

Point de vue ontologique de fonds documentaires territorialisés indexés

Eric Kergosien

▶ **To cite this version:**

Eric Kergosien. Point de vue ontologique de fonds documentaires territorialisés indexés. Recherche d'information [cs.IR]. Université de Pau et des Pays de l'Adour, 2011. Français. NNT : . tel-00720439

HAL Id: tel-00720439
https://theses.hal.science/tel-00720439

Submitted on 24 Jul 2012

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

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

ÉCOLE DOCTORALE DES SCIENCES EXACTES ET DE LEURS APPLICATIONS ED 211

Point de vue ontologique de fonds documentaires territorialisés indexés

THÈSE

présentée et soutenue publiquement le 2011

pour l'obtention du

Doctorat de l'Université de Pau et des Pays de l'Adour

(spécialité informatique)

par

Eric Kergosien

Composition du jury

Rapporteurs :	Jérôme GENSEL	STEAMER, LIG à l'Université Joseph Fourier, Grenoble
	Amedeo NAPOLI	ORPAILLEUR, LORIA - INRIA Nancy Grand EST
Présidente :	Nathalie Aussenac-Gilles	IC3, IRIT à l'Université Paul Sabatier, Toulouse
Examinatrice :	Chantal Reynaud	IASI, LRI à l'Université Paris XI, Orsay
Directeur :	Mauro GAIO	T2I, LIUPPA à l'Université de Pau et des Pays de l'Adour, Pau
Co-Encadrant :	Alain Du Boisdulier	Dirigeant, entreprise Document Image Solutions, Bidart

Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour — EA 3000

LIUPPA

Mis en page avec la classe thloria.

Remerciements

Je tiens ici à remercier toutes les personnes, proches ou lointaines qui ont fait que ces travaux de thèse ont pu être menés à bien. C'est un tel plaisir d'écrire ces remerciements envers tant de personnes qui m'ont soutenue que je risque de m'étaler un peu.

Je commencerai par remercier Mauro Gaio mon directeur de thèse pour avoir cru en moi tout au long de ma thèse. Merci pour m'avoir aider à m'ouvrir l'esprit au niveau scientifique et pour ces nombreux échanges ô combien intéressants à la frontière entre l'informatique et la géographie. Je tiens à le remercier également pour avoir su me « bousculer » dans des moments où cela était nécessaire.

Je remercie également Alain Du Boisdulier qui m'a accuelli au sein de la société DIS et qui a également été très présent tout au long de ces années. Il a joué un rôle moteur, notamment dans les objectifs de ces travaux. Je tiens à le remercier également pour son ouverture d'esprit, sa compréhension et ses conseils qui m'ont été très précieux. J'ai beaucoup appris à l'entreprise DIS et c'est en partie grâce à lui.

Je remercie vivement Jérôme Gensel et Amedeo Napoli qui m'ont fait l'honneur de rapporter ma thèse, Nathalie Aussenac-Gilles et Chantal Reynaud qui ont accepté de faire partie de mon jury. Je suis très honoré d'avoir un jury si renommé !

Vient maintenant le moment de remercier tout l'entourage professionnel. Je tiens à remercier l'équipe des experts documentalistes de la Médiathèque Intercommunale à Dimension Régionale de Pau et notamment Nicolas Barbey pour nous avoir permi d'accéder et de travailler avec un volume important de documents annotés. Merci également pour leurs connaissances expertes nécessaires au bon déroulement de ce travail scientifique. J'espère que le projet d'industrialisation de notre application TerridocViewer de navigation au sein de fonds documentaires verra le jour et permettra à la MIDR de Pau de disposer d'un moteur de recherche plus adapté à leurs collections.

Je remercie l'équipe d'experts RAMEAU de la Bibliothèque nationale de France (BnF) et notamment Michel Mingam pour nous avoir donné accès à la ressource thésaurus RAMEAU, pour ses conseils d'expert et pour nous avoir permis de présenter notre approche à plusieurs reprises aux différentes équipes travaillant autour du catalogage au sein même de la BnF. J'espère que nous pourrons continuer à collaborer et que l'idée de faire évoluer l'application TerridocViewer pour l'utiliser à la BnF afin de naviguer à travers le thésaurus RAMEAU va pouvoir se concrétiser.

Je souhaite aussi remercier nos collègues du projet ANR Géonto, avec qui nous avons pu mener une collaboration très enrichissante et fructueuse (plusieurs articles, dont l'un des meilleurs papiers de JFO'2009 nous donnant la possibilité de publier nos travaux dans la revue TSI).

Je remercie l'ensemble des membres du laboratoire LIUPPA, ceux du département informatique de l'Université de Pau ainsi que ceux de l'IUT informatique de Bayonne pour m'avoir accueilli et appuyé.

Ce fut un grand plaisir de travailler au sein de l'équipe DESI, devenue T2I en 2009, et je remercie Marie-Noelle, Christian, Thierry, Christophe, Patrick, Pantxika, les deux Philippe, Albert, Congduc, et Annig pour leurs conseils aux différentes étapes impor-

tantes de la thèse. Merci à tous les anciens doctorants pour leur aide et leurs conseils : les deux Julien, Pierre, Christine, Cyril et Damien. Je souhaite aussi encourager tous les doctorants encore au laboratoire : Thanh Vu, Van Tien, Minh Duc, Nour, Youssef, Julien, John, Camille, Keling, Hui et Muhammad. Merci également à l'association des doctorants AMDA avec qui il fut possible de collaborer pour tenter d'améliorer la vie des doctorants. Merci à la « team » des doctorants de divers horizons (Adil, Sylvain, Guillaume, Lionel, rachid et les autres) pour l'ambiance quotidienne et pour avoir partagé tous ces repas du midi. Merci les amis et à vous de jouer pour la plupart.

Je remercie également les membres de l'équipe DIS (Agnès, Nathalie, Marlène et Sylvie) pour tous ces moments passés durant la thèse.

Merci particulièrement à Stéphanie, Sabrina et Christophe d'avoir accepté de relire ce manuscrit pour corriger les (trop) nombreuses fautes restantes !

Je souhaite bien entendu remercier tous mes amis qui m'ont supporté et encouragé pendant ce travail ! Merci aux « Locs » et notamment à Jeff et Gio pour tout ! Merci aux « chanziens » et notamment les deux Julien, Remo, Majda, Domenico et tous les autres. Merci à la « Dossou Team » pour tous ces moments inoubliables que l'on a passé et je sais d'avance que ce n'est rien comparé à ce qui viendra par la suite. Il va d'ailleurs falloir penser à trouver autre chose que « étudiant à durée indeterminée » pour me chambrer :-p. Merci à « l'équipe de choc » des vacataires du Crous, à celle du Tropical Café ainsi qu'aux amis palois (Natacha, Aurel, Marlène et les autres) pour tous ces fous rires et cette entraide. J'aimerais vous remercier tous individuellement mais ce n'est pas possible ici. Sachez tout de même que mes pensées sont pour chacun d'entre vous !

Pour finir, je remercie vivement ma famille et notamment ma mère et mon père pour leur éducation et pour m'avoir ouvert l'esprit très jeune avec tous ces voyages. Je remercie également à nouveau ma mère et mon frère pour leur soutien, mon oncle Gilles ainsi que sa femme Martine et sa fille Orlane, mes oncles et tantes Noellie, Georges et Pascale qui forment le cœur de ma famille. Merci également à ma deuxième famille de « K'assos » et notamment à Eric, Raynia, Medhy, Marie-Belle et bien entendu Rafik, Bertrand, Gael et Sabrina : ils sont devenus mes frères et soeurs !

Mes remerciements enfin à Nirina, avec qui j'ai partagé tous les moments les plus forts de ces quatre dernières années. Je tiens à la remercier de tout mon cœur, je sais combien mon humeur, mes préoccupations sont pesantes depuis un an. Elle a su être patiente et m'a, elle aussi, toujours soutenu dans tous mes choix afin de me donner confiance en moi. Celui qui n'a pas vécu une thèse ne peut s'imaginer à quel point cela est difficile pour celui qui partage la vie du doctorant, et je ne la remercierai jamais assez pour tout le soutien apporté !

Je dédie cette thèse à ma famille.

Table des matières

Table des figures 1

Liste des tableaux 5

Partie I Introduction générale 7

Chapitre 1
Contexte du projet

1.1 Introduction 9
1.2 Origine du projet 11
1.3 Problématiques générales 13
1.3.1 Appréhender le travail des experts documentalistes 14
1.3.2 Modélisation d'un territoire à partir de ressources textuelles indexés 14
1.4 Travaux existants 15
1.5 Contribution 17
1.5.1 Représentation sémantique : le choix de l'ontologie 17
1.5.2 Méthodologie globale 18
1.6 Organisation du mémoire 21

Partie II Etat de l'art 23

Chapitre 2
L'indexation : une forme d'annotation dans les centres documentaires

2.1 L'annotation de documents 26
2.1.1 Définitions 26
2.1.2 Le travail d'annotation des experts et ses spécificités 28
2.2 Les différents types de vocabulaires contrôlés 29
2.2.1 Synthèse des vocabulaires contrôlés les plus utilisés 29
2.2.2 L'ontologie : une structure formelle pour représenter des connaissances 32
2.3 Discussions 38

Chapitre 3
Construction d'ontologies

3.1 « Critères » reconnus pour construire une ontologie 41
3.2 Méthodes de construction d'ontologies 43
3.2.1 Construction manuelle d'ontologies 44
3.2.2 Construction semi-automatique et automatique d'ontologies . 48
3.3 Les langages autour de l'ontologie 54
3.3.1 Langages pour la représentation de connaissances 54
3.3.2 Langages associés pour la gestion des ontologies 54
3.3.3 Les langages à balises définis autour du Web Sémantique . . . 58
3.4 Discussion 65

Chapitre 4
Définition du domaine cible : le territoire

4.1 Introduction 69
4.2 Le territoire 70
4.2.1 Le territoire des géographes 70
4.2.2 Le territoire comme outil de la géographie sociale 71
4.2.3 L'architecture du territoire 72
4.2.4 Les composantes d'un territoire 73
4.2.5 Implications dans nos travaux 74
4.2.6 Le territoire en géomatique 75
4.3 L'information géographique dans les textes 76
4.3.1 Définitions et modélisation 76
4.3.2 Le TALN pour l'extraction d'informations géographiques . . . 77

4.4 Construction d'ontologies géographiques 80
4.4.1 Les différents types d'ontologies géographiques 80
4.4.2 Le « cas » Territoire 81
4.5 Discussion 82

Chapitre 5
Synthèse et proposition d'un modèle du territoire

5.1 Synthèse de l'état de l'art 85
5.1.1 Un fonds documentaire annoté comme base de travail 87
5.1.2 Choix de l'ontologie pour la représentation de la connaissance 87
5.2 Le territoire dans un espace documentaire 88
5.2.1 Définitions 88
5.2.2 Noyau de modèle du territoire 88
5.2.3 La composante spatiale 90
5.2.4 La composante temporelle 91
5.2.5 La composante thématique 92
5.2.6 Modèle pour la représentation d'un territoire 93

Partie III Contribution **99**

Chapitre 6
Méthodologie opérationnalisée pour l'émergence d'une ontologie d'un territoire

6.1 Une première ontologie construite à partir de la connaissance des bibliothécaires 103
6.1.1 Analyse des besoins 104
6.1.2 Constitution du corpus 105
6.1.3 Analyse de la connaissance experte 106
6.1.4 Normalisation en réseau sémantique 108
6.1.5 Formalisation du réseau sémantique 120
6.1.6 Tests et Bilans 121
6.2 TALN pour la représentation d'un territoire 123
6.2.1 Chaîne de TAL pour l'indexation d'entités géographiques . . . 124

6.2.2 Application sur le contenu des notices descriptives 125
6.2.3 Application sur le contenu des documents 127
6.3 Enrichissement de l'ontologie à partir du contenu des documents . . . 129
6.3.1 Enrichissement de l'ontologie par des concepts 130
6.4 Discussion 131

Partie IV Implémentations **133**

Chapitre 7
Création d'une ontologie légère décrivant le territoire des Pyrénées

7.1 Chaîne de traitement TERRIDOC 137
7.1.1 Description de TERRRIDOC 137
7.1.2 Technologies et langages utilisés 138
7.2 Construction d'une première ontologie à partir de la connaissance structurée des bibliothécaires 139
7.2.1 Constitution du jeu d'essai 139
7.2.2 Module (1) d'extraction de la connaissance experte : création du vocabulaire vocTerridoc 144
7.2.3 Module (2) de structuration des termes : normalisation en thésaurus tTerridoc 145
7.2.4 Module (3) de transformation du thésaurus tTerridoc en une première ontologie légère 152
7.2.5 Module (4) d'instanciation de l'ontologie pour l'émergence d'un territoire 154
7.3 TALN pour la représentation d'un territoire 156
7.3.1 Modules (5 et 6) pour l'enrichissement de la représentation d'un territoire à partir du travail d'indexation d'experts 158
7.3.2 Modules (7 et 8) pour l'enrichissement de la représentation d'un territoire à partir du contenu des documents 160
7.4 Enrichissement de l'ontologie à partir du contenu des documents . . . 161

Chapitre 8
Usages liés à notre approche

8.1 Géonto : la méthodologie Terridoc pour l'enrichissement d'une ontologie de domaine . . . 164
8.1.1 Contribution au projet Geonto . . . 164
8.1.2 Une chaine de traitement pour l'indexation de documents textes 165
8.1.3 Enrichissement de l'ontologie géographique . . . 168
8.2 TERRIDOCViewer : Une application pour l'industrie permettant de naviguer dans des fonds documentaires . . . 170
8.2.1 Analyse des besoins . . . 170
8.2.2 Architecture de l'application et Technologies . . . 179
8.2.3 Schéma des bases de données . . . 182
8.2.4 Fonctionnalités implémentées . . . 185
8.2.5 Fonctionnement de l'application . . . 191
8.2.6 Premiers bilans . . . 192
8.3 Vers une application d'aide à l'indexation de documents pour les experts 192
8.3.1 Un premier contrôle . . . 193
8.3.2 Propositions : vers une aide à la correction semi-automatisée . 194
8.3.3 premiers bilans liés à l'analyse du travail d'indexation . . . 195

Partie V Conclusion générale **197**

Chapitre 9
Synthèse et perpectives

9.1 Synthèse générale . . . 199
9.1.1 Modélisation d'un territoire . . . 200
9.1.2 Choix de l'ontologie pour la représentation sémantique d'un territoire . . . 202
9.1.3 Découverte incrémentale d'un territoire à partir de documents annotés . . . 203
9.2 Usages mis en place et perspectives applicatives . . . 207
9.2.1 Application TerridocViewer pour la RI . . . 207
9.2.2 Enrichissement d'ontologie géographique et améliorations de l'indexation spatiale de fonds documentaires . . . 209
9.3 Perspectives scientifiques . . . 209

9.3.1 Prise en compte des relations spatiales pour l'enrichissement de l'ontologie d'un territoire 209
9.3.2 Création d'une ontologie d'un territoire adaptée à plusieurs fonds documentaires 211
9.3.3 Perspective à plus long terme : Application de la méthodologie Terridoc dans une autre langue 211

Bibliographie **213**

Table des figures

1.1 Axes de recherche de la thèse . . . 13
1.2 Méthodologie générale TERRIDOC pour la construction d'une ontologie d'un territoire . . . 19
1.3 Organisation du mémoire . . . 21

2.1 Travail d'indexation du bibliothécaire . . . 28
2.2 Représentation formelle de la relation d'inclusion . . . 30
2.3 Exemple de taxonomie dans le domaine des Mathématiques . . . 30
2.4 Exemple de thésaurus dans le domaine des Mathématiques . . . 32
2.5 Représentation formelle mathématique de la relation de transitivité . . . 39

3.1 Comparaison de méthodologies [FLGPJ97] . . . 45
3.2 Les couches du Web Sémantique . . . 59
3.3 Exemple de triplet RDF . . . 61
3.4 Extrait d'un concept RDF . . . 61
3.5 Exemple d'un fichier RDF/XML . . . 61

4.1 Définition du Modèle Pivot [Les07] . . . 77

5.1 Positionnement de nos travaux au sein de l'état de l'art présenté . . . 86
5.2 Noyau de modèle du territoire . . . 89
5.3 Définition de la composante spatiale . . . 90
5.4 Définition de la composante temporelle . . . 91
5.5 Définition de la composante thématique . . . 92
5.6 Modèle du territoire . . . 93
5.7 Définition d'un territoire vis à vis du thésaurus . . . 94
5.8 Définition d'un territoire vis à vis de l'ontologie . . . 96
5.9 Exemple schématisé d'une ontologie d'un territoire . . . 96

6.1 Rappel de la méthodologie générale TERRIDOC pour l'émergence d'une ontologie d'un territoire . . . 102
6.2 Extrait de notice descriptive 1 . . . 105
6.3 Extrait de notice descriptive 2 . . . 106
6.4 Forme d'un Patron Structurel . . . 107

6.5 Exemples de concepts extraits des notices descriptives . . . 107
6.6 Extrait de thésaurus avec le terme « Stations climatiques, thermales, etc. » 109
6.7 Extrait de la structuration en thésaurus à titre d'exemple . . . 112
6.8 Patron Structurel permettant d'enrichir la structure de tTerridoc à partir des relations génériques . . . 113
6.9 Extrait de notice descriptive 3 . . . 113
6.10 Enrichissement du vocabulaire identifié par les termes « génériques » . . . 114
6.11 Extrait de l'ontologie générée . . . 120
6.12 Chaîne de traitement d'informations géographiques dans des documents textuels . . . 125
6.13 Extrait de l'ontologie enrichie . . . 127
6.14 Extrait d'un fichier texte indexé par la chaîne de TAL . . . 128
6.15 Algorithme général d'enrichissement de l'ontologie . . . 130

7.1 Technologies utilisées pour l'implémentation de la méthodologie Terridoc . 136
7.2 Chaîne de traitement TERRIDOC . . . 137
7.3 Fonds documentaire de la MIDR . . . 140
7.4 Notice descriptive exemple réalisée par la MIDR . . . 141
7.5 Notices RAMEAU de la vedette « Stations climatiques, thermales, etc. » . 142
7.6 Module de création du jeu d'essai . . . 143
7.7 Définition de la liste de termes . . . 144
7.8 Extrait des termes provenants des notices descriptives . . . 145
7.9 Enrichissement du vocabulaire (création du thésaurus) . . . 146
7.10 Extrait de la liste de termes formalisée en SKOS . . . 147
7.11 Prise en compte des termes rejetés selon le formalisme SKOS . . . 148
7.12 Extrait SKOS du thésaurus tTerridoc intégrant la relation hiérarchique . 149
7.13 Extrait SKOS du thésaurus tTerridoc intégrant la relation associative . . 149
7.14 Structuration du thésaurus tTerridoc via les vedettes génériques . . . 150
7.15 Visualisation sous Protégé d'un extrait du thésaurus tTerridoc . . . 151
7.16 Formalisation du concept Station_climatique,_thermale,_etc en OWL-Lite 152
7.17 Formalisation du concept Barèges_(Hautes-Pyrénées) en OWL-Lite . . . 153
7.18 Formalisation du concept Barèges_(Hautes-Pyrénées) en OWL-Lite . . . 154
7.19 Formalisation des entités spatiales en OWL . . . 155
7.20 Visualisation sous Protégé d'un extrait de l'ontologie formalisée en OWL . 155
7.21 Chaîne de traitement linguistique implémentée sous Linguastream . . . 156
7.22 Règle de marquage des noms toponymiques . . . 157
7.23 Règle de grammaire DCG pour une analyse sémantique . . . 157
7.24 Patrons lexico-syntaxique pour capter les entités géographiques . . . 158
7.25 Extrait du traitement Linguatream . . . 159
7.26 Résultats de l'application de la chaîne de TAL sur les notices descriptives 159
7.27 Résultats de l'application de la chaîne de TAL sur les documents . . . 160
7.28 Résultats de l'application de l'algorithme d'enrichissement sur les documents 161

8.1 Démarche d'enrichissement d'ontologie dans Géonto . . . 166

8.2 Extrait de l'ontologie géographique . . . 167
8.3 Exemple de notice descriptive dans RAMEAU décrivant le terme « Grottes » 168
8.4 Toponymes candidats et termes associés . . . 169
8.5 Diagramme général des cas d'utilisations . . . 171
8.6 Navigation dans le graphe de termes . . . 172
8.7 Structure du graphe . . . 173
8.8 Visualisation du graphe . . . 174
8.9 Sélection d'un terme . . . 175
8.10 Diagramme de séquence pour la navigation dans le graphe . . . 176
8.11 Visualisation des documents . . . 177
8.12 Consultation d'une fiche RAMEAU . . . 178
8.13 Paramétrage de l'affichage . . . 179
8.14 Architecture de l'affichage . . . 180
8.15 Architecture de l'affichage selon le framework MVC Struts 2 . . . 181
8.16 Structure de la base de données *thesaurus* . . . 183
8.17 Schéma de la base de données *bdterridoc* . . . 184
8.18 Schéma de la base de données *paramétrage* . . . 185
8.19 Visualisation du graphe . . . 186
8.20 Légende du graphe . . . 186
8.21 Graphe au niveau N+3, N-3 . . . 187
8.22 Eléments de paramétrage . . . 188
8.23 Barre de recherche. (a) Auto complétion. (b) Terme non proposé avec l'auto complétion. . . . 189
8.24 Liste hiérarchique en mode fonds documentaire . . . 190
8.25 Fiche RAMEAU dans TERRIDOCViewer pour le terme « Glace de mer » 190
8.26 Affichage d'un document . . . 191
8.27 Résultat du traitement de vérification pour validation de l'indexation . . . 194

9.1 Modélisation du territoire . . . 202
9.2 Méthodologie incrémentale de construction d'ontologie d'un territoire à partir de documents annotés . . . 204
9.3 Chaîne de traitement d'informations spatiales dans les documents textuels 206

Liste des tableaux

2.1 Relations entre termes dans un thésaurus (d'après [Her05]) 31

6.1 Gestion des termes vedettes 109
6.2 Gestion des termes rejetés 110
6.3 Gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc 111
6.4 Gestion des relations associatives entre termes du vocabulaire vocTerridoc 111
6.5 Regroupement des termes en concepts 115
6.6 Structuration via les relations hiérarchiques 116
6.7 Structuration via les relations associatives 118
6.8 Instanciation de l'ontologie pour la description un territoire 119
6.9 Instanciation de l'ontologie à partir des notices descriptives 126
6.10 Enrichissement de l'ontologie par des concepts 130

Première partie

Introduction générale

Chapitre 1

Contexte du projet

Sommaire

1.1 Introduction 9
1.2 Origine du projet 11
1.3 Problématiques générales 13
1.3.1 Appréhender le travail des experts documentalistes 14
1.3.2 Modélisation d'un territoire à partir de ressources textuelles indexés 14
1.4 Travaux existants 15
1.5 Contribution 17
1.5.1 Représentation sémantique : le choix de l'ontologie 17
1.5.2 Méthodologie globale 18
1.6 Organisation du mémoire 21

1.1 Introduction

Que ce soit au niveau international ou national, de nombreux travaux de numérisation massive de collections voire de productions directes numériques (romans en version unique électronique, prises de vues par les photographes, etc.) sont mis en oeuvre. Des collections numériques sont ainsi constituées, d'un volume souvent très considérable et qu'il importe par conséquent, aux vu des investissements réalisés, de pérenniser. Le fait que les originaux existent, n'exonère pas de la conservation des copies numériques, en raison de leur propre dégradation, patente notamment pour les supports analogiques textuels. Outre la conservation, cette phase de numérisation permet également de prévoir un accès simplifié à l'information via des outils informatiques.

