J. A. Hendler, "Trust networks on the semantic web," in *Cooperative Information Agents VII, 7th International Workshop, CIA 2003, Helsinki, Finland, August 27-29, 2003, Proceedings*, pp. 238–249, 2003.

[43] O. Hartig, "Trustworthiness of data on the web," in *Proceedings of the STI Berlin & CSW PhD Workshop*, 2008.

[44] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," in *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pp. 87–102, 2012.

[45] E. Ruckhaus, M. Vidal, S. Castillo, O. Burguillos, and O. Baldizan, "Analyzing linked data quality with liquate," in *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pp. 488–493, 2014.

[46] D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann, "Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data," in *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*, pp. 265–272, 2013.

[47] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri, "Test-driven evaluation of linked data quality," in *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pp. 747–758, 2014.

[48] A. Langegger and W. Wöß, "Rdfstats - an extensible RDF statistics generator and library," in *Database and Expert Systems Applications, DEXA, International Workshops, Linz, Austria, August 31-September 4, 2009, Proceedings*, pp. 79–83, 2009.

[49] C. Bizer, A. Jentzsch, and R. Cyganiak, "State of the lod cloud," *Version 0.3 (September 2011), http://lod-cloud.net/state/*, vol. 1803, 2011.

[50] S. Khatchadourian and M. Consens, "Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud," in *The Semantic Web: Research and Applications*, vol. 6089 of *Lecture Notes in Computer Science*, pp. 272–287, Springer Berlin Heidelberg, 2010.

[51] Z. Abedjan, T. Grütze, A. Jentzsch, and F. Naumann, "Profiling and mining RDF data with prolod++," in *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pp. 1198–1201, 2014.

[52] C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl, and D. Sonnabend, "Profiling linked open data with prolod," in *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, pp. 175–178, 2010.

[53] E. Mäkelä, "Aether–generating and viewing extended void statistical descriptions of rdf datasets," in *The Semantic Web: ESWC 2014 Satellite Events*, pp. 429–433, Springer, 2014.

[54] B. Fitzpatrick, B. Slatkin, and M. Atkins, "Pubsubhubbub core 0.3–working draft," *Project Hosting on Google Code, available at http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.3.html*, 2010.

[55] A. Passant and P. N. Mendes, "sparqlpush: Proactive notification of data updates in RDF stores using pubsubhubbub," in *Proceedings of the Sixth Workshop on Scripting and Development for the Semantic Web, Crete, Greece, May 31, 2010*, 2010.

[56] U. Bojars, A. Passant, F. Giasson, and J. G. Breslin, "An architecture to discover and query decentralized RDF data," in *Proceedings of the ESWC'07 Workshop on Scripting for the Semantic Web, SFSW 2007, Innsbruck, Austria, May 30, 2007*, 2007.

[57] S. Tramp, P. Frischmuth, T. Ermilov, S. Shekarpour, and S. Auer, "An architecture of a distributed semantic social network," *Semantic Web*, vol. 5, no. 1, pp. 77–95, 2014.

[58] H. V. de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar, "Memento: Time travel for the web," *CoRR*, vol. abs/0911.1112, 2009.

[59] H. V. de Sompel, R. Sanderson, M. L. Nelson, L. Balakireva, H. Shankar, and S. Ainsworth, "An http-based versioning mechanism for linked data," in *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, 2010.

[60] G. Cheng and Y. Qu, "Searching linked objects with falcons: Approach, implementation and evaluation," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 49–70, 2009.

[61] M. B. Ellefi, Z. Bellahsene, J. G. Breslin, E. Demidova, S. Dietze, J. Szymanski, and K. Todorov, "Dataset profiling–a guide to features, methods, applications and vocabularies," in *Major Revision Statue In the Semantic Web Journal*.

[62] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: an ontology-based approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 4, pp. 434–452, 2011.

[63] P. Castells, M. Fernandez, and D. Vallet, "An adaptation of the vector-space model for ontology-based information retrieval," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 2, pp. 261–272, 2007.

[64] J. Szymański, "Comparative analysis of text representation methods using classification," *Cybernetics and Systems*, vol. 45, no. 2, pp. 180–199, 2014.