Au niveau international, nous pouvons notamment citer le projet Google Books [1] (renommé Google Livres en français) proposant un service en ligne permettant d'accéder à des livres numérisés (plus de 15 millions de livres numérisés à ce jour). Un autre

1. http ://books.google.fr/books

projet auquel la France participe, par l'intermédiaire de la Bibliothèque nationale de France (BnF), est le projet de Bibliothèque numérique européenne nommé Europeana [2]. Europeana est une bibliothèque numérique lancée en novembre 2008 par la Commission européenne. Elle fait suite au prototype de bibliothèque en ligne développé par la (BnF) sur la base de sa propre bibliothèque numérique Gallica [3], dans le cadre du projet de bibliothèque numérique européenne.

Au niveau national, Gallica propose aujourd'hui à la consultation en ligne plus d'un million de documents de type texte (livres, périodiques, revues et journaux), image, son et vidéo avec un rythme de 1 500 documents numérisés par jour. Au niveau local, les Bibliothèques-Médiathèques ont aussi la double fonction de service documentaire encyclopédique et de mémoire historique et culturelle de leur commune et plus largement de leur région. Ces deux missions doivent être envisagées dans la double dimension de desserte de la population locale et du rayonnement des communes et régions concernées. Toutefois, même si ces organismes renferment des corpus documentaires conséquents qui deviennent de plus en plus facilement disponibles au format électronique, leur accessibilité afin de comprendre ces données reste encore problématique.

Comme l'ensemble des bibliothèques et des médiathèques en France, la MIDR [4] de Pau (Pyrénées-Atlantiques) a comme souci majeur de structurer au mieux la connaissance présente dans ses fonds documentaires afin d'en faciliter l'accès et le partage par une large communauté d'utilisateurs. En effet, la mise à disposition de ces fonds doit permettre de proposer une offre de services, aussi bien locaux que régionaux, destinée à informer, éduquer et divertir les populations locales. Ces services peuvent être aussi bien proposés au grand public qu'à des publics spécifiques pour des applications particulières (formation de groupes scolaires sur l'histoire et la géographie locale, guides touristiques pour des personnes curieuses de découvrir la région, etc.). Une caractéristique importante de ce type de fonds documentaire est qu'ils contiennent d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire. Il est primordial de valoriser ces spécificités territoriales pour répondre à ces objectifs d'information et d'éducation. [Hil06] estime qu'au moins 70% des documents textuels contiennent des références à des lieux géographiques sous forme de toponymes. Une expérimentation réalisée au sein de notre équipe de recherche et notamment par [Les07] sur un extrait de corpus texte constitué de récits de voyages et de périodiques de la MIDR a révélé une connotation géographique prédominante. Sur 10 livres extraits du corpus (soit 600 000 mots), près de 10 000 entités nommées ont une connotation spatiale.

Cette dimension territoriale se symbolise dans les documents par une fréquence importante d'entités mentionnées (fleuves, villes, édifices publics, etc.), de faits relatés (événements politiques, sportifs, etc.) ou d'observations décrites (de nature sportive, par exemple) qui sont, d'une manière ou d'une autre, liés à un nom de lieu et/ou une référence temporelle dans un espace géographique. Dans la plupart des cas, l'informa-

2. http ://www.europeana.eu/portal/
3. http ://gallica.bnf.fr/
4. Médiathèque Intercommunale à Dimension Régionale

tion géographique n'est pas ou peu utilisée dans les systèmes informatiques des centres documentaires (notamment lors de la recherche de documents via des outils tels que Google Maps [5], Bing [6], etc.), alors qu'elle peut s'avérer incontournable dans une volonté de valoriser un territoire. Une étude bibliographique réalisée dans les domaines de la géographie et de la sociologie nous permet de proposer section 5.2 page 88 une définition de ce que nous entendons par « territoire ». Ce type d'informations peut en effet servir dans des applications diverses telles que la recherche d'informations à des fins pédagogiques pour une population locale soucieuse de découvrir l'histoire de sa région, à des fins touristiques pour des visiteurs souhaitant découvrir un espace géographique ou encore à des fins d'analyse et de validation pour des experts bibliothécaires soucieux de mettre à disposition du grand public des fonds documentaires accompagnés d'un travail d'indexation de qualité.

1.2 Origine du projet

Cette thèse fait l'objet d'une collaboration, initiée lors du stage recherche [Ker06], entre l'entreprise Document Image Solutions [7] (DIS) localisée à Bidart et le Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour (LIUPPA) localisé à Pau.

Dans le cadre du projet « Pyrénées Itinéraires Virtuels (PIV) » que nous avons mené en partenariat avec la MIDR de Pau (Pyrénées-Atlantiques), l'équipe DESI (Document Electronique, Sémantique et Interaction) du LIUPPA, devenue depuis l'équipe Traitement des Interactions et des Informations (T2I), a tenté de répondre à différentes problématiques liées à l'extraction, à l'analyse et à l'exploitation d'informations géographiques provenant de documents textes à des fins de recherche documentaires adaptées à ce type d'informations. Le projet PIV reposait sur un domaine de recherche dans lequel d'ores et déjà sont proposées des techniques et des méthodes autour de tâches bien identifiées d'extraction et de recherche d'informations respectivement connues sous les acronymes d'IE (Information Extraction) et de GIR (Geographic Information Retrieval). Différentes approches furent développées s'appuyant sur des méthodes de compréhension ciblée du « contenu » des documents généralement appréhendés de manière partielle pour des raisons d'efficacité évidente. Les travaux au sein de l'équipe, parmi lesquels nous pouvons notamment citer [LSG06, SGLL07, LGN07] ont apporté des résultats fort intéressants dans le domaine de l'indexation automatique de documents (marquage des informations selon les composantes spatiales et temporelles) ainsi que dans la recherche de documents textuels incluant des outils de visualisation sur cartes géographiques [GSE+08, EMC06, MESB08] et chronologies [LPLGS07]. Les problématiques de structuration et modélisation de ces informations ainsi que la recherche de l'information structurée dans un fonds documentaire hybride (textes, images, sons et vidéos) territorialisé tel que celui proposé par la MIDR sont, après le projet PIV, encore bien présentes. Le travail de thèse s'intègre donc pleinement dans la continuité du projet PIV

5. http ://maps.google.fr/
6. http ://www.bing.com/ ?cc=fr
7. http ://www.docimsol.eu

et a pour objet l'extraction, la gestion et la visualisation de connaissances dans un fonds documentaire territorialisé.

Acteur de la collaboration, DIS édite des composants métiers et intègre des solutions sur mesure dans les domaines de la lecture automatique, de l'imagerie, de la gestion électronique de documents et de la reconnaissance vocale. DIS couvre les principaux éléments clés du cycle de vie documentaire en proposant des solutions en lecture automatique de documents, en indexation automatique plein texte et en application Web pour la consultation de fonds. Si les solutions actuelles apportent des résultats intéressants, des évolutions peuvent être envisagées en ce qui concerne la structuration de fonds documentaire, notamment afin d'améliorer les phases de classification et de recherche documentaire. Les besoins sont notamment importants en ce qui concerne l'identification et l'extraction d'entités géographiques pour remplir la base de connaissances concernant un territoire donné. Cette thèse doit donc permettre de proposer des prototypes avancés offrant un moyen efficace de structurer la connaissance extraite de fonds documentaires volumineux ainsi que des outils simples et attractifs permettant la navigation dans un territoire à travers les documents, aussi bien dans le monde bibliothécaire que dans le monde industriel.

Pour l'ensemble des acteurs que nous venons de présenter et qui collaborent dans ces travaux, la nécessité est de mettre à disposition des utilisateurs des outils pour valoriser les ressources décrivant un territoire donné. Nos travaux de recherche s'inscrivent dans cette volonté de **valoriser un territoire implicitement décrit par un fonds documentaire en tentant de proposer une représentation sémantique intégrant les spécificités territoriales contenues dans les documents**.

Nos travaux s'articulent autour du large domaine qu'est la gestion de connaissances appliquée au domaine de la géographie, domaine connu sous le nom de géomatique. La gestion de connaissances est l'ensemble des initiatives et des techniques permettant d'identifier, d'analyser, d'organiser, et de partager des connaissances entre les membres des organisations. La géomatique, domaine d'application de notre démarche, a pour objet la gestion des données à référence spatiale et qui fait appel aux sciences et aux technologies reliées à leur acquisition, à leur stockage, à leur traitement et à leur diffusion. Nous schématisons l'intégration de nos travaux dans ce domaine par la figure 1.1.

Nous utilisons pour nos expérimentations un fonds documentaire hybride indexé fourni par la MIDR mettant en avant le territoire des Pyrénées entre le XVIII^{ème} et le XIX^{ème} siècle. L'indexation est réalisée en s'appuyant sur le langage d'indexation RAMEAU qui intègre un thésaurus qui couvre l'ensemble des disciplines scientifiques et contient aussi les termes traitant des loisirs, des arts, etc. Cependant, notre démarche se veut générique et applicable à tout fonds documentaire, en prenant en compte les possibles variations dans le travail d'indexation. Afin d'appréhender au mieux ce travail d'indexation, nous avons travaillé en collaboration avec les experts bibliothécaires de la MIDR ainsi qu'avec les responsables du Centre national RAMEAU qui fait partie, au sein de la BnF, du département de l'information bibliographique et numérique. Le centre national RAMEAU, responsable de la gestion intellectuelle du langage d'indexa-

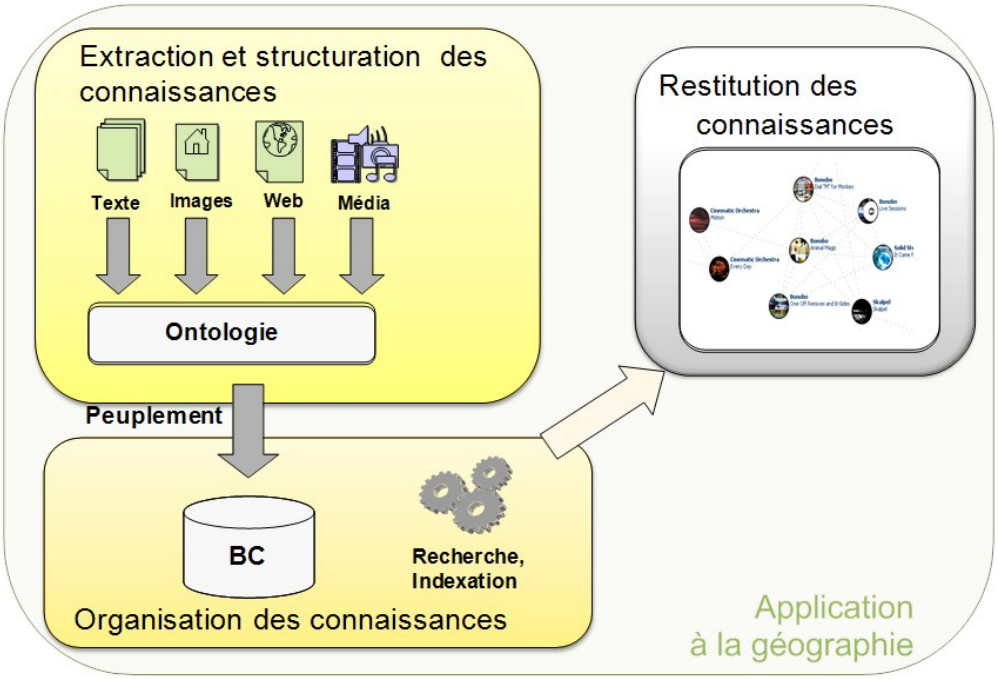


FIGURE 1.1 – Axes de recherche de la thèse

tion, nous fournit la ressource RAMEAU. Il voit dans ces travaux un moyen attractif de valoriser le travail de gestion du thésaurus RAMEAU et une possibilité intéressante de mettre en place notre approche dans d'autres centres documentaires.

1.3 Problématiques générales

L'intérêt de disposer d'un fonds documentaire et de pouvoir ensuite proposer à des utilisateurs, experts bibliothécaires ou grand public, d'accéder aux informations nécessaires pour leur activité est primordial. Dans le cadre de nos travaux menés en partenariat avec la MIDR de Pau, nous avons identifié avec les experts bibliothécaires de multiples besoins liés à leur travail d'indexation ainsi qu'à leur volonté de faire découvrir un territoire à tout type de visiteurs à travers les corpus indexés. Parmi ces besoins, il y a notamment :

- la nécessité de posséder une base de connaissances représentant le travail d'annotation des experts ;
- la nécessité d'une ressource pour une aide à la validation du travail d'indexation ;
- la possibilité d'échanger des bases de connaissances d'experts de différents centres documentaires ;
- et la mise en place d'une base de connaissances pour la découverte du fonds documentaire pour tout type d'utilisateurs.

Dans ce sens, la première difficulté est d'appréhender le travail d'indexation des experts bibliothécaires ainsi que le vocabulaire contrôlé utilisé (*cf.* chapitre 2 page 25). La deuxième difficulté est d'identifier, d'extraire et de structurer des informations pour en produire une représentation sémantique (*cf.* chapitre 3 page 41) qui se rapporte à un territoire (*cf.* chapitre 4 page 69).

1.3.1 Appréhender le travail des experts documentalistes

Bien que fastidieux, le travail manuel d'annotation est indispensable pour les centres documentaires tels que les bibliothèques et les médiathèques qui ont pour objectif premier de fournir des informations de qualité pour faciliter le travail de recherche documentaire. Si l'indexation automatique apporte des résultats intéressants pour traiter des gros volumes de documents tout en limitant les coûts, le travail manuel d'indexation reste incontournable lorsque l'on préconise avant tout un travail de haute précision et de qualité [CDMV89], [APC01], [Sav05]. L'expert discerne plus aisément l'information essentielle des éléments secondaires d'un document et il est plus sélectif dans le choix des termes pour décrire son contenu. Aussi, l'indexation d'œuvres volumineuses constituant les fonds documentaire de bibliothèques et médiathèques nécessite une lecture complète et un travail d'analyse difficilement automatisable. Ce travail d'analyse prend notamment en compte la maitrise de la ressource de type vocabulaire contrôlé sur laquelle s'appuient les experts pour sélectionner les termes décrivant le contenu des documents. Comme pour beaucoup de centres documentaires, le travail d'indexation réalisé par les bibliothécaires de la MIDR nécessite des connaissances du langage d'indexation RAMEAU et l'utilisation de la ressource thésaurus associée. Le thésaurus est un outil qui fait partie de la famille des vocabulaires contrôlés permettant l'accès par sujet aux catalogues et aux bases de données bibliographiques [Nie03]. Un vocabulaire contrôlé est notamment caractérisé par le fait qu'un terme le constituant n'a qu'un seul sens précis et cet élément sera le seul à avoir ce sens, impliquant le contrôle de la synonymie, de l'homonymie et de la polysémie.

Une fois l'indexation réalisée sur un ensemble de documents, la relecture des notices est inévitablement longue et fastidieuse et il est difficile, voire impossible pour les experts, de se représenter l'ensemble du travail réalisé et encore moins du territoire décrit dans ces documents. La première difficulté consiste donc à définir une méthodologie et des outils pour définir automatiquement une représentation sémantique synthétisant le travail de l'ensemble des experts. Dans ce travail d'indexation, les références géographiques sont utilisées comme les autres termes et elles ne sont pas enrichies par des éléments caractéristiques nous permettant d'identifier un territoire.

1.3.2 Modélisation d'un territoire à partir de ressources textuelles indexés

Lors de la phase d'annotation de documents, les bibliothécaires décrive le contenu des documents mais ils n'ont pas pour objectif de mettre directement en avant un territoire.

Cependant, les fonds documentaires mis à disposition par un grand nombre de centres culturels sont constitués d'une quantité importante de documents territorialisés. Nous entendons ici par document territorialisé, tout document qui décrit/raconte un territoire : récits de voyage, contes, cartes postales, etc. Le problème est alors de savoir comment identifier les éléments relatifs au territoire dans les documents et comment ensuite les utiliser pour proposer une représentation de ce territoire.

1.4 Travaux existants

La notion de territoire est une notion complexe qui intéresse de nombreux chercheurs dans différents domaines ayant des actions liées à l'implantation de l'Homme tels que la géographie, l'économie, la médecine, la géologie, la sociologie, le droit, etc. L'encyclopédie Universalis définit le mot territoire comme l'« étendue de la Terre sur laquelle vit un groupe humain ». Cependant, sa définition est encore sujet à discussion comme le montrent ces différents travaux [Pio92, Sch94, LL03, GV04, DMB05, Gui07] et elle évolue en fonction des domaines d'activités.

Dans nos travaux, nous cherchons à identifier les éléments communs aux différentes définitions proposées dans l'ensemble des domaines pour en faire émerger une vue « grand public » d'un territoire décrit dans un fonds documentaire annoté. Nous nous appuyons tout d'abord sur les travaux des géographes, pour qui le territoire est une notion centrale, afin de suggérer une définition, reconnue quel que soit le domaine. [Gui07] met en avant la relation entre l'Homme et la Terre en définissant la notion de **territoire** comme la superposition d'un espace et de pratiques sociales. [DMB05] indique que le territoire regroupe et associe des lieux. [DMB05] indique également que le territoire permet d'appréhender le phénomène qu'est le contact vécu de l'homme avec le milieu. Dans ce cadre, le contact vécu de l'Homme fait référence à la composante thématique et également à la composante temporelle. Le milieu fait lui référence à la composante spatiale et correspond à un ensemble de lieux qui forment un espace. [Ent96, BE98] mettent également en avant la relation entre les composantes sujet et temporelle avec la composante spatiale ; la composante spatiale étant vue alors comme un ensemble de lieux formant un espace défini selon la composante sujet.

A notre connaissance, il n'existe pas de définition formelle de la notion de territoire que l'on puisse exploiter ensuite dans notre approche pour construire une vue territoriale à partir d'un ensemble de documents. Nous proposons dans ce mémoire une synthèse succincte des travaux visant à définir cette notion complexe qu'est le territoire et nous tentons ensuite de tirer parti de ces définitions dans notre domaine pour décrire les approches permettant d'identifier, d'extraire et de structurer un ensemble d'informations caractéristiques d'un territoire à partir de documents textes.

Pour construire une représentation d'un territoire, nous nous intéressons plus particulièrement aux méthodologies permettant de créer une ontologie de façon automatique, intégrant néanmoins en fin de traitement une étape de validation par des experts. Ces travaux émergent comme un sous-domaine de l'ingénierie des ontologies. Un moyen très largement utilisé pour atteindre cet objectif est de partir d'éléments préexistants dans

le domaine. [MS01] distinguent différents types d'approches en fonction du support sur lequel elles se basent : ressources structurées de type vocabulaire contrôlé (taxonomies, thésaurus, etc.), normes ou fragments d'ontologie préexistants, ou encore des corpus textuels. Certains travaux reposent sur l'analyse de textes tels que ARCHONTE [Bac00], TERMINAE [BS99, AGBS00, SBAG02, BAGC04], KAON [VOS03], Text2Onto [CV05] afin d'aider à la construction automatique ou semi-automatique des ontologies.

De notre côté, nous cherchons à construire une ontologie d'un domaine cible à partir de ressources textuelles peu ou pas structurées et d'un vocabulaire contrôlé de type thésaurus. Nous nous intéressons plus particulièrement aux méthodes permettant de transformer un vocabulaire contrôlé de type thésaurus en ontologie du domaine [WSWS01, SLL+04, Her05, CHGM06]. Nous nous rapprochons plus précisément des travaux de [CHGM06] qui, sur la base de la méthodologie TERMINAE, propose de transformer un thésaurus du domaine cible en une ontologie. Le thésaurus utilisé permet alors d'identifier un ensemble de concepts et de relations entre ces concepts. Cependant, seul, il ne permet pas réellement de définir une représentation précise d'un domaine cible comme un territoire.

En géomatique, des travaux récents parmi lesquels nous pouvons citer [AM10,RLB+04, BBG+07] s'appliquent à constuire une ontologie géographique de domaines cibles. Cependant, il ne semble pas exister d'approche permettant de construire une ontologie du territoire à partir d'une ressource documentaire préexistante. En effet, parmi l'ensemble des ressources structurées (thésaurus ou ontologies géographiques), beaucoup sont organisées de façon administrative. Par ailleurs, elles restent très générales et ne permettant pas de représenter de façon suffisament précise le territoire décrit dans des documents. Nous nous rapprochons des travaux de [COL09, Sal09] qui s'appliquent actuellement à construire à partir de textes une ontologie du territoire mais ces travaux cherchent à représenter les évolutions économiques d'une région et donc un sous ensemble de ce qu'est un territoire.