[65] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to wikipedia," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 1375–1384, 2011.

[66] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 509–518, ACM, 2008.

[67] E. M. Voorhees, "Using wordnet for text retrieval," *Fellbaum (Fellbaum, 1998)*, pp. 285–303, 1998.

[68] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology learning from text: An overview*, vol. 123. 2005.

[69] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Information processing & management*, vol. 43, no. 4, pp. 866–886, 2007.

[70] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 389–396, ACM, 2009.

[71] C. B. Jones, H. Alani, and D. Tudhope, "Geographical information retrieval with ontologies of place," in *Spatial information theory*, pp. 322–335, Springer, 2001.

[72] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, "Survey of temporal information retrieval and related applications," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 15, 2014.

[73] O. Alonso, J. Strötgen, R. A. Baeza-Yates, and M. Gertz, "Temporal information retrieval: Challenges and opportunities.," *TWAW*, vol. 11, pp. 1–8, 2011.

[74] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: a spatially and temporally enhanced knowledge base from wikipedia," *Artificial Intelligence*, vol. 194, pp. 28–61, 2013.

[75] A. Abdelali, J. Cowie, D. Farwell, B. Ogden, and S. Helmreich, "Cross-language information retrieval using ontology," in *Proceedings of TALN*, 2003.

[76] B. Koopman, P. Bruza, L. Sitbon, and M. Lawley, "Towards semantic search and inference in electronic medical records: an approach using concept-based

information retrieval," *The Australasian medical journal*, vol. 5, no. 9, p. 482, 2012.

[77] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[78] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature reviews genetics*, vol. 7, no. 2, pp. 119–129, 2006.

[79] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biology*, vol. 2, no. 11, p. e309, 2004.

[80] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[81] J. M. Abasolo and M. Gomez, "Melisa: An ontology-based agent for information retrieval in medicine," in *Proceedings of the first international workshop on the semantic web (SemWeb2000)*, pp. 73–82, 2000.

[82] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome biology*, vol. 6, no. 7, p. 224, 2005.

[83] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann, "User-driven quality evaluation of dbpedia," in *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pp. 97–104, 2013.

[84] H. Paulheim and C. Bizer, "Improving the quality of linked data using statistical distributions," *Int. J. Semantic Web Inf. Syst.*, vol. 10, no. 2, pp. 63–86, 2014.

[85] D. Fleischhacker, H. Paulheim, V. Bryl, J. Völker, and C. Bizer, "Detecting errors in numerical linked data using cross-checked outlier detection.," in *Semantic Web Conference (1)*, pp. 357–372, 2014.

[86] D. Wienand and H. Paulheim, "Detecting incorrect numerical data in dbpedia," in *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pp. 504–518, 2014.

[87] H. Paulheim, "Identifying wrong links between datasets by multi-dimensional outlier detection," in *Proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2014, co-located with*

*11th Extended Semantic Web Conference (ESWC 2014), Anissaras/Hersonissou, Greece, May 26, 2014.*, pp. 27–38, 2014.

[88] V. Bryl and C. Bizer, "Learning conflict resolution strategies for cross-language wikipedia data fusion," in *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pp. 1129–1134, 2014.

[89] W. Yuan, E. Demidova, S. Dietze, and X. Zhou, "Analyzing relative incompleteness of movie descriptions in the web of data: A case study," in *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pp. 197–200, 2014.

[90] J. Bleiholder and F. Naumann, "Data fusion," *ACM Comput. Surv.*, vol. 41, pp. 1:1–1:41, Jan. 2009.

[91] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich, "Data summaries for on-demand queries over linked data," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 411–420, ACM, 2010.

[92] "Fedsearch: Efficiently combining structured queries and full-text search in a sparql federation," vol. 8218 of *Lecture Notes in Computer Science*, pp. 427–443, Springer Berlin Heidelberg, 2013.

[93] O. Hartig, C. Bizer, and J. C. Freytag, "Executing sparql queries over the web of linked data," in *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pp. 293–309, 2009.

[94] N. K. Yeganeh, S. Sadiq, and M. A. Sharaf, "A framework for data quality aware query systems," *Inf. Syst.*, vol. 46, pp. 24–44, Dec. 2014.

[95] A. Wagner, D. T. Tran, G. Ladwig, A. Harth, and R. Studer, "Top-k linked data query processing," in *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pp. 56–71, 2012.

[96] O. Görlitz and S. Staab, "Splendid: Sparql endpoint federation exploiting void descriptions," in *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), Bonn, Germany, October 23, 2011*, 2011.

[97] S. Elbassuoni, M. Ramanath, R. Schenkel, M. Sydow, and G. Weikum, "Language-model-based ranking for queries on rdf-graphs," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, (New York, NY, USA), pp. 977–986, ACM, 2009.

[98] F. Naumann, *Quality-driven Query Answering for Integrated Information Systems*. Berlin, Heidelberg: Springer-Verlag, 2002.

[99] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni, "The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems," *Information Systems*, vol. 29, no. 7, pp. 551 – 582, 2004. Data Quality in Cooperative Information Systems.

[100] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, "Improving efficiency and accuracy in multilingual entity extraction," in *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pp. 121–124, 2013.

[101] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *TACL*, vol. 2, pp. 231–244, 2014.

[102] A. N. Ngomo and S. Auer, "LIMES - A time-efficient approach for large-scale link discovery on the web of data," in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 2312–2317, 2011.

[103] A. Jentzsch, R. Isele, and C. Bizer, "Silk–generating rdf links while publishing or consuming linked data," in *9th International Semantic Web Conference (ISWC'10)*, Citeseer, 2010.

[104] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pp. 722–735, 2007.

[105] H. R. de Oliveira, A. T. Tavares, and B. F. Lóscio, "Feedback-based data set recommendation for building linked data applications," in *Proc. of the 8th ISWC*, pp. 49–55, ACM, 2012.

[106] C. Baron Neto, K. Müller, M. Brümmer, D. Kontokostas, and S. Hellmann, "Lodvader: An interface to lod visualization, analyticsand discovery in real-time," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 163–166, 2016.

[107] A. Nikolov and M. d'Aquin, "Identifying relevant sources for data linking using a semantic web index," in *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India*, 2011.

[108] L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze, "Identifying candidate datasets for data interlinking," in *Proc. of the 13th ICWE, Aalborg, Denmark*, pp. 354–366, 2013.

[109] G. Lopes, L. A. Paes Leme, B. Nunes, M. Casanova, and S. Dietze, "Two approaches to the dataset interlinking recommendation problem," in *Proc. of 15th on WISE 2014*, 2014.

[110] M. Mehdi, A. Iqbal, A. Hogan, A. Hasnain, Y. Khan, S. Decker, and R. Sahay, "Discovering domain-specific public SPARQL endpoints: a life-sciences use-case," in *Proc. of the 18th IDEAS 2014, Porto, Portugal*, pp. 39–45.

[111] M. Röder, A.-C. N. Ngomo, I. Ermilov, and A. Both, "Detecting similar linked datasets using topic modelling," in *International Semantic Web Conference*, pp. 3–19, Springer, 2016.

[112] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[113] M. B. Ellefi, Z. Bellahsene, S. Dietze, and K. Todorov, "Beyond established knowledge graphs-recommending web datasets for data linking," in *Web Engineering - 16th International Conference, ICWE, Lugano, Switzerland, June 6-9, 2016. Proceedings*, pp. 262–279, 2016.

[114] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[115] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web.," 1999.

[116] A. Markov, "Extension of the limit theorems of probability theory to a sum of variables connected in a chain," 1971.

[117] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[118] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991.

[119] L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese, "Umbc_ebiquity-core: Semantic textual similarity systems," in *Proc. of the *SEM*, Association for Computational Linguistics, 2013.

[120] M. B. Ellefi, Z. Bellahsene, S. Dietze, and K. Todorov, "Dataset recommendation for data linking: An intensional approach," in *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, pp. 36–51, 2016.

[121] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. of the 32Nd ACL*, pp. 133–138, 1994.