Les informations manipulés en géomatiques sont appelées informations géographiques, et sont définies comme des molécules formées d'une composante spatiale, d'une composante temporelle et d'une composante thématique ou phénomène [Gal01,UTC04,PSA07]. Nous présentons dans ce mémoire une synthèse des travaux visant à identifier et extraire l'information géographique de un corpus documentaire [Van86,Bor98,Tal00,Par70,Sto02, AHV05, Les07]. Nous souhaitons nous appuyer sur les travaux réalisés au sein de notre laboratoire dans le cadre du projet PIV [SBLG07]. Ces travaux proposent une chaîne de traitement linguistique pour le traitement d'informations spatiales. La chaine proposée intègre une étape d'identification et d'annotation des entités nommées (EN) constituant une information spatiale avec une phase intérmédiaire de désambiguïsation d'EN à partir d'un lexique externe. Nous nous rapprochons des travaux de [BP06] qui présentent également une approche de désambiguïsation d'EN mais le lexique utilisé, en l'occurrence la ressource encyclopédique Wikipédia, est très générale et n'est pas réellement représentative du domaine cible. Dans nos travaux, nous tentons de traiter finement les EN en nous appuyant sur un lexique structuré représentant le domaine cible pour faciliter le marquage des thèmes dans le texte. L'utilisation d'un lexique structuré doit également nous

permettre d'identifier plus précisément les relations entre la composante thématique et les composantes spatiales et temporelles formant l'information géographique.

1.5 Contribution

Hypothèse de départ : Nous faisons l'hypothèse qu'en utilisant un point de vue géographique pour modéliser un ensemble de ressources terminologiques utilisées pour indéxer un fonds documentaire, il est alors possible de faire émerger la représentation du territoite qui y est implicitement décrit.

Nous nous appuyons pour cela sur le travail d'indexation d'experts réalisé à partir d'un vocabulaire contrôlé, contenant des termes porteurs de sens et des relations entre ces termes, car, comme [Sve86], [APC01], nous pensons qu'il apporte un plus sémantique indéniable à la représentation du domaine étudié. En effet, ce travail d'expertise visant à décrire de façon très générale le document ainsi que son contenu, réalisé sur la base d'un vocabulaire contrôlé qui fait office de référence dans le monde des centres documentaires en France et maintenant dans des pays étrangers, est une source de connaissance intéressante dans notre « quête du sens ». Le territoire est à l'origine implicitement décrit par les documents et notices descriptives attachées constituant le fonds. Notre objectif est de créer une première représentation sémantique et d'y intégrer ensuite les références géographiques contenues dans les documents.

Comme indiqué figure 1.1 (*cf.* page 13), notre contribution fait le lien entre les domaines *Extraction et structuration des connaissances* et *Organisation des connaissances*. La partie *Restitution des connaissances* fait référence dans nos travaux à l'utilisation des résultats (intérmédiaires ou finaux) de notre approche dans des interfaces de recherche documentaire.

Les références géographiques que nous souhaitons identifier sont définies comme des « entités géographiques ». Nous reprenons la définition provenant de la géomatique [UTC04], [Gai01] indiquant qu'une entité géographique est une « molécule composée non seulement d'une composante spatiale, mais aussi d'une composante temporelle et d'une composante thématique ou phénomène ». L'exploitation d'une entité géographique, qu'elle soit une donnée ou une métadonnée, nécessite des formalismes de représentation et de raisonnement adaptés aux particularités de ce type d'information et, pour faciliter les échanges, compatibles avec le Web sémantique [BHL01].

1.5.1 Représentation sémantique : le choix de l'ontologie

Les outils permettant de modéliser la connaissance sont nombreux. Nous pouvons notamment citer les modèles à base de logique, ceux s'appuyant sur les frames, ceux sur les graphes conceptuels, ceux basés sur la logique de description ou encore ceux de type vocabulaires contrôlés qui ont évolué à travers le Web Sémantique. Parmi ces derniers, nous pouvons citer la taxonomie, le thésaurus ou encore l'ontologie (du plus simple au plus spécifique).

Dans un premier temps, notre choix de représentation s'est porté sur le thésaurus. En effet, il fait partie de la famille des vocabulaires contrôlés qui sont utilisés par les experts pour indexer les documents et ce type de modèle apparaît ainsi comme une base intéressante de connaissances structurées pour créer une première représentation sémantique d'un fonds documentaire. Aussi comme le vocabulaire contrôlé, le thésaurus permet de définir un ensemble de termes caractérisant le contenu du fonds documentaire ; comme la taxonomie, il permet de représenter les relations de types spécifique et générique existantes entre ces termes ; et enfin il donne la possibilité de représenter des relations sémantiques entre termes connexes.

Cependant, le thésaurus reste limité dans la représentation sémantique car il ne permet pas d'ajouter des propriétés à des éléments constituants (termes, relations, etc.). Les relations de type « terme associé » ne peuvent être explicitées pour mettre en avant des caractéristiques géographiques entre deux termes. Le formalisme de modélisation qui s'est alors imposé, pour associer du sens aux ressources, est celui des ontologies [Gru93]. Nous proposons de transformer le premier thésaurus obtenu en ontologie dite « légère ». Une ontologie est un modèle sémantique contenant un ensemble de concepts correspondant à l'ensemble des entités généralement (et consensuellement) reconnues comme représentant le domaine d'application traité, leurs définitions et les relations qui les lient. Les ontologies « légères » se distinguent des ontologies dites « lourdes » par le fait qu'elles n'intègrent pas de règles et de restrictions définies sur les concepts et relations pour en préciser leur intention. Ce modèle offre les outils nécessaires pour intégrer les spécificités territoriales que nous pouvons identifier et extraire des documents et notices attachées. La notion de propriétés nous permet de prévoir une description détaillée des informations géographiques. Aussi, l'ontologie donne la possibilité supplémentaire d'expliciter les relations entre concepts pour préciser le lien que nous pouvons identifier entre deux concepts. L'ontologie, par ses différents composants, est un outil qui nous permet d'offrir une représentation complète d'un territoire, et nous allons maintenant présenter de façon succincte la méthodologie définie pour construire une ontologie d'un territoire à partir d'un fonds documentaire indexé.

1.5.2 Méthodologie globale

En accord avec [Bac00], nous pensons que la construction d'ontologies dépend du domaine cible et des applications pour lesquelles elles sont définies. Dans notre cas, l'ontologie de territoire doit être construite pour représenter un territoire donné décrit dans un ensemble de documents. Notre contribution ne correspond pas uniquement à instancier une ontologie pour un contexte documentaire donné afin de faire émerger le territoire mais s'attache également à proposer une méthodologie complète automatisée permettant de construire une ontologie d'un territoire pour un fonds documentaire. La méthodologie que nous proposons (figure 1.2) est générique et automatisée. Elle peut être appliquée sur des fonds documentaires provenant d'autres centres documentaires. Cette méthodologie, que nous nommons « TERRIDOC », se décompose en quatre étapes :

1. Construction d'une ontologie légère à partir de la connaissance experte ;

2. Peuplement de l'ontologie à partir d'une chaîne de Traitement Automatique de la Langue Naturelle [8] (TALN) pour l'identification d'entités géographiques. La chaine est exécutée sur les notices descriptives résultant du travail d'indexation d'experts ;
3. Peuplement de l'ontologie à partir de la chaîne linguistique appliquée sur le contenu des documents pour proposer une représentation plus précise (constituée d'une nombre plus important d'informations géographiques) du territoire ;
4. Proposition d'enrichissements de l'ontologie d'un territoire à partir de documents textuels.

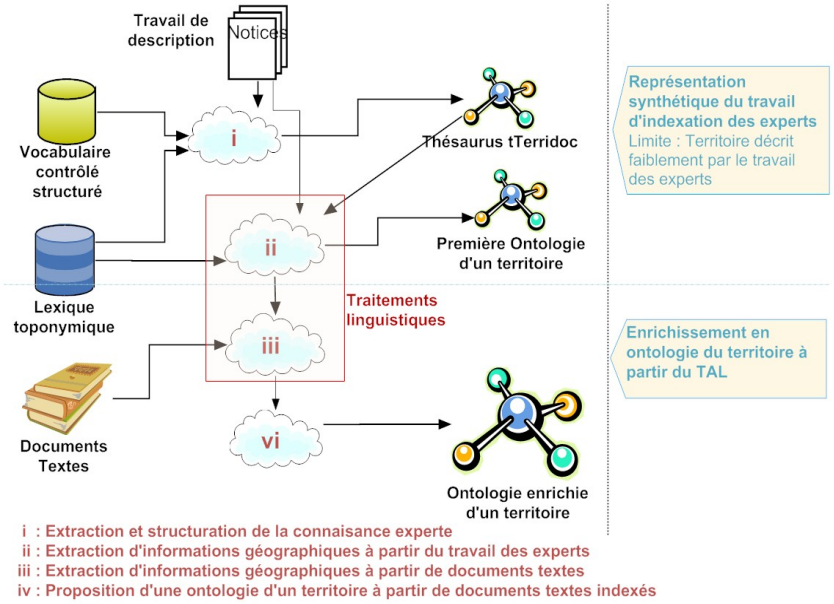


FIGURE 1.2 – Méthodologie générale TERRIDOC pour la construction d'une ontologie d'un territoire

Dans notre démarche, nous nous appuyons dans une première étape sur le travail d'indexation des experts. Ce travail est matérialisé sous forme de notices descriptives, pouvant être au format TXT, XML, etc., ainsi que sur le vocabulaire contrôlé utilisé par les experts pour construire ces notices descriptives. L'exploitation de ces ressources nous permet de produire une première représentation sémantique synthétique du fonds documentaire. Bien que le travail d'indexation des experts n'ait pas pour but de mettre en avant le territoire décrit par les documents, les spécificités liées aux fonds documentaires dits territorialisés nous permettent de penser qu'indirectement, ce travail de description forme une base de connaissances intéressante sur laquelle nous pouvons nous appuyer pour construire une première représentation d'un territoire. Les expérimentations réalisées montrent que ceci n'est que partiellement vrai et cela malgré les spécificités du

8. discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain (http ://fr.wikipedia.org/wiki/TALN)

type de fonds traité. La validation du travail d'identification des entités géographiques est réalisée en s'appuyant sur des ressources de type gazetteers [9].

Dans un deuxième temps, nous proposons une chaîne de traitement linguistique dont le but est d'identifier des entités géographiques et nous l'appliquons sur le contenu des notices descriptives afin d'instancier notre ontologie. Une entité géographique peut contribuer à instancier l'ontologie lorsqu'un lien est identifié entre l'une des composantes de cette entité (thème, espace ou temps) et les éléments constituants de l'ontologie (une composante de l'entité géographique est présente dans l'ontologie, en tant que label de concept par exemple). Nous sommes en mesure à ce niveau de traiter l'ensemble du fonds documentaire, en nous appuyant sur les notices descriptives attachées aux documents, et d'en fournir une ontologie peuplée d'un territoire décrit par l'ensemble des documents.

Dans une troisième étape, nous appliquons la chaîne de traitement linguistique sur le contenu des documents constituant le fonds documentaire traité. Il est important de préciser ici que, n'ayant pas les compétences nécessaires pour traiter le contenu des documents tels que les images, sons (beaucoup en langues locales) et vidéos, nous utilisons pour valider nos travaux le sous-ensemble de documents textuels avec leurs notices attachées. L'analyse linguistique réalisée sur le document permet d'extraire un nombre important d'entités géographiques que nous prenons en compte pour enrichir sémantiquement la première représentation du territoire obtenue.

Enfin, parmi cet ensemble d'entités géographiques identifiées lors de l'analyse linguistique, un nombre non négligeable d'entre elles ne peuvent rattachées à l'ontologie car aucun lien n'est repéré avec l'un des éléments la constituant. Dans une quatrième étape, nous proposons une analyse automatisée de ces entités géographiques dans le but d'identifier un lien sémantique avec l'un des éléments de l'ontologie. Le cas échéant, un enrichissement est alors proposé sous forme de label, de propriété ou encore de relation. Dans cette étape, il est nécessaire que des experts bibliothécaires interviennent pour valider les propositions.
Seule, l'ontologie résultante est difficilement exploitable par le grand public ou même les experts bibliothécaires pour observer et analyser le fonds documentaire. Généralement, les utilisateurs visiteurs de centres documentaires n'ont pas une connaissance précise du fonds documentaire et du territoire qui y est décrit. Pour mettre en avant ces connaissances, la visualisation de l'information apparaît comme l'une des voies les plus intéressantes pour « produire du sens » dans l'observation des masses de données. [Nor93] souligne le fait que l'utilisation d'aides externes et notamment visuelles augmente considérablement la puissance cognitive de l'esprit humain. Dans le même sens, [Bar00] montre que l'œil permet de percevoir un ensemble de signaux simultanément et d'effectuer un grand nombre de traitements instantanément avant même de mettre en place des mécanismes cognitifs comme le raisonnement ou la mémorisation.

Nous nous appuyons sur cet état de fait pour définir un système de recherche d'informations que nous nommons « TERRIDOCViewer ». L'application intègre notamment un module de recherche présenté sous forme d'une carte de concepts. Ce type de représen-

9. Répertoire toponymique fournissant les coordonnées géographiques correspondant au nom d'un lieu

tation, obtenu sur la base de l'ontologie créée à partir de la méthodologie TERRIDOC, doit permettre d'appréhender de façon synthétique l'ensemble des documents du corpus et de découvrir le territoire décrit dans les documents. Un usage sous-jacent de nos travaux est la possibilité pour les experts du domaine de naviguer à travers les documents via une représentation sémantique synthétisant leur travail d'indexation.

1.6 Organisation du mémoire

Notre mémoire de thèse se compose de quatre parties principales, que nous schématisons figure 1.3.

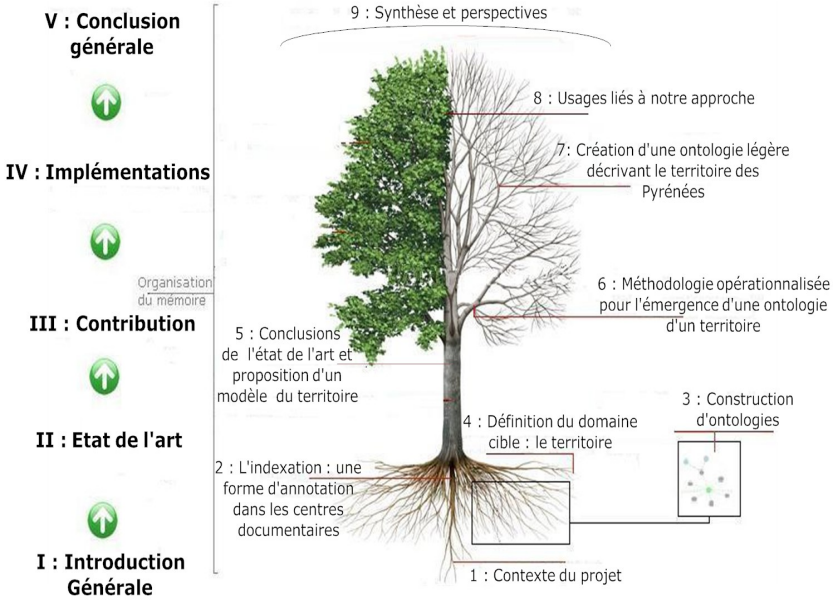


FIGURE 1.3 – Organisation du mémoire

La première partie que nous concluons maintenant comporte le **chapitre d'introduction** (*cf.* page 9) présentant le contexte de ces travaux, la problématique générale et les différents aspects de nos travaux qui seront traités dans ce mémoire.

La deuxième partie est constituée de trois chapitres présentant l'état de l'art en lien avec le cadre de notre thèse et d'un dernier chapitre présentant notre positionnement par rapport à l'ensemble des travaux scientifiques présentés. Le **deuxième chapitre** (*cf.* page 25) concerne les notions et concepts autour du travail d'annotation documentaire réalisé par des experts bibliothécaires dans des centres de type bibliothèques et médiathèques. Ce chapitre présente également l'ensemble des outils de type vocabulaire contrôlé utilisés pour aider à indexer les documents et nous nous focalisons notamment sur le thésaurus et l'ontologie. Le **troisième chapitre** dresse un état des travaux proposant d'extraire et de structurer la connaissance en une ontologie de domaine en privilégiant les approches s'appuyant sur des ressources textuelles et des ressources structurées

(*cf.* page 41). Le **quatrième chapitre** (*cf.* page 69) spécifie le domaine cible qu'est le territoire et présente un ensemble de travaux visant à définir cette notion complexe. Nous présentons ensuite dans ce quatième chapitre l'ensemble des travaux qui, à notre connaissance, permettent de constuire une ontologie géographique. Nous concluons cette partie dans un **cinquième chapitre** en proposant une synthèse de l'état de l'art dans lequel nous positionnons notre approche. Nous proposons ensuite un premier modèle du territoire que nous chercherons ensuite à implémenter dans notre approche (*cf.* page 85).

La troisième partie regroupe le détail de notre contribution visant à construire une première ontologie de domaine à partir de ressources structurées puis à l'enrichir par des informations géographiques résultant d'un traitement linguistique appliqué sur le corpus de textes. Le **chapitre six** présente la méthodologie opérationnalisée pour construire une ontologie d'un territoire à partir du travail d'indexation des experts formalisé sous forme de notices descriptives (*cf.* chapitre page 101). Ce chapître intègre une approche de peuplement d'ontologie pour faire émerger un domaine cible, incluant tout d'abord une analyse linguistique d'un corpus textuel afin d'identifier et d'extraire l'information pertinente (l'information géographique dans notre cas) présente dans les documents. Nous proposons également des premiers traitements d'enrichissement de l'ontologie résultante par des concepts identifiés lors de l'analyse linguistique.

La quatrième partie, que nous décomposons en deux chapitres, décrit l'implémentation de l'ensemble des modules et outils permettant de valider notre approche. Le **chapitre sept** présente l'implémentation complète de notre approche permettant de construire une ontologie du territoire des Pyrénées (*cf.* page 135). Le **chapitre huit** présente des premiers usages scientifiques et applicatifs découlant de la méthodologie proposée (*cf.* page 163). Un premier usage scientifique, réalisé dans le cadre du projet ANR GEONTO qui vise le domaine spécifique des ontologies géographiques, décrit une application de notre approche d'enrichissement d'une ontologie géographique afin d'améliorer la phase d'identification de la représentation spatiale adaptée à l'entité spatiale identifiée. Au niveau applicatif, nous proposons notamment une application Web, développée dans le cadre de la thèse, qui s'appuie sur le travail d'extraction et de structuration de la connaissance pour naviguer à travers les documents du corpus. Un deuxième scénario applicatif, destiné aux experts bibliothécaires, coïncide à les aider dans leur travail de validation en leur proposant de façon automatique un ensemble d'erreurs identifiées et des solutions correspondantes.

La cinquième partie conclut ce mémoire dans un **neuvième et dernier chapitre**. Elle synthétise l'ensemble de notre proposition et met en avant des perspectives scientifiques et applicatives (*cf.* page 199).

Deuxième partie

Etat de l'art

Chapitre 2

L'indexation : une forme d'annotation dans les centres documentaires

Sommaire

2.1 L'annotation de documents 26

2.1.1 Définitions 26

2.1.2 Le travail d'annotation des experts et ses spécificités . . . 28

2.2 Les différents types de vocabulaires contrôlés 29

2.2.1 Synthèse des vocabulaires contrôlés les plus utilisés 29

2.2.2 L'ontologie : une structure formelle pour représenter des connaissances 32

2.3 Discussions 38

Les documents constituant les fonds documentaires de bibliothèques et de médiathèques possèdent des informations de description sélectionnées par des experts bibliothécaires et formalisées sous forme de notices descriptives. Afin de bien comprendre les problématiques liées à ce travail d'expertise, nous définissons dans un premier temps le travail d'annotation de documents en insistant sur les spécificités du travail d'experts en centre documentaire. Nous présentons ensuite les divers outils, de type vocabulaire contrôlé, mis à disposition des experts bibliothécaires pour réaliser ce travail d'annotation de documents. Nous nous attardons sur la notion d'ontologie qui, encore peu utilisée dans les centres documentaires, permet de modéliser la connaissance d'un domaine de façon précise tout en respectant les règles propres aux vocabulaires contrôlés. Enfin, nous concluons ce chapitre en reprenant, parmi les différentes ressources que nous avons à disposition, celles qui semblent le plus adaptées pour construire le plus automatiquement possible une structure sémantique complète modélisant un domaine cible.

2.1 L'annotation de documents

Dans ce chapitre, nous définissons, dans un premier temps, les notions importantes qui sont liées au travail d'annotation de documents. Ensuite, nous décrivons le travail réalisé par les experts dans les centres documentaires de type bibliothèques ou médiathèques ainsi que les différents outils mis à disposition des experts pour les aider. Enfin, nous concluons ce chapitre par une analyse du travail d'indexation et des outils disponibles pour le réaliser qui nous permet de faire un choix sur la structure sémantique que nous souhaitons utiliser pour décrire un territoire.

2.1.1 Définitions

Plusieurs notions importantes s'articulent autour du travail d'annotation de documents. Tout d'abord, le **document** constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document.

Le **fonds documentaire** est défini dans le dictionnaire de l'Académie Française [10] comme « l'ensemble des documents, des livres, des manuscrits ou des œuvres d'art provenant d'un même donateur, ou se rapportant à un même thème ». Le fonds documentaire (ou collection de documents, corpus) constitue l'ensemble des informations exploitables et accessibles. Dans le cas général et dans un souci de faciliter l'accès aux documents, un travail d'annotation est réalisé pour chaque document du corpus sous forme de représentations simplifiées mais suffisantes. Ces représentations sont étudiées de telle sorte que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base puissent être réalisées dans les meilleures conditions.

Une **annotation** est définie par le Petit Robert comme une « note critique ou explicative qui accompagne un texte ; une note de lecture qu'on inscrit sur un livre ». Le grand dictionnaire terminologique [11] précise cette définition dans le domaine de l'informatique comme un « type d'analyse documentaire où, dans le dessein d'une recherche ultérieure de l'information, on exprime le contenu des documents au moyen de descripteurs (parfois appelés mots clés) ».