[122] D. Lin, "An information-theoretic definition of similarity," in *Proc. of ICML*, pp. 296–304, 1998.

[123] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, Nov. 1995.

[124] T. Gottron, M. Knauf, S. Scheglmann, and A. Scherp, "A systematic investigation of explicit and implicit schema information on the linked open data cloud," in *Proc. of ESWC*, pp. 228–242, 2013.

[125] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[126] M. B. Ellefi, Z. Bellahsene, and K. Todorov, "Datavore: A vocabulary recommender tool assisting linked data modeling," in *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference ISWC, Bethlehem, PA, USA, October 11, 2015.*, 2015.

[127] T. Berners-Lee, "Relational databases on the semantic web," 2013.

[128] T. Lebo and J. McCusker, "csv2rdf4lod," tech. rep., Technical report, Tetherless World, RPI, 2012. https://github. com/timrdf/csv2rdf4lod-automation/wiki, 2012.

[129] E. Munoz, A. Hogan, and A. Mileo, "Dreta: extracting rdf from wikitables," in *Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035*, pp. 89–92, CEUR-WS. org, 2013.

[130] E. Kalampokis, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris, and K. Tarabanis, "Exploiting linked data cubes with opencube toolkit," in *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pp. 137–140, CEUR-WS. org, 2014.

[131] T. Morris, T. Guidry, and M. Magdinie, "Openrefine: A free, open source, powerful tool for working with messy data," tech. rep., Technical report, The OpenRefine Development Team, 2015. http://openrefine. org, 2015.

[132] S. Das, S. Sundara, and R. Cyganiak, "{R2RML: RDB to RDF Mapping Language}," 2012.

[133] C. Bizer and A. Seaborne, "D2rq-treating non-rdf databases as virtual rdf graphs," in *Proceedings of the 3rd international semantic web conference (ISWC2004)*, vol. 2004, Citeseer Hiroshima, 2004.

[134] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, *et al.*, "The nhgri gwas catalog,

a curated resource of snp-trait associations," *Nucleic acids research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[135] G. A. Atemezing and R. Troncy, "Information content based ranking metric for linked open vocabularies," in *Proceedings of the 10th International Conference on Semantic Systems*, pp. 53–56, ACM, 2014.

[136] A. S. Butt, "Ontology search: finding the right ontologies on the web," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 487–491, ACM, 2015.

[137] M. Fernández, I. Cantador, and P. Castells, "Core: A tool for collaborative ontology reuse and evaluation," 2006.

[138] M. M. Romero, J. M. Vázquez-Naya, C. R. Munteanu, J. Pereira, and A. Pazos, "An approach for the automatic recommendation of ontologies using collaborative knowledge," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 74–81, Springer, 2010.

[139] J. Schaible, T. Gottron, and A. Scherp, "Termpicker: Enabling the reuse of vocabulary terms by exploiting data from the linked open data cloud," in *International Semantic Web Conference*, pp. 101–117, Springer International Publishing, 2016.

[140] C. A. Knoblock, P. Szekely, J. L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, and P. Mallick, "Semi-automatically mapping structured sources into the semantic web," in *ESWC*, pp. 375–390, Springer, 2012.

[141] P.-Y. Vandenbussche and B. Vatant, "Metadata recommendations for linked open data vocabularies," *Version*, vol. 1, pp. 2011–12, 2011.

[142] P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalón, and B. Vatant, "Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web," *Semantic Web*, no. Preprint, pp. 1–16.

[143] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

[144] J. Schaible, P. Szekely, and A. Scherp, "Comparing vocabulary term recommendations using association rules and learning to rank: A user study," in *International Semantic Web Conference*, pp. 214–230, Springer, 2016.

[145] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao, "Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'.," in *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, (Madrid, Spain), 2009.

[146] F. Maali, R. Cyganiak, and V. Peristeras, "Enabling interoperability of government data catalogues.," in *EGOV* (M. Wimmer, J.-L. Chappelet, M. Janssen,

and H. J. Scholl, eds.), vol. 6228 of *Lecture Notes in Computer Science*, pp. 339–350, Springer, 2010.