De façon générale, le travail d'annotation peut être réalisé à partir d'un vocabulaire ouvert ou d'un vocabulaire fermé. Un **vocabulaire ouvert** peut être mis à jour par des annotateurs à tout instant sans règle ni contrôle. Au sein de la famille des vocabulaires ouverts, nous pouvons notamment citer la folksonomie [12] [VW07]. L'expression francisée officielle est « indexation personnelle » [13]. Parmi les évolutions liées à la mutation

10. Dictionnaire de la langue française, dont la rédaction - sous la direction de son secrétaire général perpétuel - et la diffusion constituent l'une des missions de l'Académie française. http ://www.academie-francaise.fr/dictionnaire/index.html

11. http ://www.granddictionnaire.com/btml/fra/r_motclef/index800_1.asp

12. Thomas Vander Wal a défini ce terme en combinant la taxonomie (taxonomy en anglais) et les usagers (folk).

13. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000021530619&dateTexte=&categorieLien=id

du Web (réseaux sociaux, encyclopédies ouvertes, etc.), la folksonomie constitue une des fonctionnalités phare du Web 2.0 en permettant aux utilisateurs de décrire des ressources (billets de blog, pages Web, photos, vidéos, etc.) par des mots clés choisis librement. On parle alors de classification collaborative décentralisée spontanée. Parmi les exemples les plus connus de folksonomies, le site de « deliciousSocial bookmarking » [14] permet de partager des signets et le site Flickr [15] offre la possibilité de rendre accessible des photos, annotées par des millions d'utilisateurs à travers le monde. Bien qu'intéressant, les vocabulaires ouverts sont difficiles à organiser et à maintenir. Pour ces raisons, ils sont peu utilisés dans le monde des bibliothèques.

A l'inverse du vocabulaire ouvert, un vocabulaire fermé ou **vocabulaire contrôlé** se compose d'un ensemble limité de termes. Il est défini par un groupe (une communauté de pratiques) afin de pouvoir annoter des contenus. La signification des termes n'est pas forcément explicitée et il n'y a pas nécessairement d'organisation logique des termes entre eux [LM01]. Cependant, dans les centres documentaires les vocabulaires sont définis de façon précise. Généralement, ils sont constitués de termes, de relations existant entre ces termes, de contraintes, de définitions, etc. La description de ces différents éléments est exprimée à la fois par la structure du document et par le langage naturel. Le vocabulaire contrôlé est une structure sémantique définie à plusieurs niveaux selon des règles précises présentées dans les normes NF Z 47-100 [16] et ISO 2788 [17] ainsi que dans le document élaboré par la section de classification et d'indexation de IFLA [18].

Au niveau sémantique tout d'abord, un élément du vocabulaire contrôlé n'aura qu'un seul sens précis dans le vocabulaire d'indexation donné et cet élément sera le seul à avoir ce sens. Ceci implique un contrôle de la synonymie (présence des descripteurs et des non-descripteurs autrement appelés « descripteurs rejetés » ou « employés pour ») et de la polysémie (utilisation d'une syntaxe particulière pour différencier les différents sens). L'élément choisi dans le vocabulaire contrôlé pour représenter un sens (ou un concept) est aussi appelé **vedette** et l'ensemble des termes du vocabulaire contrôlé qui portent le même sens sont des termes rejetés. De même, le contrôle de l'homonymie et de la polysémie est effectué par divers moyens (qualificatifs, adjectifs complémentaires, notes d'application, etc.). Parmi les vocabulaires contrôlés, différentes structures sémantiques intègrent des relations hiérarchiques. Un contrôle sémantique supplémentaire est alors possible en analysant la mise en relation des éléments qui composent ces vocabulaires.

Au niveau terminologique, des règles permettent notamment d'indiquer le choix de la forme en français ou en langue étrangère, l'utilisation du singulier ou du pluriel, une forme en langage courant ou, au contraire, en langage technique plus ou moins spécialisé.

Au niveau syntaxique, des règles permettent au vocabulaire de traiter les sujets complexes ou nouveaux, pour lesquels il n'existe pas de descripteurs directs. Ce dernier

14. http ://delicious.com/
15. http ://www.flickr.com/
16. Règles d'établissement des thésaurus monolingues.
17. Principes directeurs pour l'établissement et le développement de thésaurus monolingues.
18. Working Group On Principles Underlying Subject Heading Languages. Principles Underlying Subject Heading Languages (SHLs).

niveau de contrôle est applicatif et peut s'exercer pendant le processus d'indexation.

Nous nous intéressons dans nos recherches au travail réalisé par les experts bibliothécaires dans les centres documentaires sur la base de vocabulaires contrôlés.

2.1.2 Le travail d'annotation des experts et ses spécificités

Les experts bibliothécaires réalisent un travail d'annotation dite descriptive visant à valoriser des fonds documentaires (figure 2.1).

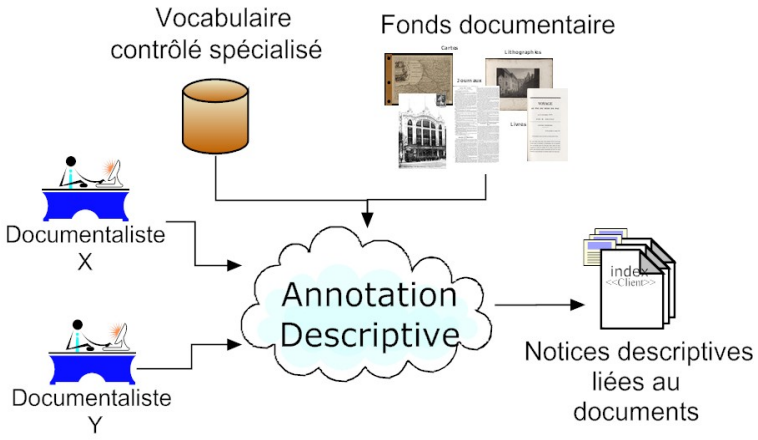


FIGURE 2.1 – Travail d'indexation du bibliothécaire

Cette action d'annotation consiste d'une part à décrire le document en indiquant tous les renseignements bibliographiques relatifs à la source (auteur, titre légende, etc.) et d'autre part à « indexer » le document en proposant une brève description sur le contenu du document en s'appuyant sur un vocabulaire contrôlé d'indexation. Un exemple de notice descriptive est présenté 7.4. Dans ce contexte, « l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse » [Mén93]. L'intérêt est ici de pouvoir ajouter du sens au travail de description pour améliorer ensuite les résultats lors de la phase de recherche documentaire. On parle alors d'informations sémantiques sur les données ou plus généralement d'**annotation sémantique** (*i.e.*, l'annotation à l'aide de termes définis dans des vocabulaires contrôlés).

Dans nos travaux, nous nous intéressons plus précisément à l'étape d'indexation du contenu des documents. Cette étape s'appuie sur un ou plusieurs vocabulaire(s) contrôlé(s) complet(s) défini(s) au préalable. Ces vocabulaires peuvent être définis en local dans un centre documentaire ou à plus grande échelle pour un groupe de centres documentaires. Il est important de préciser ici que le terme sélectionné à partir du vocabulaire externe pour décrire le contenu d'un document ne se trouve pas forcément

dans ce document. L'intérêt est d'offrir le plus de concordance possible entre la description du contenu d'un document et le terme choisi par un utilisateur pour obtenir des informations sur un concept décrit dans ce document.

2.2 Les différents types de vocabulaires contrôlés

La famille des vocabulaires contrôlés est grande comme le montre notamment [Dub04]. Nous présentons de manière synthétique les structures sémantiques les plus utilisées. Puis nous nous attardons ensuite sur l'ontologie. Encore peu utilisé dans les bibliothèques et médiathèques, nous verrons que l'ontologie correspond à une vocabulaire contrôlé et organisé qui intègre des outils permettant d'expliciter de façon précise la sémantique.

2.2.1 Synthèse des vocabulaires contrôlés les plus utilisés

2.2.1.1 Le glossaire

Simple liste définie de termes (ou mots-clés) qui décrit des sujets restreints fréquemment à un domaine particulier ou une application précise. Le glossaire est une spécialisation d'un lexique qui peut se définir comme un ensemble de mots d'une langue, et qui a donc une portée plus générale. Le glossaire est composé d'un ensemble de termes souvent classés par ordre alphabétique et il est introduit par un index. Aussi dans le glossaire, une définition est associée à chaque terme et celle-ci peut varier de celle présentée dans un dictionnaire car on donne la définition du mot dans le contexte dans lequel il est utilisé, ou le domaine auquel il se rapporte. L'exemple du terme *baie* provenant d'un glossaire de la géographie [19] qui a comme seule définition :« ouverture, arc de cercle du littoral. » alors que dans un autre contexte comme par exemple dans le dictionnaire français Larousse [20], ce terme peut avoir pour définition « Fruit charnu, indéhiscent, qui contient directement les graines, tel que la groseille, le raisin ou la myrtille. ».

2.2.1.2 La taxonomie

Aussi appelée **taxinomie** (du grec taxis : rangement et nomos : loi), la taxonomie est l'étude théorique de la classification, de ses bases, de ses principes, des méthodes et des règles. A l'origine le terme « taxonomie » [Can13] ne s'intéresse qu'à la classification biologique. Aujourd'hui la taxonomie élargit son champ d'application à la science en général avec pour premier but de repérer et classifier les objets du monde pour les étudier et les comprendre [Cha02]. [BAGC04] met en avant cette évolution pour les sciences naturelles pour répondre à un besoin de classifier des espèces, etc.

Cette structure sémantique définit un ensemble de termes qui sont arrangés dans une hiérarchie de généralisation-spécialisation [Tex05]. En mathématiques, la relation de généralisation-spécialisation est appelée relation d'inclusion (*cf.* figure 2.2) signifiant qu'un ensemble A est un sous-ensemble ou une partie d'un ensemble B, ou encore que

19. http ://www.dijon.iufm.fr/spip.php ?article255
20. http ://www.larousse.fr/dictionnaires/francais/baie

B est sur-ensemble de A, si tout élément du sous-ensemble A est aussi élément du sur-ensemble B. Il peut par contre y avoir des éléments de B qui ne sont pas éléments de A.

$$A \subset B \quad \text{signifie} \quad \forall\, x\, (x \in A \Rightarrow x \in B)$$

FIGURE 2.2 – Représentation formelle de la relation d'inclusion

Cette relation d'inclusion peut se lire de plusieurs façons : « A est contenu dans B », « A est une partie de B » ou « A est un sous-ensemble de B ».

Dans une taxonomie, le vocabulaire contrôlé est organisé sous forme hiérarchique simple. Cette hiérarchisation correspond souvent à une spécialisation. Il existe donc un lien précis entre un terme du vocabulaire et ses enfants. Ce lien donne un sens supplémentaire, une signification. La taxonomie peut être définie comme un vocabulaire organisé de termes respectant les règles d'un vocabulaire contrôlé. Ce vocabulaire organisé est limité dans le sens où il ne peut pas définir les attributs de ces termes, ni leurs relations. Voici figure 2.3 un exemple de taxinomie en Mathématiques.

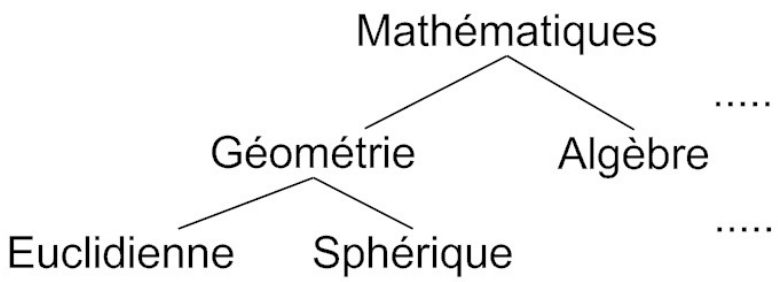


FIGURE 2.3 – Exemple de taxonomie dans le domaine des Mathématiques

2.2.1.3 Le thésaurus

Selon la définition de l'AFNOR [21], un thésaurus est un langage documentaire fondé sur une structuration hiérarchique d'un ou plusieurs domaines de connaissances et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre les notions par des signes conventionnels. Le thésaurus multilingue associé au projet HEREIN [22] en est un bon exemple. Ce thésaurus est dédié à la recherche au sein d'un corpus particulier de documents multilingues sur les politiques nationales des pays de l'Union Européenne. Il permet d'interroger les textes constitutifs du corpus dans

21. AFNOR. Documentation : règles d'établissement des thésaurus monolingues. NF Z47-100, 1981.

22. http ://thesaurus.european-heritage.net/sdx/herein/thesaurus/consult.xsp. L'objectif du projet est de proposer un système terminologique relatif aux politiques nationales concernant le patrimoine architectural et archéologique au sens défini par les Conventions de Grenade (octobre 1985) et de La Vallette (janvier 1992).

leur intégralité et dans chacune des langues. Nous pouvons également citer le thésaurus bilingue anglais-français MeSH [23] couvrant le domaine biomédical ou encore le thésaurus multilingue AGROVOC défini par la FAO [24] pour couvrir les domaines liés à l'agriculture, à la pêche, à l'alimentation et aux domaines connexes. Dans le tableau 2.1 sont décrites les relations les plus usuelles d'un thésaurus.

t1 **terme préféré dans** t2, t3,...,tN	t1 est le terme préféré pour désigner l'ensemble des synonymes t2, t3,...,tN
t1 **note** texte	Remarque sur le terme t1 (usage exceptionnel, contexte d'utilisation)
t1 **utiliser plutôt** t2	t2 est utilisé pour désigner t1
t1 **utilisé pour** t2	t1 est utilisé pour désigner t2
t1 **plus spécifique que** t2	le terme désigné par t1 est plus spécifique que le terme désigné par t2
t1 **plus générique que** t2	le terme désigné par t1 est plus générique que le terme désigné par t2
t1 **est lié à** t2	t1 est un terme lié ou associé à t2

TABLE 2.1 – Relations entre termes dans un thésaurus (d'après [Her05])

Un thésaurus peut donc être vu comme une taxonomie enrichie. La taxonomie permettait d'obtenir une spécialisation des termes employés. Le thésaurus donnera de l'information sur les sujets connexes également. On pourra donc restreindre ou élargir le champ de connaissance. Cet élargissement se fait en donnant les termes relatifs. Des liens permettent la spécialisation. On pourra alors dire : c'est une **sous-catégorie** (spécialisation) ou est « **relatif à** » ou « **terme associé** » ou encore « **voir également** » (élargissement). Aussi, les thésaurus dans leur définition font régulièrement la différence entre un même terme au singulier et au pluriel. Un terme au singulier représente le sens du terme alors que son pluriel va représenter une catégorie. Si nous prenons l'exemple du terme *Château(x)*, le singulier *Château* définit le concept *Château* alors que son pluriel (*Châteaux*) regroupe un ensemble de châteaux.

La figure 2.4 présente un exemple de thésaurus, enrichissant l'exemple de taxinomie présenté figure 2.3, contenant des relations de types hiérarchiques ainsi que des relations de types « terme associé ».

Il est important de noter que les relations hiérarchiques incluent la relation générique (genre-espèce), la relation partitive (tout-partie), la relation d'instance et les relations poly-hiérarchiques. [Fis98] souligne cette ambiguïté par le fait que la définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation,

23. Medical Subject Headings : http ://mesh.inserm.fr/mesh/presentation.htm. Le thésaurus est régulièrement mis à jour par la NLM(National Library of Medicine).

24. Organisation des Nations unies pour l'alimentation et l'agriculture, http ://www.fao.org/index_fr.htm

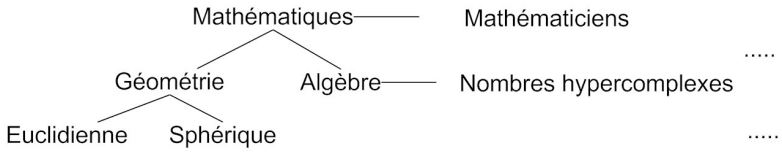


FIGURE 2.4 – Exemple de thésaurus dans le domaine des Mathématiques

recherche), et non par la formalisation de la connaissance du domaine.

Les glossaires, taxonomies et thésaurus sont utilisés dans le monde des bibliothèques pour annoter les documents depuis de nombreuses années. Dans les années 1990, un nouveau type de vocabulaire contrôlé nommé **ontologie** voit le jour. Jusqu'à aujourd'hui, le principe d'un vocabulaire contrôlé organisé sous forme d'une ontologie est encore utilisé de manière expérimentale au sein des centres documentaires de type bibliothèques et médiathèques. Nous allons voir maintenant que cette structure de vocabulaire contrôlé possède les outils permettant de modéliser la connaissance implicitement décrite dans les fonds documentaires.

2.2.2 L'ontologie : une structure formelle pour représenter des connaissances

Cette section présente la notion d'ontologie et montre que ce type de vocabulaire contrôlé possède des outils permettant de couvrir la sémantique d'un domaine cible de façon précise. Dans un premier temps, nous définissons la notion d'ontologie en tentant de mettre en avant ses spécificités. Nous décrivons ensuite les différents éléments consituant une ontologie et nous terminons cette section en présentant les différentes catégories d'ontologies existantes classées selon la richesse de la structure [LM01] puis en fonction du niveau d'abstraction (travaux initiés par [VHSW97]).

2.2.2.1 Définitions

A l'origine, l'ontologie est une notion philosophique, sous domaine de la métaphysique, qui étudie la nature et l'organisation de ce qui existe. Le terme **ontologie**, du grec ancien ontos (être) et logos (discours), est défini comme : « étude de l'être en tant qu'être ». Le terme n'apparaît qu'au 19$^{\text{ème}}$ siècle mais fait référence à la notion qu'Aristote désignait il y a près de 2300 ans comme théorie de l'être en tant qu'être [25]. L'étymologie renvoie ainsi à l'étude des propriétés générales de ce qui existe (qualité, substance), indiquant une volonté d'expliquer les concepts qui existent dans le monde et comment ces concepts s'imbriquent et s'organisent pour donner du sens. Aujourd'hui, ce terme signifie la « science des étants » c'est-à-dire l'ensemble des objets reconnus comme existants dans un domaine [26].

25. Définition provenant de l'encyclopédie Universalis.
26. Bruno Bachimont, http ://www.technolangue.net/article.php3 ?id_article=280

Au début des années 1990, le terme ontologie voit donc le jour en Intelligence Artificielle (IA). [NFF+91] présente la notion d'ontologie de la façon suivante : « *An ontology defines the basic terms and relations to define extensions to the vocabulary* ». En 1993, Thomas Gruber propose sa définition : « *An ontology is an explicit specification of a conceptualization* » [Gru93], qui devient et reste jusqu'à nos jours la définition référence en Ingénierie des Connaissances. En 1997, [Bor97] modifie légèrement la définition de Thomas Gruber : « *Une ontologie est définie comme étant une spécification formelle d'une conceptualisation partagée* ». [SBF98] explique les notions clés de ces deux définitions de la façon suivante :

- **explicite** : « type des concepts et les contraintes sur leurs utilisations sont explicitement définies » ;
- **formelle** : « renvoie au fait que la spécification doit être lisible par une machine » ;
- **partagée** : « capture la connaissance consensuelle, qui n'est pas propre à un individu mais validée par un groupe » ;
- **conceptualisation** : « un modèle abstrait d'un certain phénomène du monde reposant sur l'identification des concepts pertinents de ce phénomène ».

Parmi les nombreuses définitions de la notion d'ontologie qui ont suivi en ingénierie des connaissances, nous reprenons celle de Guarino [Gua98] indiquant qu'une ontologie est un vocabulaire partagé, plus une spécification ou caractérisation du sens « convenu » de ce vocabulaire.

Nous remarquons que la notion d'ontologie, vue comme une science en philosophie (l'Ontologie), est définie comme un objet (une ontologie) en informatique. F. Gandon [Gan02] enrichit cette définition en indiquant qu'une ontologie informatique est une représentation de propriétés générales de ce qui existe que l'on peut formaliser et qui peut supporter un traitement rationnel.

A partir du milieu des années 1990, les ontologies sont utilisées dans de nombreux domaines. Guarino [Gua98] liste les domaines suivants : l'ingénierie des connaissances, la modélisation qualitative, l'ingénierie des langages, la conception de bases de données, la recherche d'information, l'extraction d'information, la gestion et l'organisation de connaissances. Depuis, grâce à l'essor du Web, elles sont utilisées dans le domaine du commerce en ligne et sont au centre du Web Sémantique [BHL01]. L'intérêt est ici d'ajouter au Web un ensemble de connaissances permettant des recherches d'information au niveau sémantique et non plus au simple niveau lexical et/ou syntaxique.

Depuis, de nombreux travaux de recherche visant à proposer de nouvelles méthodes de construction d'ontologies, vont dans ce sens et enrichissent la définition de la notion d'ontologie [CJB99], [Bac00], [NSD+01], [CBT05], [AGM04], etc.). En nous imprégnant de l'ensemble de ces définitions, nous nous accordons avec Nathalie Aussenac [AGM04] pour dire qu'une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine.

En nous appuyant sur les définitions citées ci-dessus ainsi que sur les travaux de

[Dia06], nous pouvons caractériser une ontologie de la façon suivante :

- Les ontologies sont *formelles* : les concepts sont exprimés dans une langue intégrant des spécifications précises permettant de les traiter par des programmes informatiques. Les concepts ou les objets qui existent dans des techniques de modélisation traditionnelles (schéma relationnel et UML, par exemple) sont seulement semi-formels et ne permettent pas d'expliciter aisément leur sémantique.
- Les ontologies sont *compréhensibles par les humains*. Ceci signifie, qu'avec la connaissance d'un langage adapté (par exemple OWL [27]), elles peuvent être manipulées (construction, lecture, échange via un ordinateur) par les communautés d'experts de domaine ainsi que des utilisateurs potentiels.
- Les ontologies sont *vastes*. L'objectif est d'inclure toute la signification appropriée des concepts liés à un domaine ou une application. L'intérêt est que la sémantique intégrée dans l'ontologie puisse être comprise, modifiée, et contrôlée par n'importe quel expert de domaine.
- Les ontologies sont *partageables*. De plus en plus, les ontologies sont définies à partir de bibliothèques communes de concepts fondamentaux et sont utilisables à travers de multiples domaines d'application. Aussi, l'évolution du Web vers le Web sémantique a permis de faire émerger des outils pour représenter une ontologie et la partager. Ceci aide la communication entre les systèmes d'information qui doivent partager des informations basées sur des concepts communs.

Une ontologie est de façon générale constituée de **concepts**, regroupant un vocabulaire de termes décrivant un domaine ou un système plus précis et pouvant être caractérisés par des **propriétés** telles que des définitions. Ces concepts sont reliés entre eux par des **relations explicites** afin de former une structure sémantique modélisant la connaissance du domaine ou de l'application cible. Nous allons maintenant présenter ces différents éléments.

2.2.2.2 Ingrédients d'une ontologie

Nous détaillons les différents éléments que nous venons de citer qui constituent une ontologie, à savoir les concepts, les propriétés, les relations, les axiomes, et les instances [AGPLTP99], [Tro04].

- **Les concepts** : Comme l'indique Uschold et King [UK95], un concept peut représenter un objet matériel (par exemple une église, un monastère, etc.), une notion (par exemple, le poids) ou bien une idée. Dans le vocabulaire commun des ontologies, la notion est aussi appelée intention. [Bac04] étend cette définition sur trois niveaux :
  - Son sens : la position du concept dans la structure sémantique permet de l'identifier, de le comprendre et de le différencier par rapport aux autres concepts ;
  - Sa construction : Comprendre un concept revient à construire l'objet dont il est le concept.

27. Ontology Web Language

  - Sa prescription : Comprendre un concept revient à exécuter l'action qu'il entreprend.

  Le concept (ou classe) est défini en fonction du domaine ou de l'application cible pour l'ontologie. La notion de concept est différente de celle de terme : un terme est une représentation de concept (sous forme de mot ou groupes de mots), un concept pouvant être décrit par plusieurs termes. Un concept est donc représenté par un terme aussi appelé **label** mais il peut aussi être symbolisé par une liste de termes, permettant ainsi d'étendre la portée du concept. Par exemple, le concept *Eglise* peut avoir pour label les termes *Eglise* et *Etablissements Religieux*. Concernant les aspects terminologiques, il est d'usage de ne pas utiliser de pluriel dans la définition de concepts contrairement aux autres vocabulaires contrôlés.
- **Les propriétés** : ce sont des attributs valués attachés aux concepts ;
- **Les relations** : ce sont des liens permettant de structurer les concepts entre eux pour former une représentation enrichie du domaine ou de l'application cible. Ces relations peuvent être de type hiérarchique (relations genre-espèce (est_un), partitive (partie_de), d'instance (instance_de) ou poly-hiérarchiques) pour former une taxonomie, de type associative comme dans un thésaurus avec la possibilité supplémentaire de qualifier précisément la relation.
- **Les instances** : ce sont des éléments singuliers qui véhiculent les connaissances concernant l'application ou l'application cible. Ces objets, définis par un concept, sont appelés extensions de ce concept.
- **Les axiomes** : Ces éléments permettent de modéliser les assertions acceptées comme toujours vraies dans l'ontologie. Ils peuvent être utilisés en tant que règles générales ou lors de la définition des concepts ou des relations. Les axiomes indiquent les intentions des concepts et des relations du domaine, mettant en avant les connaissances ne possédant pas de spécificités exclusivement terminologiques [SM00].

L'ontologie est un outil sémantique permettant de modéliser la connaissance d'un domaine ou d'une application de façon très précise.

2.2.2.3 Les différents types d'ontologie

Différents types de classification des ontologies sont proposées dans la littérature parmi lesquels nous pouvons notamment citer la classification selon le niveau conceptuel [VHSW97] ou encore la classification selon la richesse de la structure sémantique [LM01]. Ces deux classifications sont importantes à prendre en compte lorsque l'on souhaite construire une ontologie.

Une première forme de classification peut être réalisée selon le niveau conceptuel. On distingue différents niveaux d'ontologies selon le domaine modélisé et éventuellement les applications pour lesquelles elles sont conçues [VHSW97], [Gua98], [Miz98], [CJB99], [GP99], [BB01] :

Supérieure (haut niveau) [VHSW97], [GP99] : L'ontologie supérieure définit les primitives qui permettent de décrire l'ontologie générique ou de domaine. Elle se veut

neutre, d'un niveau d'abstraction élevé, et ne fait pas référence aux entités du monde réel. [Gua98] met en avant le fait que les ontologies de représentation sont indépendantes des différents domaines de connaissances, puisqu'elles décrivent des primitives cognitives communes aux divers domaines. Les ontologies supérieures décrivent les idées utilisées dans toutes les ontologies pour spécifier les connaissances, telles que les substances, les concepts, les relations etc. Par exemple, la « Frame-Ontology » utilisé dans Ontolingua [Gru93] est une ontologie de haut niveau proposant une définition formelle des concepts utilisés principalement dans les langages à base de frames : classes, sous-classes, attributs, valeurs, relations et axiomes.

Générique [Gua97], [Sow95], [GGPPSF07] du projet NeOn [28] : D'un niveau d'abstraction moins élevé que l'ontologie supérieure, l'ontologie générique se rapproche de l'ontologie de domaine à l'exception près qu'elle définit des concepts pouvant être considérés comme communs à différents domaines. Aussi appelé méta-ontologie, l'ontologie générique doit pouvoir être rattachée aux sommets d'ontologies de domaine, décrivant les concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc. Elle peut modéliser des connaissances factuelles ou encore des connaissances visant à résoudre des problèmes dans différents domaines. [Cha02] met en avant le fait que le haut niveau d'abstraction des ontologies génériques impliquant une normalisation légère et forcément limitée rend difficile leur réutilisation. Wordnet [29] est un exemple d'ontologie générique composée d'ensembles de synonymes (synsets), des termes regroupés en classes sémantiques d'équivalence, chaque terme appartenant à une catégorie lexicale donnée (nom, verbe, adverbe, adjectif). C'est une ontologie linguistique qui couvre la plupart des mots anglais ordinaires, couvrant ainsi une multitude de domaines.

Domaine : décrit le vocabulaire lié à un domaine générique (comme la médecine, ou les automobiles) en spécialisant les concepts présentés dans les ontologies de haut niveau. Elle donne une représentation formelle des concepts du domaine étudié ainsi que des différentes relations qui lient ces derniers, offrant ainsi une représentation de la connaissance plus précise que les ontologies de plus haut niveau. Définies pour être réutilisables dans différentes applications d'un même domaine, les ontologies de domaine sont indépendantes du type de traitement qui va être opéré sur les connaissances les constituant. L'ontologie pose ici des règles régissant la structure et le contenu de la connaissance du domaine cible, formant une sorte de méta-modèle de connaissance dont les concepts et propriétés sont de type déclaratif. La majorité des ontologies existantes sont de niveau générique. Parmi les nombreux travaux existants, [NV06] propose la création d'une ontologie de domaine mettant en avant l'héritage culturel.

Tâches [MKSK00] : conceptualise des tâches spécifiques dans les systèmes, telles que les tâches de configuration, de planification, de conception, soit tout ce qui concerne

28. www.neon-project.org
29. http ://wordnet.princeton.edu/

la résolution de problèmes. Selon Mizoguchi [MKSK00], l'ontologie de tâches caractérise l'architecture computationnelle d'un système à base de connaissances qui réalise une tâche. L'ontologie des objectifs d'apprentissage (Learning Goal Ontology) est un exemple d'utilisation de l'ontologie de tâches dans le domaine de l'éducation [ISI+00], décrivant les rôles des apprenants et des agents dans le cadre d'un apprentissage collaboratif.

Application : contient l'ensemble des définitions nécessaires à la construction de la connaissance dans une application particulière. L'ontologie d'application, plus spécifique, est difficilement réutilisable car elle est définie pour une application particulière. Cette ontologie est la plus spécifique. Les concepts dans l'ontologie d'application correspondent souvent aux rôles joués par les entités du domaine tout en exécutant une certaine activité [Mäd02]. Notons tout de même qu'une ontologie d'application peut être définie sur la base d'une ontologie de domaine.

Cette première forme de classification s'appuyant sur le niveau d'abstraction des ontologies amène jusqu'à aujourd'hui au débat sur la nature et la réutilisabilité des ontologies qui est bien présent en ingénierie ontologique [CBBZ96], [Bac00], [Gua97], [VHSW97]. Un premier courant, porté par N. Guarino [Gua97] notamment, se place dans la construction d'ontologies génériques, voire supérieures en cherchant à définir une ontologie universelle réutilisable dans tout type d'applications quel que soit le domaine d'application. C'est notamment le cas de l'ontologie Cyc [LDG90] définie pour représenter un ensemble maximal de connaissances, reliées sémantiquement.

Un deuxième courant porté notamment par B. Bachimont [Bac00], [BAG03] préconise la construction d'ontologies de domaine régional dépendantes du domaine d'application et des applications pour lesquelles elles sont définies. B. Bachimont précise que « *définir une ontologie pour la représentation des connaissances, c'est définir, pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée* » [Bac00]. Parmi les travaux de recherche formant le premier courant, certains remettent finalement en cause le fait qu'une ontologie puisse être réutilisée telle quelle dans différents domaines distincts et pour une multitude d'applications diverses [VHSW97].

Une deuxième forme de classification, tout aussi importante, peut être réalisée de façon transversale à la classification selon le niveau conceptuel. Lassila et McGuinness [LM01] repris ensuite par [GPCFL04] proposent une classification des ontologies selon l'information dont l'ontologie a besoin et la richesse de sa structure interne. Cette classification consiste en un ensemble allant de structures légères aux ontologies lourdes. Les ontologies « légères » incluent des concepts, comprenant des propriétés, et organisées en taxonomies avec des relations conceptuelles ; certains auteurs considèrent les taxonomies comme des ontologies parce qu'elles fournissent des conceptualisations partagées pour des domaines donnés. Les ontologies légères se distinguent des ontologies dites « lourdes » par le fait qu'elles n'intègrent pas d'axiomes (règles) et de restrictions définies sur les concepts et relations pour en préciser leur intention. Ces restrictions, généralement formalisées dans un langage logique, peuvent être définies dans les ontolo-

gies lourdes sur les valeurs des propriétés ou entre des constituants (e.g. relations). Les ontologies lourdes modélisent un domaine de façon plus profonde avec plus de restrictions basées sur la sémantique du domaine, rendant par la même occasion ce travail de modélisation plus complexe et difficilement réutilisable pour divers besoins.

2.3 Discussions

Rappelons tout d'abord l'objectif de ce travail qui consiste à modéliser et structurer la connaissance mettant en avant un domaine cible implicitement décrit dans un fonds documentaire annoté. Les ressources à notre disposition sont un corpus documentaire, des notices descriptives produites par des experts bibliothécaires ainsi que le vocabulaire contrôlé utilisé pour réaliser le travail de description du contenu. Dans les centres documentaires tels que les bibliothèques et les médiathèques, l'action d'annotation consiste d'une part à décrire le document en indiquant tous les renseignements bibliographiques relatifs à la source (auteur, titre, légende, etc.) et d'autre part à « indexer » le document en proposant une brève description sur le contenu du document en s'appuyant sur un vocabulaire contrôlé d'indexation. Dans ce contexte, l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document.

Parmi les vocabulaires contrôlés, les thésaurus sont très utilisés dans les centres documentaires pour modéliser la connaissance dans un domaine car ils comportent des propriétés potentiellement exploitables. Les règles sémantiques, terminologiques et syntaxiques caractérisant ce type de structure sont autant d'outils qui permettent de modéliser et structurer la connaissance dans un domaine cible. Le fait que les informations de description choisies par les experts bibliothécaires pour décrire le contenu des documents soient extraites d'un vocabulaire contrôlé, régulièrement structuré sous forme d'un thésaurus, nous permet d'exploiter le travail d'experts notamment sur le choix de la sémantique, de la terminologie et la syntaxe des termes sélectionnés pour décrire des documents.

Cependant, du point de vue de la représentation des connaissances, les thésaurus ont un faible degré de formalisation et limitent les possibilités de représentation et de formalisation de la connaissance liée au fonds documentaire traité. La distinction entre un concept et sa lexicalisation n'est pas clairement établie et les relations de même niveau ne peuvent être explicitées précisément. Cela s'explique notamment par le fait que le thésaurus, initialement défini pour être utilisé dans le monde des bibliothèques et médiathèques, est utilisé dans la phase d'indexation de documents comme une terminologie complète ou un ensemble de catégories. Les relations entre termes indiquent parfois l'utilisation du thésaurus et non la sémantique réelle entre deux termes. En d'autres termes, la sémantique d'une relation entre deux termes est dépendante du contexte d'utilisation du thésaurus et dans le cas des centres documentaires de type bibliothèques, la sémantique est parfois guidée par le travail d'indexation des bibliothécaires.

Les ontologies ne posent pas ce type de problème. L'ensemble des définitions que

nous avons cité section 2.2.2.1 (*cf.* page 32) nous permettent d'identifier un ensemble d'éléments constituant une ontologie qui n'apparaissent pas dans les autres vocabulaires contrôlés. De façon simplifiée, une ontologie correspond à un vocabulaire contrôlé et organisé et à la formalisation explicite des relations créées entre les différents termes du vocabulaire. En effet, il est possible dans la définition d'une ontologie d'intervenir sur la syntaxe du langage d'indexation pour définir des concepts (pouvant ensuite être caractérisés), ainsi que de nouvelles relations entre ces concepts. On peut ainsi parler de langage formel, c'est-à-dire une grammaire qui définit la façon dont les termes peuvent être utilisés entre eux. Dans un modèle comme l'ontologie, les connaissances sont toujours vraies, quel que soit le niveau de description. Par exemple, on dira dans l'ontologie qu'un cheval est un animal (car c'est toujours vrai), mais on ne lui donnera pas une couleur, car cela change d'un cheval à un autre. Lorsqu'on utilise une ontologie pour décrire un domaine ou une application précise, il est alors possible de factoriser la description de plusieurs occurrences d'un concept. Par exemple, dans une ontologie décrivant les constructions, on attache la localisation d'un château aux instances et non au concept *château* car cette connaissance est propre à l'instance. Par contre, on factorise la connaissance « un château est un monument » entre les concepts *château* et *monuments* dans l'ontologie car c'est toujours vrai. Aussi, l'ontologie permet d'effectuer des inférences, limitant ainsi la quantité d'informations à structurer. Si nous reprenons l'exemple des châteaux, indiquer dans l'ontologie qu'un *château fort* est un *château* implique indirectement qu'un *château fort* est un *monument*. Cette relation, appelée relation transitive « $\mathcal{R}$ » sur E en mathématiques, se définie par le fait que lorsqu'un premier élément de E est en relation avec un deuxième élément lui-même en relation avec un troisième, le premier élément est aussi en relation avec le troisième (*cf.* figure 2.5).

$$\forall(x, y, z) \in E^3, [(x\mathcal{R}y) \wedge (y\mathcal{R}z)] \Rightarrow (x\mathcal{R}z)$$

FIGURE 2.5 – Représentation formelle mathématique de la relation de transitivité

Enfin, les connaissances dans une ontologie de domaine ou d'un niveau d'abstraction plus élevé peuvent être utilisées et réutilisées entièrement ou partiellement. Ce modèle sémantique est alors un moyen efficace de modéliser le travail d'indexation réalisé par les experts bibliothécaires, et de l'utiliser dans divers systèmes, sans avoir à redéfinir à chaque fois des règles et informations communes.

De ce fait, nous souhaitons mettre en place une méthodologie pour extraire cet ensemble de connaissances structurées et les « projeter » le plus automatiquement possible dans une structure de type ontologie. L'ontologie propose les outils pour répondre aux limites identifiées dans les autres vocabulaires contrôlés : manque de sémantique dans la définition des propriétés des concepts ainsi que dans la définition des relations. Nous positionnons nos travaux dans la construction d'une ontologie légère. Nous souhaitons construire l'ontologie de façon automatique et l'intégrer ensuite dans un système de recherche d'informations destiné à tout type d'utilisateurs et cela nous contraint à ne pas

complexifier la structure de l'ontologie par des axiomes ou des restrictions. En accord avec [Bac00] précisant que l'usage prévu de l'ontologie contraint et encadre sa construction pour un domaine cible, nous cherchons à construire une ontologie légère de domaine, le domaine dans notre cas étant le territoire. Notons tout de même que cette conceptualisation est souvent qualifiée de partielle car il semble présomptueux de croire pouvoir formaliser, dans une même structure sémantique, toute la complexité d'un domaine.

Nous allons aborder dans le prochain chapitre les principales méthodes relatives à la construction d'ontologies.

Chapitre 3

Construction d'ontologies

Sommaire

3.1 « Critères » reconnus pour construire une ontologie . . . 41
3.2 Méthodes de construction d'ontologies 43
3.2.1 Construction manuelle d'ontologies 44
3.2.2 Construction semi-automatique et automatique d'ontologies 48
3.3 Les langages autour de l'ontologie 54
3.3.1 Langages pour la représentation de connaissances 54
3.3.2 Langages associés pour la gestion des ontologies 54
3.3.3 Les langages à balises définis autour du Web Sémantique . 58
3.4 Discussion 65

Nous cherchons à construire une ontologie légère de domaine offrant une représentation d'un domaine cible implicitement décrit dans un fonds documentaire. Nous présentons maintenant les principales méthodes existantes visant à créer une ontologie de domaine et qui nous semblent pertinentes dans nos travaux. Nous en profitons pour présenter tout d'abord les règles, définies sur la base des travaux de [Gru93], généralement respectées dans le processus de création d'ontologie. Enfin, nous énumérons un ensemble d'outils permettant de formaliser une ontologie dans un langage de représentation des connaissances.

3.1 « Critères » reconnus pour construire une ontologie

Différents critères, définis par [Gru93] et régulièrement intégrés maintenant dans la phase de construction d'une ontologie, permettent de mettre en évidence des aspects importants d'une ontologie :

– La clarté et la complétude : les termes utilisés dans l'ontologie doivent être définis de façon objective et accompagnés d'une documentation en langage naturel. Une définition complète regroupant plusieurs conditions essentielles et suffisantes est privilégiée à une définition formée par exemple d'une seule condition, pouvant paraître incomplète ensuite pour les utilisateurs de l'ontologie ;

- La cohérence : pas de contradiction possible entre les définitions ;
- L'extensibilité : les extensions doivent être anticipées. On ne doit pas interférer dans le fondement d'une ontologie. L'ajout d'un terme ou d'un concept par exemple ne doit pas engendrer la modification de la structure et des définitions existantes ;
- Une déformation de l'encodage minimal : l'influence de la spécification sur la conceptualisation doit être évitée autant que possible ;
- Un engagement ontologique minimal : l'ontologie n'a pas pour objectif de répondre à tous les problèmes d'un domaine, elle doit intégrer le vocabulaire nécessaire pour représenter ce pour quoi elle est définie. L'ontologie doit caractériser au minimum la signification des termes qui la constituent, offrant la possibilité ultérieure de la spécifier ou de l'instancier en fonction des besoins.

Ces critères s'imposent d'autant plus qu'ils sont validés par bon nombre d'autres chercheurs. Parmi eux, Mike Uschold [Usc96] indique notamment que « les critères de Gruber pour construire des ontologies sont pertinents et peuvent être intégrés dans toute autre méthodologie ». Différents travaux [BGM96, BLC96, AGPTP98], visant à créer des ontologies, ont enrichi la liste des critères généralement acceptés maintenant dans la communauté.

- Principe de distinction ontologique [BGM96] : les classes dans une ontologie doivent être disjointes. Le critère utilisé pour isoler le noyau de propriétés considérées comme invariables pour une instance d'une classe est appelé le critère d'identité.
- Modularité [BLC96] : la modularité mesure la possibilité de découper l'ontologie en parties (modules indépendants) pouvant être utilisées par d'autres ontologies. Un module est défini comme étant une partie ou une sous-hiérarchie relativement indépendante. La modularité facilite la maintenance et l'enrichissement de la structure de même que la réutilisation de l'ontologie.
- Diversification des hiérarchies [AGPTP98] : certains chercheurs comme [MKSK00] sont opposés à l'idée d'héritage multiple en ingénierie ontologique. Il semble plus aisé d'ajouter de nouvelles connaissances et notamment de nouveaux concepts lorsque l'héritage multiple est évité, laissant place à un nombre plus important de classifications.
- Distance sémantique minimale [AGPTP98]. Il s'agit de la distance minimale entre les concepts fils d'un même concept père. Les concepts similaires sont groupés et représentés comme concepts fils d'un concept père, et doivent être définis en utilisant les mêmes propriétés, considérant que les concepts qui sont moins similaires sont représentés plus loin dans la hiérarchie.
- Normaliser les noms [AGPTP98]. Ce principe indique qu'il est préférable de normaliser les noms autant que possible (par exemple dans la gestion du singulier-pluriel).

Nous remarquons que l'objet ontologie dans le processus de création d'ontologies est l'aboutissement d'un travail important respectant un ensemble de critères qui font maintenant référence en ingénierie ontologique. Un nombre important de travaux propose une méthodologie pour construire une ontologie. Nous verrons cependant que si un cadre commun semble apparaître entre les différentes méthodes recensées, il n'existe pas à ce jour de méthode unique pour créer une ontologie permettant de répondre aux besoins

de tout type d'utilisateurs dans les différents domaines.

3.2 Méthodes de construction d'ontologies

Depuis plusieurs années, les ontologies sont créées et utilisées dans le domaine de l'Ingénierie des Connaissances (IC) et notamment leur représentation. Le champ d'application est très large [Gru93] : d'une manière générale dans l'indexation et la recherche d'information [Gen00], et plus particulièrement dans le domaine médical [Zwe94], dans le domaine touristique [VFM01], dans le domaine de l'éducation [MIS97], dans le domaine de l'héritage culturel [NV06], dans le domaine de la géographie [SM01]. Guarino [Gua98] présente aussi les utilisations des ontologies dans de multiples domaines, du génie logiciel à l'intelligence artificielle, et naturellement, en ingénierie des connaissances.

La phase de création d'ontologies est un processus complexe et il n'existe pas encore de normes ou de méthodes consensuelles en ingénierie des connaissances. Cependant, de plus en plus de travaux proposent des méthodes de construction d'ontologies. Comme l'indique [Ban07], il existe une quantité importante de méthodes permettant de représenter les connaissances :

> « Bien que la plupart des méthodologies initient le processus de construction par l'identification, puis l'organisation et la structuration des concepts et des relations à représenter, les ontologies réalisées sont très différentes les unes des autres. Faut-il faire l'hypothèse qu'il y ait autant de manières de représenter les connaissances d'un domaine qu'il y a d'ontologues ? »

Nous allons cependant présenter les principales méthodes afin d'identifier une méthodologie ou des éléments de méthologies existantes que nous souhaitons appliquer dans notre approche. Des travaux tels que [CFLGP03], [FLGP02], [FL99] et [PM04] en font une synthèse. Nous relevons un certain nombre de travaux visant à proposer une méthode de création d'ontologies : CYC [LDG90], CommonKADS [HMW+93], TOVE [GF94, GF95], Entreprise Ontology [UK95], METHONTOLOGY [FLGPJ97], Sensus [SRKR97], OTK [SSSS01] [Miz98], TERMINAE [AGBS00, AGDS08], B. Bachimont et ses collègues ont proposé une méthode dans le cadre du projet MENELAS [30] [BBCZ95], méthode qui fut ensuite théorisée [Bac00]. La construction d'ontologies peut se réaliser selon trois modes : le mode manuel (cf. 3.2.1), le mode automatique (cf. 3.2.2) et le mode mixte (cf. 3.2.2) qui sont présentés ci-après.

30. Projet européen piloté de 1992 à 1995 ayant pour objectif principal la conception et l'implémentation d'un système pilote capable d'accéder à des rapports médicaux rédigés en langage naturel dans trois langues, l'anglais, le français et le néerlandais. Ce système devait pouvoir analyser le contenu de rapports médicaux (comptes rendus d'hospitalisation ou CRH) et l'archiver dans une base de données sous la forme d'un ensemble de structures conceptuelles (graphes conceptuels de Sowa [Sow84]) pour consultation. http ://www.med.univ-rennes1.fr/Menelas/french_mnl.html

3.2.1 Construction manuelle d'ontologies

La méthode manuelle donne la possibilité aux experts de définir, de façon consensuelle, les concepts et relations en fonction de leur vision du domaine ou de l'application cible. Ce type de méthodologie est encore régulièrement utilisé pour créer ou étendre une ontologie existante pour des domaines spécialisés. Cependant, cette méthode est très coûteuse en temps et en moyens et pose aussi et surtout des problèmes de maintenance et de mise à jour.

Parmi les travaux existants dans la création manuelle d'ontologie en ingénierie des connaissances, nous présentons brièvement Enterprise Ontology, TOVE, OTK et Mikrokosmos, définies dans les années 1995, 1996 apparaissant plus comme des recommandations que comme de réelles méthodologies.

3.2.1.1 Enterprise Ontology

Uschold et King [UK95] ont proposé une méthode de construction d'ontologies basée sur l'expérience acquise lors du développement de l'ontologie d'Entreprise, *the Enterprise Ontology*[31]. La méthode repose sur trois grandes étapes : (i) Identification du « pourquoi » de l'ontologie ; (ii) Construction de l'ontologie (identification des concepts clef ; modélisation informelle ; formalisation) et intégration d'ontologies existantes ; (iii) Evaluation et documentation de l'ontologie.

Ces recommandations, offrant un cadre de méthodologie, s'inspirent du développement de Système à Base de Connaissances (SBC). Le but de la première étape est de clarifier la finalité de l'ontologie à créer (réutilisation, partage, utilisation comme une partie d'une base de connaissances, etc.) ainsi que les utilisateurs potentiels de l'ontologie. L'expert peut définir un modèle conceptuel de l'ontologie en partant de sa connaissance des concepts du domaine. Les recommandations peuvent ensuite guider l'expert dans une démarche de généralisation et d'instanciation des concepts connus. Les étapes suivantes permettent de formaliser et d'implémenter le modèle conceptuel. En effet, dans cette recommandation, Uschold et King proposent trois approches qui seront régulièrement reprises par la suite :

- Approche descendante : partir de concepts abstraits que l'on spécialise en concepts plus spécifiques ;
- Approche ascendante : partir de tous les concepts spécifiques que l'on généralise en concepts abstraits ;
- Approche intermédiaire : les concepts se structurent autour de concepts intermédiaires, ni trop généraux, ni trop spécifiques.

Cependant, les étapes sont décrites de façon abstraite et les sous-tâches ne sont pas précisées (ex : comment identifier les concepts clef ? Quel langage de formalisation utiliser ?). Le reproche majeur fait à ces travaux concerne le manque de conceptualisation. Les travaux présentés ci-après autour de METHONTOLOGY ont levé ce manque.

31. http ://www.aiai.ed.ac.uk/project/enterprise

3.2.1.2 METHONTOLOGY

METHONTOLOGY est également l'une des méthodologies les plus représentatives. Le processus de construction d'ontologie envisagé, décrit dans les travaux suivants [FLGPJ97] [LGPSS99], [GPCFL04], assiste « l'ontologue » dans la capture et la structuration de l'information. Les quatre étapes principales sont : (i) spécification (but et utilisateurs visés de l'ontologie), (ii) conceptualisation (structuration du domaine au niveau des connaissances), (iii) formalisation (traduction automatique du modèle conceptuel en utilisant des traducteurs) et (iv) implémentation (expression du modèle formel à l'aide d'un langage d'implémentation).

METHONTOLOGY caractérise les ontologies au niveau des connaissances et travaille à partir de représentations intermédiaires des connaissances lors de la phase de conceptualisation, sans nécessiter une connaissance *a priori* de concepts. WebODE [BFGPGP98] est l'environnement qui implémente METHONTOLOGY. Des guides sont proposés, constitués d'un ensemble de tables prédéfinies facilitant l'acquisition des connaissances et leur conceptualisation. Le schéma figure 3.1 tiré de [FLGPJ97] permet de comparer les deux méthodologies précédentes.

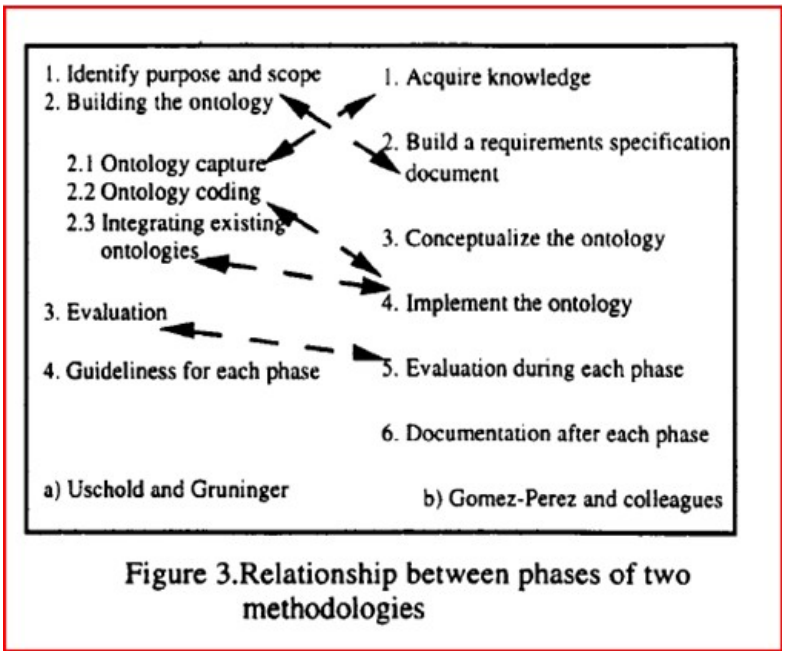


FIGURE 3.1 – Comparaison de méthodologies [FLGPJ97]

Qu'est ce que la conceptualisation ? Le fait de structurer la connaissance du domaine dans un schéma conceptuel (renommé dans le domaine des bases de données). Pour METHONTOLOGY, cette phase consiste à construire un glossaire exhaustif de tous les

termes du vocabulaire du domaine. A partir de ces termes, des groupes de concepts et verbes sont réalisés afin de bâtir un arbre de classification des concepts et un diagramme des verbes qui sont ensuite reliés via des règles et des formules.

« *...The conceptualization activity in METHONTOLOGY organizes and converts an informally perceived view of a domain into a semi-formal specification using a set of intermediate representations (IRs) based on tabular and graph notations that can be understood by domain experts and ontology developers...* » [CFLGPLC05].

METHONTOLOGY met notamment l'accent sur la conceptualisation et sur la réutilisation.

Une mise en œuvre de cette méthodologie est détaillée dans [CFLGPLC05] portant sur la construction d'une ontologie du droit. Au niveau logiciel, un éditeur nommé **ODE** (Ontology Design Environment), et développé à l'Université de Polytechnique de Madrid, permet de mettre en place la méthodologie. Son successeur pour le Web **WebODE** [ACFLGP03] propose de couvrir l'ensemble des sous-domaines de l'ingénierie ontologique à travers les différentes activités liées au cycle de vie d'une ontologie : acquisition de connaissances à partir du Web, édition d'ontologies, test de la consistance d'une ontologie, alignement et fusion d'ontologies, import et export dans des formats variés. Le modèle de représentation de connaissances utilisé associe un modèle de type frame (concepts et attributs) avec des relations entre concepts. L'éditeur permet aussi la création de modèles conceptuels *ad hoc*.

3.2.1.3 TOVE

Le projet TOVE [32] [GF95] a pour objectif principal de créer un modèle d'entreprise exprimé dans une ontologie, permettant à un système utilisant cette ontologie de gérer les connaissances liées à l'organisation et aux activités des entreprises. Le projet TOVE propose un processus de création en cinq étapes : (i) identification de scénarios (problèmes) dépendants d'une application ; (ii) formulation de questions informelles (basées sur les scénarios) auxquelles l'ontologie doit permettre de répondre ; (iii) spécification d'une terminologie à partir des termes apparaissant dans les questions ; (iv) spécification formelle (en KIF [33] [GF92]) des axiomes et des définitions pour les termes de la terminologie ; (v) évaluation de la complétude de l'ontologie.

La principale critique formulée envers TOVE est qu'elle se limite à la formalisation des connaissances, sans proposer une étape préalable de modélisation conceptuelle de l'ontologie. De plus, les étapes intégrant le processus de création d'ontologies ne sont pas spécifiées et sont peu décrites. Avec TOVE, l'expert doit en effet structurer le domaine en utilisant des concepts exprimés au moyen d'un langage formel (basé sur la logique du premier ordre), directement à partir du corpus.

32. TOronto Virtual Enterprise
33. Knowledge Interchange Format : langage destiné à faciliter la communication entre systèmes disparates.

3.2.1.4 OTK : On-To-Knowledge

On-To-Knowledge [SSSS01] est un projet proposant d'appliquer les ontologies aux informations et ressources textuelles disponibles sur internet, en extranet ou encore en intranet. L'objectif est d'améliorer la qualité de la gestion de la connaissance au sein des organisations imposantes et distantes. Cette méthodologie tend à construire une ontologie très dépendante de l'application, qui tient compte du cycle de vie et de la future utilisation de l'ontologie.

La méthode On-To-Knowledge propose cinq étapes pour construire une ontologie : (i) Etude de faisabilité (appliquée sur l'application entière et sert de base à l'étape suivante) ; (ii) Description des spécifications des besoins de l'ontologie (domaine, objectifs, sources de connaissance, etc.) ; (iii) Raffinement (production d'une application conformément aux spécifications données à l'étape de Kickoff et production d'un document de spécifications des besoins d'ontologie) intégrant une phase de validation des experts et une étape de formalisation) ; (iv) Evaluation (preuve de l'utilité du développement de l'ontologie et les applications associées) ; (v) Maintenance (maintenance de l'ontologie effectuée comme une partie de l'application).

3.2.1.5 Ontogeny

[MDH05] proposent une méthode de création d'ontologies du domaine sur deux niveaux, conceptuel et logique, le conceptuel étant dédié à la compréhension humaine, le logique pour la machine. Afin de ne pas perdre d'information durant la traduction du niveau conceptuel vers le niveau logique, les auteurs ajoutent un niveau intermédiaire permettant d'exprimer l'ontologie en logique de 1er ordre.

La méthodologie comprend quatre étapes : (i) Préparation (identifier les besoins et collecter les données), (ii) Identification de concepts et structuration ; (iii) Evaluation de l'ontologie proposée ; (iv) Documentation de l'ontologie conceptuelle.

Le traitement est manuel, voire semi-automatique. Les auteurs classent leur méthodologie dans les méthodologies universelles pouvant servir à créer tout type d'ontologie.

3.2.1.6 Autres méthodes

D'autres méthodes, proposant un cadre plus ou moins précis, sont régulièrement reprises dans les travaux en ingénierie ontologique et méritent d'être cités.

La méthode Mikrokosmos est née du développement de l'ontologie Mikrokosmos [MN95, Mah96] au cours duquel une trentaine de lignes directrices ont été mises au point pour le développement. Ces lignes directrices apparaissent donc comme des recommandations à suivre lors du développement d'ontologies.

L'approche Ontolingua intègre des recommandations sur l'affichage, le développement, la maintenance et le partage des ontologies, en tentant de privilégier la réutilisation de l'ontologie. Un guide d'utilisation est disponible sur le serveur Ontolingua [FFR97].

Une autre méthode nommée CommonKADS [HMW+93] met l'accent sur l'analyse et la modélisation des connaissances partagées, dans laquelle le cycle de vie de la connaissance et l'interaction système/utilisateurs prennent une place centrale. Cette approche est largement utilisée pour le développement des systèmes à base de connaissances dans lesquels les ontologies jouent un rôle important. Elle préconise le principe de développement modulaire, qui est de plus en plus populaire en ingénierie ontologique. CommonKADS propose une approche structurée pour la documentation des connaissances. Un langage de représentation formelle nommé CommonKADS est proposé pour formaliser la méthodologie du même nom, considérée comme un standard européen pour l'analyse des connaissances et le développement de systèmes orientés connaissances.

Enfin, la méthode Sensus [SRKR97] propose une méthodologie visant à construire une ontologie de domaine à partir d'une ontologie générique, l'ontologie Sensus. Le projet Cyc [LDG90] présente un scénario permettant de développer une ontologie couvrant l'ensemble des connaissances de sens commun. Si le principe est intéressant, il semble présomptueux de penser qu'il soit possible de définir une ontologie générique pour l'ensemble des domaines et des applications.

3.2.1.7 Premiers bilans

Cette revue de méthodologie permet de distinguer deux niveaux de modélisation :
- une modélisation pour donner du sens, autrement dit, une modélisation des connaissances ontologiques conduisant à la définition d'une ontologie conceptuelle ;
- une modélisation pour implémenter un système conduisant à une ontologie computationnelle.

Bien que manuelles, les approches présentées imposent un cadre pour la définition d'ontologies. Elles sont d'ailleurs généralement reprises en partie dans les méthodologies de construction semi-automatiques ou automatiques d'ontologies que nous allons maintenant présenter.

3.2.2 Construction semi-automatique et automatique d'ontologies

La méthode automatique implique l'utilisation de techniques d'extraction des connaissances pouvant provenir de supports divers. Les éléments extraits candidats à constituer l'ontologie sont vérifiés par des inférences réalisées par des modules/programmes informatiques. La méthode mixte ou semi-automatique combine les deux approches en proposant à des utilisateurs, généralement experts du domaine, de vérifier la structure sémantique définie automatiquement dans une première étape. Généralement, la construction d'ontologies dite automatique implique néanmoins une vérification de la terminologie et/ou de la structure par des experts du domaine ou de l'application visée.

Parmi les nombreux travaux existants dans la création semi-automatique voire automatique d'ontologies, nous centrons notre analyse sur les approches reposant sur l'analyse de textes [Bac00, BIT02, Ban07, VOS03, CV05, AGBS00, SBAG02] ainsi que sur les travaux de [WSWS01, Mäd02, SLL+04, SGD04, Her05, CHGM06] s'appuyant également sur des ressources sémantiques structurées.

3.2.2.1 Construire une ontologie à partir de ressources textuelles

Le choix du corpus est une étape primordiale dans le processus de création d'ontologies, comme l'affirme [Bac00, WC03] en ajoutant qu'il existe un lien étroit qui unit le corpus et les résultats obtenus. En ingénierie ontologique, les textes apparaissent comme des sources de connaissances intéressantes, et leur analyse automatique offre un moyen efficace pour accélérer la construction d'ontologies, et pour valider par la même occasion le fait qu'elles reflètent le vocabulaire de corpus ou de domaines cibles.

Les méthodes de construction d'ontologies à partir de textes privilégient souvent l'analyse du texte proprement dit, que ce soit selon une approche statistique ou linguistique [NN06, AGDS08, Bac00, Mäd02, BCM05]. Nous présentons les méthodes ARCHONTE, KAON et TERMINAE qui comptent parmis les méthodes les plus citées :

- **ARCHONTE** (ARCHitecture for ONTological Elaborating) est proposée pour construire des ontologies en s'appuyant sur la sémantique différentielle [Bac00, BIT02].
  « *La sémantique différentielle permet de décrire les unités entre elles par les identités qui les unissent et les différences qui les distinguent* » [Bac00].
  Ici, la construction d'une ontologie comporte trois étapes qui définissent 3 niveaux d'ontologie : (i) choisir les termes pertinents du domaine et normaliser leur sens puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarité et de différence que chaque concept entretient avec ses concepts frères et son concept père. On obtient alors *l'ontologie régionale* ; (ii) formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation, etc. On obtient alors *l'ontologie référentielle* ; (iii) l'opérationnalisation dans un langage de représentation des connaissances. On obtient *l'ontologie computationnelle*.
  [Ban07] met en œuvre cette méthodologie pour construire une ontologie de la pneumologie. Un ensemble de traitements pour la construction de hiérarchies terminologiques fondées sur deux méthodologies de Traitement Automatique de la Langue (TAL), adaptées chacune à un type et genre de corpus ont été utilisées. Il s'agit de l'analyse distributionnelle sur un corpus redondant et riche en termes spécialisés, et de l'extraction par patrons lexico-syntaxiques sur un corpus didactique à la structure régulière. La complémentarité de ces deux modes d'analyse de la langue facilite la structuration hiérarchique des concepts de l'ontologie.
  La méthode ARCHONTE insiste sur le fait que les ontologies vraiment intégrables dans des applications ne peuvent être que des ontologies régionales, dans lesquelles le choix et l'organisation des concepts tiennent compte de leur utilisation dans l'ap-

plication. Selon cette analyse, reprise dans les travaux du groupe TIA [34] et dans TERMINAE, les ontologies n'ont pas de portée universelle mais ce sont des représentations de définitions consensuelles des concepts retenus au sein d'un domaine en vue d'un objectif précis [Cha02].

- **TERMINAE** [AGBS00,SBAG02,AGDS08] est une approche permettant de construire des ontologies dans un sous domaine de l'ingénierie des connaissances impliquant des traitements linguistiques afin de pouvoir sélectionner les concepts, leurs propriétés, les relations et leur regroupement. Les termes, identifiés de façon automatique, sont regroupés suivant leur contexte et facilitent la création de concepts et de relations sémantiques. Les concepts et relations sont ensuite formalisés dans un modèle sémantique. La méthodologie proposée se décompose en cinq étapes : (i) analyse des besoins ; (ii) constitution du corpus ; (iii) analyse linguistique ; (iv) normalisation en réseau sémantique ; (v) formalisation du réseau sémantique.
  TERMINAE insiste sur l'importance de faire appel à des experts du domaine à chaque étape du processus pour valider les résultats intermédiaires. La méthodologie met également en avant l'importance de l'étape d'analyse linguistique dans le processus de construction d'ontologies. Une plateforme logicielle du même nom, documentée [35], est développée et maintenue pour permettre à des experts de construire une ontologie. La plateforme intègre notamment des outils d'analyse linguistique applicables sur un corpus de référence : un premier outil nommé LEXTER [BGMG96] permet d'extraire les termes candidats à partir de leurs dépendances syntaxiques, et un deuxième nommé Caméléon [SAG99] permet d'extraire des relations entre termes à partir de patrons linguistiques.
- **KAON** [36] [VOS03] est définie pour gérer toute la vie d'une ontologie. Elle est composée de six phases : (i) création, (ii) stockage, (iii) raffinage, (iv) exploitation, (v) maintenance, et (vi) application d'ontologies. Cette approche se spécifie notamment par le fait qu'elle intègre un module nommé TextToOnto [37] [CV05] proposant des techniques de fouille de texte (extraction de termes, extraction de relations conceptuelles et algorithme de validation des éléments sélectionnés) sur des corpus de texte préalablement sélectionnés. L'intérêt est ici de pouvoir définir des méthodes statistiques ou des expressions régulières pour identifier les termes candidats à devenir des concepts dans l'ontologie. Il est ensuite possible de définir des règles (degré de proximité des termes, etc.) pour extraire des relations qui pourront permettre de construire l'ontologie. TextToOnto permet également de déduire une ontologie propre à un domaine à partir d'une structure terminologique générique, comme WordNet.
  Un environnement modulaire du même nom est développé pour permettre la construction et la maintenance d'ontologies. Il intègre la méthode de construction d'ontologies KAON et permet l'édition des hiérarchies de concepts et de relations

34. Terminologie et Intelligence Artificielle : http ://www-test.biomath.jussieu.fr/TIA/
35. lien : http ://www-lipn.univ-paris13.fr/ szulman/TERMINAE.html
36. KArlsruhe ONtology, http ://kaon.semanticweb.org/
37. http ://www.architexturez.net/FILES/archive/sub.gate.archive/kaon/TextToOntoPaper.pdf

dans le cadre de la logique des frames, ainsi que l'expression d'axiomes algébriques. Orientée vers l'utilisation des ontologies sur le Web, l'application KAON Portal permet la recherche et le parcours d'ontologies via un navigateur Web. Les modules les plus importants de cette suite sont : API, Query, Serveurs (d'ontologie et d'application), Générateur de portails web (basés sur les ontologies), Éditeur d'ontologie (construction et maintenance).

Parmi les méthodes d'extraction d'informations s'appuyant sur des textes, les approches statistiques consistent à étudier généralement les termes co-occurrents par analyse de leur distribution dans le corpus [AAHM00] ou encore par des mesures calculant la probabilité d'occurrences d'un ensemble de termes [VFM01, NH04]. Les relations peuvent être identifiées par des calculs de similarité entre leurs contextes syntaxiques [Hin90, Gre94], par prédiction à l'aide de réseaux bayésiens [WN07] ou de techniques de Text Mining [GKN09], ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage [GLR06]. Ces méthodes sont efficaces, mais elles nécessitent une intervention humaine fastidieuse pour le positionnement des concepts dans l'ontologie, ou n'identifient pas toujours la sémantique de la relation. Ces éléments, liés au fait que nous travaillons sur des fonds documentaires constitués de quelques milliers de documents tout au plus (ce qui limite le nombre de concepts et relations à créer), nous incitent à nous tourner vers les approches linguistiques

[MS01] (2001 - citation [GPMM03], p. 11 tiré de [Lav07]) proposent une classification des principales approches utilisées pour la construction d'ontologie à partir de corpus de textes non structurés :

- extraction à base de patron : une relation est reconnue quand une séquence de mots dans un texte correspond à un patron textuel ;
- association par règles : mise en correspondance de termes ou de concepts par le biais de règles de type implication (si X alors Y) ;
- regroupement conceptuel : les concepts sont regroupés selon la distance sémantique entre chacun afin de former des hiérarchies ;
- élagage d'ontologie : des textes généraux et spécifiques à un domaine donné sont utilisés afin d'éliminer des concepts qui ne sont pas spécifiques au domaine ;
- et apprentissage de concepts : une taxonomie donnée est mise à jour de façon incrémentale avec de nouveaux concepts dont les termes correspondants sont extraits à partir de textes.

Les travaux d'analyse linguistique ont trouvé au fur et à mesure de l'évolution des technologies informatiques des applications dans des domaines tels que la modélisation et la recherche d'information. Dans ce cadre, l'évolution des outils de Traitement Automatique de la Langue (TAL) permet maintenant de proposer une analyse fine basée sur l'interprétation de la sémantique contenue dans les documents textuels. Dans [Sag06] est dressé un état des moyens existants pour analyser automatiquement la sémantique de la langue française. Une chaîne de traitements ressort d'un ensemble de travaux s'appliquant sur l'analyse linguistique [AFG03, Bil06, Les07], se décomposant selon les

sous-processus d'analyse et d'extraction d'information suivants :

1. la lemmatisation pour segmenter les mots et identifier leur lemme ;
2. l'analyse lexicale et morphologique pour la reconnaissance des mots ;
3. l'analyse syntaxique, basée sur des grammaires, afin de trouver les relations entre les mots : permet d'identifier le rôle des termes ou des syntagmes dans la phrase ;
4. enfin l'analyse sémantique pour réaliser une interprétation plus spécifique sur les syntagmes retenus [BCEM03, WB06, WB07] : l'objectif est ici d'identifier le sens potentiel véhiculé par un mot ou un groupe de mots.

Dans le domaine de la construction d'ontologies, l'approche linguistique peut être utilisée pour découvrir des concepts via des règles d'association. [PS09] fait un état de nombreux travaux traitant de l'identification et de l'annotation des entités nommées. Une « entité nommée » est une expression linguistique autonome d'un texte qu'il est possible de classer sur le plan sémantique [Gry08, Ehr08]. [Poi03] précise cette définition en mettant en avant la notion de catégories : *ensemble des noms de personnes, d'entreprises et de lieux présents présents dans un texte donné.*

Plus largement, l'approche linguistique est mise en application dans le cadre d'identification de relations. Dans ce cas, elle fait appel à des analyses syntaxiques ou des calculs de dépendance pour identifier les relations argumentatives *(sujet, verbe, objet)* [JKT97, BL02], ou définit des patrons lexico-syntaxiques pour reconnaître les marques linguistiques des relations sémantiques [AGBS00, AB08], ou s'appuie sur la combinaison des deux approches [LGN07]. Ainsi, la sémantique des relations est bien identifiée, mais la variabilité de leur sémantique et de leur expression en corpus oblige à multiplier les patrons et rend l'approche coûteuse en temps et en ressources, à la fois lors de la conception mais aussi lors de l'exécution des patrons sur le corpus.

Ces techniques s'appliquent au niveau interne de la phrase, alors qu'il existe aussi des relations sémantiques plus diffuses entre différentes unités textuelles repérées. L'analyse doit alors se situer au niveau du texte. Ces liens peuvent s'exprimer à l'aide de relations de discours [AHB01] ou à travers la structure du texte, structure qui peut être matérialisée entre autres par la mise en forme typographique et dispositionnelle [VL01] du texte, ou par l'annotation sémantique du contenu à l'aide de langage de balises de type XML [RR06, GHM+07, KAG09].

Une approche différente consiste à s'appuyer sur une ressource structurée du domaine pour construire une ontologie. L'intégration de la connaissance du domaine d'intérêt dans le processus de création d'ontologies peut permettre d'identifier et de manipuler des informations pertinentes supplémentaires [Zar04]. Cette connaissance est dans ce type d'approche modélisée sous forme d'un ensemble structuré d'informations tel que les thésaurus ou les ontologies.

3.2.2.2 Construction à partir de textes et de ressources sémantiques structurées

Dans [WSWS01], l'approche présentée permet de transformer le thésaurus de l'art et de l'architecture AAT en ontologie de domaine pour indexer des images. Cette approche est entièrement manuelle. Deux étapes d'identification de concepts et d'augmentation des concepts grâce à des propriétés permettent de définir cette ontologie. La méthode explicitée dans [SLL+04] repose sur trois étapes. Cette dernière a permis la transformation du thésaurus AGROVOC couvrant le domaine de l'agriculture, de la forêt, de la nourriture et des domaines reliés tels que l'environnement. L'originalité de l'approche se situe dans la phase d'apprentissage permettant d'extraire des relations supplémentaires afin d'augmenter la sémantique liée au thésaurus de base.

[Her05, CHGM06] simplifient l'opération de création d'ontologies à travers une approche permettant d'enrichir un thésaurus pour créer une ontologie à partir de sources de connaissances du domaine (vocabulaires, thésaurus, etc). Ces sources formalisées, contenant des termes représentant le domaine et (pour les thésaurus) des relations entre ces termes, apportent alors un plus sémantique indéniable à la représentation du domaine étudié. Les travaux de [Her05, CHGM06] se basent sur la méthodologie TERMINAE et plus précisément sur les étapes III (analyse linguistique), IV (normalisation en réseau sémantique) et V (formalisation du réseau sémantique) pour spécifier la méthode de transformation d'un thésaurus en une ontologie. La méthodologie proposée se décompose en 3 étapes :

1. Extraction d'informations du corpus ;
2. Identification d'un ensemble de concepts présents dans le corpus ainsi que leurs variations lexicales lorsqu'elles peuvent être identifiées ;
3. Structuration des concepts à partir des relations hiérarchiques et associatives identifiées dans le thésaurus et dans le corpus de référence. Un thésaurus se caractérise par le fait que le plus haut niveau hiérarchique est généralement composé de nombreux termes et afin d'organiser l'ontologie résultante à partir d'un niveau d'abstraction comportant un nombre limité de termes, une ontologie générique est utilisée dans cette étape.

La méthodologie proposée est adaptée lorsque le thésaurus initial est construit en respectant la sémantique de la relation « est un ». En revanche et comme le souligne [CHGM06], lorsque ce n'est pas le cas, une étape supplémentaire doit être ajoutée afin de distinguer les différentes relations telles que « est une partie de » ou « est une instance de ».

Si la création d'ontologies de domaine ou d'application « from scratch » requiert beaucoup de temps et des connaissances avancées du domaine cible (*cf.* section 3.2.1, page 44), ce type de méthodes dont l'objectif est de proposer une représentation sémantique enrichie d'un niveau formel plus élevé simplifie cette opération en utilisant un thésaurus, défini sur la base d'un vocabulaire contrôlé intégrant ce travail de conception conséquent. La conception d'ontologies à partir de thésaurus présente l'avantage de reposer

sur l'ensemble des termes qu'il contient et qui ont été identifiés par des experts comme étant représentatifs du domaine. Cependant, elle doit prendre en compte les différences fondamentales entre thésaurus et ontologie.

Comme indiqué dans les méthodologies précédemment présentées et notamment dans TERMINAE (*cf.* section 3.2.2.1 page 51), la phase de construction d'une ontologie intègre une étape de formalisation dans un langage afin de pouvoir ensuite l'exploiter dans des outils informatiques. Nous allons maintenant présenter les différents langages existants pour formaliser les ontologies.

3.3 Les langages autour de l'ontologie

3.3.1 Langages pour la représentation de connaissances

Nous avons vu lors de la description des différentes méthodes permettant de construire une ontologie, qu'il est nécessaire après la phase de conceptualisation de formaliser la représentation sémantique obtenue dans un langage compréhensible par la machine. Nous avons recensé plusieurs formalismes pour la représentation des connaissances dans le cadre d'ontologies que nous pouvons classer en deux catégories : les **formalismes relationnels** (orientés objet, cadres, réseaux sémantiques, graphes conceptuels, etc.) ; les **formalismes à base de logique** (logique propositionnelle, logique des prédicats, logique floue, etc.). Les formalismes à base de logique, plus anciens, sont depuis quelques années maintenant bien intégrés dans des systèmes informatiques. Cependant, bien que plus récents, les formalismes relationnels, offrant notamment un bon degré d'expressivité et de simplicité, ont connu une évolution fulgurante via l'avènement du Web. Lors de la phase de création d'une ontologie, le choix du formalisme doit être fait en fonction de critères tels que l'expressivité, la simplicité et l'efficacité qui sont attendues.

3.3.2 Langages associés pour la gestion des ontologies

Des langages propres à la gestion des ontologies (définition, stockage, traitement, etc.) par des systèmes informatiques ont été définis. Ces langages sont principalement issus des formalismes liés aux **réseaux sémantiques**. A l'origine, la notion de réseau sémantique est définie par Quillian [Qui68] pour modéliser le fonctionnement de la mémoire. Depuis lors, cette notion est appliquée dans plusieurs domaines [LKIR94]. De façon très synthétique, un réseau sémantique est un réseau dont les sommets sont des concepts et les arêtes des relations. Il y a donc plusieurs types d'arêtes. Par la suite, les relations peuvent encore posséder des propriétés (héritage, transitivité, symétrie, être elles-mêmes en relation, etc.). Il est possible de réaliser des inférences en fonction de la nature des liens. Cependant, ce type de définition ne concerne que la structure du graphe et ne permet pas d'ajouter de l'information sémantique. [Woo75, Bra77] mettent en avant le manque de précision de ce type de graphe qui mène à des confusions entre les relations et aussi entre les classes et individus. Pour tenter de palier ces manques de précision, de nombreux langages ont vu le jour. Ces langages peuvent être regroupés en

trois catégories que nous allons maintenant présenter de façon succincte : les langages basés sur la logique du 1er ordre, ceux à base de frames, et enfin les langages à balises orientés Web Sémantique.

3.3.2.1 Les langages s'appuyant sur la logique du premier ordre

La logique du premier ordre, inventé par [Fre94], est un langage formel comportant la syntaxe qui permet de distinguer les phrases logiques parmi les assemblages de sous-phrases, et la sémantique qui attribue une signification aux phrases logiques. Ce type de langage permet d'exprimer des faits et des règles de manière précise et n'autorise pas d'ambiguïtés. Les éléments constituant d'un langage du premier ordre sont les constantes pour désigner des noms spécifiques d'éléments, les variables pour désigner des entités, des prédicats pour désigner les règles d'assemblage entre constantes et variables aussi appelées fonctions propositionnelles. Chaque prédicat a sa propre arité et donne comme résultat vrai ou faux (exemple (Gauche(x, y)) retourne vrai si x est à gauche de y et faux sinon). Souvent le langage de premier ordre intègre la notion de fonctions pour désigner, à partir d'arguments, des entités. Les fonctions peuvent être vues comme des prédicats, sur des variables qui sont utilisées comme paramètres (Place (x) retourne la position de la couleur x).

Parmi les langages s'appuyant sur la logique du premier ordre, **KIF** [38] [Gin91] est un langage de bas niveau pour l'expression des ontologies. Pour palier le manque de conceptualisation, le langage KIF intègre des extensions pour représenter des définitions et des méta-connaissances. **ONTOLINGUA** [FFR97] est une extension de KIF utilisée dans le serveur d'édition d'ontologies du même nom Ontolingua [39]. ONTOLINGUA propose notamment un outil permettant d'inclure une ontologie dans celle en cours de construction. L'inclusion consiste à ajouter à l'ontologie courante les axiomes de l'ontologie à inclure, après traduction des axiomes [FFP+95]. La traduction consiste à établir une relation d'identité entre les termes des deux ontologies qui désignent les mêmes classes ou relations. Ces termes sont tous différents entre eux car préfixés par le nom de l'ontologie à laquelle ils appartiennent.

3.3.2.2 Les langages à base de frames

Les langages à base de frame [Min74] s'appuient sur les notions de prototype (ou frame ou encore schéma), d'objet, de classe et d'instance afin de représenter des éléments reliés entre eux. Dans ce type de langage, les frames sont des structures de données complexes qui représentent des concepts. Elles ont un nom et une série d'attributs appelés des slots. Les slots sont des propriétés du frame permettant de définir la structure de données. Par exemple, un concept peut nécessiter d'avoir un type, une durée, et on utilise

38. Knowledge Interchange Format
39. Ontolingua, développé au Knowledge Systems Laboratory de l'Université de Stanford. http ://www.ksl.stanford.edu/software/ontolingua/

dans ce cas les slots pour les représenter. Les frames sont organisées dans une hiérarchie suivant un lien de spécification. A la différence des langages objet, ces langages prennent en charge l'incomplétude et l'évolutivité des connaissances d'où le concept de prototype. Dans les langages à base de frames, la classification se fait en sélectionnant un élément représentatif (ou prototype) de la classe. Les instances appartenant à la même catégorie vont partager une certaine similitude avec ce meilleur représentant. On procède donc par appariement et non via un modèle de conditions nécessaires et suffisantes. L'intérêt des frames est qu'ils permettent de représenter la façon de penser d'experts en fournissant une représentation structurée et concise des relations utiles [FK85]. L'information peut être partagée entre plusieurs frames grâce à l'héritage.

Parmi les langages définis sur la notion de frame, **F-Logic** (Frame logic) [KLW95] est un langage de bases de données orienté objet qui combine les outils de modélisation des modèles orientés objet et l'expressivité des langages de bases de données. Un deuxième langage nommé **OCML** (Operational Conceptual Modelling Language) [DMCG99] propose une combinaison des frames et de la logique du premier ordre. Il permet de formaliser les concepts, la hiérarchie entre concepts, les relations de même niveaux, les fonctions, les axiomes et les instances. OCML est notamment utilisé pour implémenter les ontologies intégrées dans l'application Web WebOnto [MBD00].

3.3.2.3 Les graphes conceptuels

Les graphes conceptuels ont été introduits par [Sow84]. A l'origine définis pour représenter la sémantique du langage naturel, les graphes conceptuels ont évolué pour devenir des systèmes complets au sens de la logique. De façon générale, un graphe conceptuel est défini comme un graphe qui a deux sortes de nœuds :

- Les concepts qui représentent des entités, des attributs, des états, des événements ;
- Les relations conceptuelles qui symbolisent les liens qui existent entre deux concepts.

Les concepts et relations sont tous deux typés et il est possible d'attribuer aux relations le nombre de concepts qui leur sont reliés. La liste des types des concepts liés à une relation correspond à la signature de cette relation. Enfin, les types sont organisés en hiérarchies structurées par une relation de subsomption.

Dans les graphes conceptuels [CM92], nous pouvons distinguer différents niveaux de représentation que sont le niveau conceptuel et le niveau d'exécution. Le niveau conceptuel peut servir de base à un langage spécialisé de communication entre les spécialistes de différentes disciplines impliquées dans un travail cognitif commun. Le niveau d'exécution peut servir de base à un outil commun de représentation employé par plusieurs modules d'un système complexe. L'une des particularités de ce langage est de permettre de représenter des connaissances sous forme graphique. Comme l'indique [Bri04], *un graphe conceptuel peut être représenté selon différentes notations : Le format graphique appelée DF (Display Form), le format d'échange CGIF (Conceptual Graph Interchange Form) et le format linéaire (Linear Form).* Il est possible de définir une représentation équivalente en logique des prédicats à l'aide du langage KIF.

3.3.2.4 Logique de descriptions

Les logiques de description (LD) associent les notions provenant à la fois des réseaux sémantiques et des langages de frame. Des correspondances existent entre les LD et les formalismes précédemment cités comme le montrent notamment [SCM03, BN03].

La présence de catégories générales d'objets et de relations fait d'ailleurs partie de l'héritage conceptuel des schémas et des réseaux sémantiques. Les primitives de modélisation sont au nombre de trois : les concepts pour représenter les objets, les rôles pour formaliser les relations entre les objets et les individus pour représenter les instanciations des objets. Dans la LD, nous pouvons distinguer deux niveaux pour la représentation des connaissances :

- les *informations terminologiques* : définition des notions basiques ou dérivées et de comment elles sont reliées entre elles. Ces informations sont « génériques » ou « globales », vraies dans tous les modèles et pour tous les individus. Ce sous-langage décrivant les concepts est appelé TBox.
- les *informations sur les individus* : ces informations sont « spécifiques » ou « locales », vraies pour certains individus particuliers. Ce sous-langage de description des instances est appelé ABox.

Les langages basés sur la logique de description, comme **LOOM** [Bri93] par exemple, peuvent être utilisés pour représenter la connaissance terminologique d'un domaine d'application d'une façon structurée et formelle, apparaissant comme une extension des frames et des réseaux sémantiques, qui ne possédaient pas de sémantique formelle basée sur la logique.

3.3.2.5 Limites de ces langages pour nos besoins

Les outils conceptuels que nous souhaitons manipuler dans nos travaux doivent permettre de représenter l'ensemble des collections volumineuses traitées dans le domaine de la RI. En effet, dans notre cas, l'approche documentaire déclinée sur l'accès aux documents, l'aide à la décision et l'indexation doit fournir des documents contextuellement compréhensibles. Les formalismes que nous venons de présenter sont difficilement adaptables dans ce cas-là. Nous nous rapprochons alors des problématiques posées par le « Web sémantique » qui cherchent à améliorer l'accès aux documents. La question de leur récupération, donc de leur indexation dans une problématique de recherche d'information est une des clés de leur accès et un certain nombre de langages existe pour tenter de répondre à ces besoins.

Le Web constitue un terrain idéal d'application des ontologies considérées en tant que spécifications partagées de connaissances, les pages Web représentant une masse de connaissances immense et hétérogène. Les ontologies y sont généralement utilisées pour l'indexation, fournissant les index conceptuels décrivant les ressources sur le Web pour en faciliter l'accès. Après avoir présenté brièvement l'évolution du Web en Web sémantique, nous présenterons les principaux langages existants dans la mouvance du Web sémantique que sont XML, RDF et OWL. Une raison et non des moindres de ce choix est que ces langages sont ou ont été pour la plupart recommandés par le World

Wide Web Consortium (le W3C [40]). Les champs d'application éventuels liés à cette évolution sont vastes : raisonnement automatique, résolution de problèmes par inférences, représentation de données structurées, traduction automatisée, etc.

3.3.3 Les langages à balises définis autour du Web Sémantique

Le Web [BLCL+94] forme une source de données et d'informations interrogée par un grand nombre d'internautes. Cette source augmente sans cesse ainsi que le nombre d'utilisateurs, qui ont des profils et des objectifs variés. Dans ce contexte, le document joue un rôle central. D'une part, il constitue le support d'information le plus familier aux internautes et représente, d'autre part, le format standard d'échange de données sur le Web. A l'origine du Web, les documents sont liés les uns aux autres par les hyperliens, selon le langage de balise HTML [41], cependant ces hyperliens qui relient ces documents n'apportent pas de réponse sur la sémantique des relations. L'ajout d'une couche sémantique au dessus de la couche HTML, qui ne peut servir qu'à décrire formellement les pages Web, est donc nécessaire. Le Web sémantique vise à pallier le manque de sémantique des documents du Web traditionnel. Selon [BHL01] : « *Le Web sémantique est une extension du Web tel qu'on le connaît dans lequel on donne à l'information un sens bien défini, permettant ainsi aux ordinateurs et aux gens de mieux travailler en coopération* ». Le Web sémantique a donc comme objectif de transformer le Web en un système ayant un contenu qui pourrait être compris par un ordinateur. Son objectif principal est d'exprimer la signification.

Une première approche est d'enrichir chaque document par des balises (des métadonnées ou des annotations) qui vont lui donner du sens, et qui sont organisées au sein de référentiels partagés (ontologies). Un premier langage nommé XML [42] voit le jour pour représenter sémantiquement les données et non plus des éléments de présentation, comme en HTML. Cependant, XML ne permet pas de représenter des concepts mais des documents et ils ne disposent pas d'une sémantique pour réaliser des manipulations appliquées ou des raisonnements sur des documents. Ces limites ont orienté les travaux de recherche vers la définition de nouveaux langages visant à étendre XML, notamment RDF et OWL. L'idée sous-jacente à ces évolutions est de lier des concepts plutôt que des documents. L'objectif commun à ces langages est de participer à une formalisation des savoirs, en permettant ainsi un meilleur partage et une transmission plus aisée. [BHL01] présente l'architecture en couches du Web sémantique en s'appuyant sur une pyramide de langages pour représenter des connaissances sur le Web (voir figure 3.2).

Dans cette architecture, un langage de la couche haute doit être une extension du langage de la couche située en dessous. Dans cette architecture, XML permet de décrire les différents éléments faisant partie des documents. XML fournit la syntaxe mais n'offre aucun élément pour décrire la sémantique sur les objets qui le composent. RDF permet

40. Consortium pour la standardisation autour du Web Sémantique, le site http ://www.w3.org décrit les différents langages que le consortium préconise.

41. Hypertext Markup Language : langage constitué de balises (tags) pour décrire des éléments de présentation.

42. eXtensible Markup Language : Recommandation du W3C depuis le 10 février 1998.

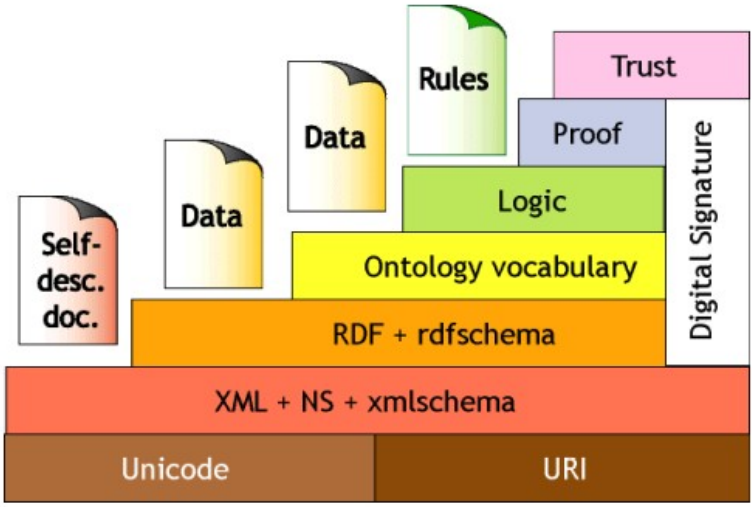


FIGURE 3.2 – Les couches du Web Sémantique

alors d'assigner un identifiant global à nos ressources permettant ainsi de référer et de modifier des affirmations faites dans d'autres documents. OWL correspond à la couche ontologie et permet de décrire une conceptualisation d'un domaine d'intérêt en particulier. La couche logique contient les règles qui permettront de faire de l'inférence à partir des concepts de l'ontologie. Enfin, la couche de preuve est conçue pour permettre à un agent logiciel intelligent d'expliquer une réponse qui a été générée à partir de la couche logique.

Nous allons maintenant présenter de façon succincte les langages que nous souhaitons utiliser pour définir une ontologie : SKOS, RDF et OWL. Nous en profitons pour présenter une solution alternative à OWL, en l'occurence XTM en indiquant ses atouts et ses limites.

3.3.3.1 SKOS

SKOS (Simple Knowledge Organisation System ou Système simple d'organisation des connaissances) [43] est un modèle de données permettant de gérer différents types de vocabulaires contrôlés, tels que les thésaurus bien sûr mais également les listes d'autorités, les schémas classificatoires, ou encore les taxonomies. Initié par l'Union européenne dans le cadre du projet SWAD-Europe (Semantic Web Advance Development for Europe), le langage SKOS a pour objectif de proposer un système permettant d'exprimer et de gérer des modèles interprétables par les machines dans la perspective du web sémantique. Ce modèle est défini comme « simple » par opposition à d'autres modèles, comme OWL, plus à même de représenter des structures sémantiques plus riches telles que les ontologies, mais de ce fait également plus complexes à utiliser. Le langage SKOS

43. http ://www.w3.org/TR/2009/REC-skos-reference-20090818/

permet :

- d'une part, la représentation dans un contexte multilingue de tout type de vocabulaire contrôlé et structuré (thésaurus, classifications, taxonomies, etc.),
- d'autre part, l'alignement de différents vocabulaires, l'objectif étant l'exploitation par la machine de ressources publiées sur le Web.

Le modèle de données SKOS a été conçu comme une ontologie OWL-Full en soi. Il est donc constitué de classes et de propriétés. SKOS n'est en aucun cas en lui même un langage formel de représentation des connaissances, mais l'expression dans un tel langage, du modèle de données adapté au traitement des vocabulaires contrôlés. SKOS n'a donc pas pour vocation l'expression d'axiomes et de faits, mais de concepts et de réseaux de liens conceptuels et sémantiques. Bien que limité pour proposer une représentation complète d'un domaine, SKOS offre une possibilité relativement simple pour représenter des vocabulaires contrôlés de type taxomonie et thésaurus sur laquelle il est possible de s'appuyer pour déterminer les classes, propriétés et individus d'une ontologie, en utilisant notamment comme point de départ les relations hiérarchiques et associatives pour créer des axiomes et des faits.

SKOS, initialement défini pour formaliser des thésaurus, est également une alternative intéressante pour formaliser de façon légère une représentation sémantique d'un domaine, ne nécessitant pas de calculer des inférences. Cependant, ce langage ne permet pas de qualifier précisément les relations de type « associés ».

Parmi les différents projets s'appuyant sur SKOS, le projet néerlandais STITCH [44] (*Semantic Interoperability To access Cultural Heritage* en anglais) intégré au sein du projet européen TELplus [45], propose une traduction du langage RAMEAU en SKOS [IB09]. Ce projet est mené en partenariat avec la BnF et le Centre national RAMEAU. Le répertoire Rameau sous SKOS est interrogeable avec un outil développé dans le cadre de ce projet [46]. De son côté, la Bibliothèque du Congrès des Etats-Unis a mis à disposition son langage d'indexation LSCH en SKOS et, grâce au projet MACS [47], elle a pu réaliser un alignement de vocabulaire avec les concepts RAMEAU, démontrant ainsi la puissance de cet outil novateur.

3.3.3.2 RDF et RDFS

RDF (Ressource Description Framework) [48] est un modèle de représentation sémantique des informations du Web utilisant la syntaxe XML. Ces représentations comportent des métadonnées sur les ressources du Web comme les auteurs de pages Web, leur date de création. Chaque ressource est pourvue d'un identifiant uniforme de ressource (Uniform Resource Identifier, URI). L'intérêt principal de RDF réside dans la définition d'un mécanisme permettant de décrire des données indépendamment de tout domaine et de toute spécificité.

44. http ://www.cs.vu.nl/STITCH/
45. http ://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/
46. http ://eculture.cs.vu.nl :48080/vocreptags/autocompleteplus.jsp
47. https ://macs.hoppie.nl/pub/
48. http ://www.w3.org/RDF/

La structure fondamentale de toute expression en RDF est une collection de triplets, chacun composé d'un sujet, un prédicat et un objet. Un ensemble de tels triplets est appelé un graphe RDF. Ceci peut être illustré par un diagramme composé de noeuds et d'arcs dirigés, dans lequel chaque triplet est représenté par un lien nœud-arc-nœud.

Considérons l'information suivante : « *l'auteur de http ://www.Dupond.fr/ est Jean Dupond* ».

On peut choisir de représenter cette information en triplet RDF par les chaines suivantes (figure 3.3).

« {auteur de http:// www.Dupond.fr, Jean Dupond, est} »,
« est(auteur de http:// www.Dupond.fr Jean Dupond) »,
« <auteur dehttp:// www.Dupond.fr ><est>< Jean Dupond > ».

FIGURE 3.3 – Exemple de triplet RDF

En RDF/XML (déclaré en Xml), cette information s'écrira comme indiqué figure 3.4.

```
<rdf:Description about="http://www.Dupond.fr/">
        <schema:auteur>Jean Dupond </schema:auteur>
</rdf:Description>
```

FIGURE 3.4 – Extrait d'un concept RDF

Le préfixe d'espace de nom « schema » correspond à un espace de nom spécifique, qui doit être indiqué par l'auteur du document RDF/XML dans une déclaration XML d'espace de nom. Un fichier RDF/XML ne contenant que l'information « l'auteur de http ://www.Dupond.fr/ est Jean Dupond » peut s'écrire de la façon présentée figure 3.5.

```
<?xml version="1.0"?>
<rdf:rdf xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:prefixe="http://source_de_schema/schema">
        <rdf:Description about="http://www.Dupond.fr/">
                <prefixe:auteur> Jean Dupond </prefixe:auteur>
        </rdf:Description>
</rdf:rdf>
```

FIGURE 3.5 – Exemple d'un fichier RDF/XML

Il est donc nécessaire, pour donner un sens aux informations stockées sous forme de triplets RDF, de se donner un vocabulaire, de définir la signification de la propriété « auteur », ainsi que son type, son champ de valeurs, etc. C'est le rôle de RDF Schema, qui permet de créer des vocabulaires de métadonnées.

On pourrait également employer un espace de nom par défaut.

Ainsi, la nécessité de déclarer l'espace de nom employé nous amène à la définition du Schéma RDF (RDFS). En effet, les schémas RDF (RDFS) permettent de définir le vocabulaire utilisé dans les descriptions RDF. Ils confèrent un formalisme de représentation

riche, incluant des classes, sous-classes, propriétés, sous-propriétés, des règles d'héritage de propriétés, etc., mais ne normalisent pas les inférences que l'on pourrait faire avec. La structure objet-classe des RDFS permet de représenter un modèle du domaine en définissant des objets du domaine et leurs relations pour rendre compte d'une ontologie.

Dans l'exemple précédent, la propriété « auteur » n'a de sens pour décrire la ressource « http ://www.Dupond.fr/ » que dans un contexte bien défini : le lecteur (utilisateur) est humain, ce lecteur comprend le français et l'information transmise par le triplet RDF « {http ://www.Dupond.fr/, Jean Dupond, auteur} » est suffisamment triviale pour comprendre que sa signification est « Jean Dupond est l'auteur de http ://www.Dupond.fr/ ».

3.3.3.3 OWL

OWL (OntologyWeb Language) [49], créé en 2001 par le W3C, héritier du langage DAML+OIL [50], doit permettre de représenter des ontologies sur le Web. OWL est destiné à être utilisé lorsque les informations contenues dans les documents doivent être traitées par des applications logicielles, c'est-à-dire lorsqu'elles ne sont pas simplement « montrées » à l'utilisateur. Une ontologie OWL est composée d'un en-tête (métadonnées), d'axiomes et de faits. Les axiomes concernent la définition complète ou partielle de concepts et de relations (ou priorités), la spécification de propriétés sur les relations (propriétés algébriques) et la définition d'axiomes sur les classes et les relations (équivalence, expression booléenne). Parmi les relations, on distingue celles dont le domaine de valeur sera de type primitif (attribut) de celles dont le domaine de valeur sera un autre concept (relation). Les faits concernent des individus pour lesquels on donne des valeurs aux propriétés des classes dont ils sont les instances.

OWL fournit trois sous-langages, d'expressivité croissante, nommés OWL Lite, OWL DL et OWL Full :

- Le langage OWL Lite peut être vu comme une extension du langage RDFS. Son principal intérêt est de permettre la modélisation d'ontologies simples, d'une complexité formelle peu élevée, de sorte qu'il soit facile d'implémenter des raisonneurs corrects et complets.
- Le langage OWL DL contient des constructeurs supplémentaires, mais il ne peut être utilisé qu'avec certaines restrictions. Par exemple, une classe ne peut pas être une instance d'une autre classe. Il en résulte un langage un peu plus expressif et toujours décidable.
- Le langage OWL Full dispose des mêmes constructeurs que OWL DL mais il les interprète de manière plus large. Ainsi, une classe peut cette fois être vue comme un ensemble d'individus (définition extensionnelle) ou comme un individu à lui tout seul (définition intensionnelle) qui pourra, par exemple, donner une valeur à une propriété. Toutefois, le langage OWL Full n'est plus décidable.

« *OWL étend RDF-Schema (ou RDFS) pour permettre l'expression de relations complexes entre différentes classes RDFS, ainsi que l'expression de contraintes plus précises*

49. http ://www.w3.org/TR/owl-features/
50. http ://www.daml.org/about.html

sur des classes et des propriétés spécifiques »[51]. Nous pouvons notamment noter :«

- *la manière de limiter les propriétés de la classe au regard du nombre et du type* ;
- *les moyens d'induire que ces termes assortis de propriétés diverses soient des membres d'une classe particulière ou non* ;
- *les moyens de déterminer si tous les membres d'une classe auront une propriété particulière, ou seulement quelques uns d'entre eux* ;
- *les moyens de séparer des relations de types un-à-un de relations de type plusieurs-à-un ou un-à-plusieurs, permettant ainsi de représenter des « clés étrangères » d'une base de données dans une ontologie* ;
- *les moyens d'exprimer des relations entre des classes définies dans différent documents sur le Web* ;
- *les moyens de construire de nouvelles classes en dehors de toutes unions, intersections et compléments avec d'autres classes* ;
- *les moyens de contraindre un domaine à des combinaisons classe/propriété spécifiques* ».

OWL est actuellement le langage le plus utilisé en ingénierie ontologique pour la représentation sémantique d'ontologies. C'est un standard du W3C. C'est-à-dire qu'à défaut d'être une norme ISO, il en est l'équivalent dans le monde des industriels. La déclinaison d'OWL en trois langages offre une large gamme d'outils nous permettant d'envisager à terme de proposer une représentation sémantique riche du territoire implicitement décrit par le fonds documentaire traité. Il nous intéresse dans notre démarche car c'est en même temps un formalisme efficace pour proposer une première représentation plus légère du domaine étudié.

L'évolution du Web a cependant permis à d'autres formalismes d'être créés, parfois plus simples à définir et à utiliser, qui permettent de proposer une représentation sémantique légère d'un domaine. Nous pouvons notamment citer les formalismes permettant de représenter des « cartes de concepts ». Nous présentons dans ce cadre le formalisme XML Topics Map [Top01], capable de représenter une ontologie en différenciant aisément les niveaux conceptuels et physiques (les documents eux-mêmes).

3.3.3.4 XTM

TopicMaps.Org est un consortium indépendant développant un modèle applicatif des **Topic Maps** (TM) (ISO13250[52]) destiné au World Wide Web, cela en utilisant la spécification XML et les technologies associées. Le standard spécifie également le XML Topic Map (XTM) [Top01]. Le langage peut être vu comme la sérialisation des Topics Maps en XML, servant à l'échange et à l'utilisation de topic maps par des agents logiciels.

Notre volonté étant de représenter les connaissances de fonds documentaire sous forme d'une ontologie caractérisant au mieux le territoire, nous ne prenons dans nos choix que les langages ontologiques. OWL apparaît comme le successeur de RDF/RDFS en ajoutant un niveau de sémantique permettant de détailler la description des éléments

51. http ://www.w3.org/2003/08/owlfaq.html.fr
52. http ://www.isotopicmaps.org/rm4tm/

qui les constituent. Cela dit, le langage XTM qui se positionne au même niveau conceptuel que le langage OWL, se rapproche bien plus de la conception, au niveau sémantique, que l'on se fait de l'ontologie que l'on souhaite mettre en place, en proposant notamment un langage de description proche de notre mémoire. De plus, contrairement aux autres langages ontologiques, XTM offre la possibilité de définir des relations autres que binaires. Il est possible dans les TM de définir des relations ternaires particulières entre topics associant les deux premiers à un contexte particulier grâce à l'utilisation du terme « Scope ». Notons tout de même que malgré la difficulté, cela reste faisable en OWL mais de manière bien moins déductible. [PH02] met en avant les éléments déterminants suivants :

- Il est possible d'envisager les topics comme des concepts ou des instances de concepts ;
- Il est possible d'envisager les associations, les contextes (scope) et les occurrences comme des relations entre concepts ;
- Les associations n'ont pas de limitation dans le nombre de leurs membres ;
- La relation occurrence permet d'attacher des ressources directement à un concept (une même ressource peut bien sûr apparaître dans plusieurs relations d'occurrence et être accessible à partir de plusieurs concepts) ;
- Les relations (associations, occurrence et les libellés des concepts) peuvent être définies à l'intérieur d'un contexte. Ceci permet d'implémenter simplement les notions d'annotation ou de points de vue dans la mémoire.

En comparant les TM aux autres formalismes du Web sémantique,nous pouvons différencier son rôle de celui de RDF. En effet, tandis que RDF est un modèle de représentation de métadonnées pour l'annotation de ressources Web, Topic Map est un modèle d'organisation de connaissances pour pouvoir naviguer dans les ressources Web. [Ouz06] montre que les Topic Maps sont d'un grand apport pour le développement du Web sémantique. Elles permettent de représenter des connaissances pour faciliter la navigation dans l'univers de l'information modélisée. Les notions de rôle dans une association et de contexte (scope) jouent un rôle très important notamment dans le contexte des données partagées.

Les Topic Maps ne peuvent suffire à elles seules pour la réalisation du Web sémantique. Nous avons vu qu'elles manquent de sémantique et de mécanismes d'inférence logique. De plus, le formalisme de représentation utilisé pour maintenir une ontologie doit assurer sa consistance. Cette fonctionnalité n'est pas considérée par les Topics Maps. Elles permettent tout type d'inconsistance pouvant être spécifié par les auteurs. C'est la raison pour laquelle les Topic Maps ont été très critiquées par la communauté des logiques formelles. Pour remédier à cet inconvénient, la communauté des Topics Maps a proposé un langage d'expression de contraintes appelé TMCL [53] (Topic Maps Constraint Language) qui ne reste à l'heure actuelle qu'une spécification de l'ISO.

La place de XML Topic Maps dans l'architecture en couches du Web sémantique est celle occupée par RDF. Certes, les Topics Maps sont plus souples et plus expressifs pour modéliser les connaissances relatives aux ressources du Web, mais RDF(S) est

53. http ://www.isotopicmaps.org/tmcl/

actuellement très utilisé et sert de base à la définition des langages SKOS et OWL.

3.3.3.5 Editeurs d'ontologie

L'ingénierie des ontologies est un sous domaine de l'ingénierie des connaissances en plein essor et la multitude des outils permettant de manipuler les ontologies en est un bon exemple. Nous pouvons citer Ontolingua [FFR97], WebOnto [Dom98], HOZO [MKSK00], Protégé 2000 [NFM00], Oiled [BHGS01], KAON (intégrant l'outil OntoEdit [Mäd02, SAS02]), WebODE [ACFLGP03], Swoop [KPH05], NeOn Toolkit [54], et parmi les outils industriels Transinsight [55], le plug-in eclipse TopBraid Composer [56], Synaptica [57], Be Informed Suite [58], etc. Dans la plupart des cas, ces outils permettent de créer puis d'éditer une ontologie et acceptent différents formalismes tels que RDF, OWL, etc. [GPCFL04] fait un inventaire intéressant des outils permettant de traiter les ontologies. Ils définissent dans cette étude différents critères permettant d'évaluer les outils à disposition :

- La description générale des outils : informations générales par rapport à l'outil et ses développeurs, les versions, la disponibilité de l'outil, etc. ;
- L'architecture logicielle et l'évolution de l'outil : Standalone, 3-Tiers ou n-tiers. Il faut indiquer ici si l'application donne la possibilité d'intégrer de nouvelles fonctionnalités ou des plug-ins, et s'il est possible de faire des sauvegardes pour un possible retour en arrière « backup ». Enfin, ce critère met en avant la façon de stocker l'ontologie (système de fichiers ou base de données) ;
- Manipulation de l'ontologie : Il est nécessaire de précisr ici les fonctionnalités d'édition et de navigation de l'ontologie fournies par l'outil, avec les librairies utilisables, les services d'inférence, les options de documentations, les possibilités de constructions collaboratives, etc. ;
- La représentation des connaissances : évaluation du modèle de connaissance de l'outil pour identifier quelle connaissance peut être modélisée par l'outil et comment ;
- L'interopérabilité : L'outil proposé est-il en mesure d'échanger avec d'autres outils de développement d'ontologie et des langages d'ontologies : les formats d'importation et d'exportation, langages supportés. L'étude analyse notamment comment intégrer les ontologies développées avec d'autres éditeurs dans différents Systèmes d'Informations.

3.4 Discussion

L'ontologie dans le processus de création d'ontologies est l'aboutissement d'un travail important respectant un ensemble de critères qui font maintenant référence en ingénierie ontologique. Un nombre important de travaux propose une méthodologie pour construire

54. http ://neon-toolkit.org/wiki/Main_Page
55. http ://www.transinsight.com/products
56. http ://www.topquadrant.com/
57. http ://www.synaptica.com/Overview.asp
58. http ://www.beinformed.nl/BeInformed/website/nl ?init=true

une ontologie. Si un cadre commun semble apparaître entre les différentes méthodes recensées, il n'existe pas à ce jour de méthode unique pour créer une ontologie permettant de répondre aux besoins de tout type d'utilisateur dans les différents domaines.

Pour notre part, nous souhaitons proposer un processus automatisé de construction d'ontologies dont le résultat sera analysé et validé ensuite par les experts bibliothécaires à travers des outils permettant de naviguer dans des fonds documentaires qu'ils auront indexés. Nous nous intéressons plus particulièrement aux méthodologies permettant de créer une ontologie de façon automatique, intégrant néanmoins en fin de traitement une étape de validation par des experts. Ces travaux émergent comme un sous-domaine de l'ingénierie des ontologies. Un moyen très largement utilisé pour atteindre cet objectif est de partir d'éléments préexistants dans le domaine. [MS01] distinguent différents types d'approches en fonction du support sur lequel elles se basent : ressources structurées de type vocabulaire contrôlé (taxonomies, thésaurus, etc.), normes ou fragments préexistants d'ontologies, ou encore des corpus textuels. Certains travaux reposent sur l'analyse de textes tels que ARCHONTE [Bac00], TERMINAE [BS99,AGBS00,SBAG02, BAGC04], KAON [VOS03], Text2Onto [CV05] afin d'aider à la construction automatique ou semi-automatique des ontologies.

De notre côté, nous cherchons à construire une ontologie d'un domaine cible à partir de ressources textuelles peu ou pas structurées et d'un vocabulaire contrôlé de type thésaurus. Nous nous intéressons plus particulièrement aux méthodes permettant de transformer un vocabulaire contrôlé de type thésaurus en ontologie du domaine [WSWS01, SLL+04, Her05, CHGM06]. Nous nous rapprochons plus précisément des travaux de [CHGM06] qui, sur la base de la méthodologie TERMINAE, propose de transformer un thésaurus du domaine cible en une ontologie. Le thésaurus utilisé nous permet d'identifier un ensemble de concepts et de relations entre ces concepts.

Afin de construire une ontologie d'un domaine cible la plus précise possible, nous souhaitons, dans l'étape d'analyse linguistique, nous appuyer sur un ensemble de documents textuels annotés afin d'identifier de nouveaux concepts et de nouvelles relations caractérisant le domaine visé. Nous nous appuyons notamment sur les travaux de [LGN07] qui propose de combiner des analyses syntaxiques et des calculs de dépendance pour identifier les relations argumentatives *(sujet, verbe, objet)* [JKT97, BL02], à une approche intégrant des patrons lexico-syntaxiques pour reconnaître les marques linguistiques des relations sémantiques [AGBS00, AB08].

Concernant la formalisation de nos résultats, nous faisons le choix d'utiliser le langage SKOS pour les étapes intérmédiaires dans lesquelles nous manipulons un thésaurus. Ce choix vient du fait que SKOS permet d'exprimer toutes les composantes d'un thésaurus. Et, dans le cadre du projet néerlandais STITCH mené en partenariat avec la BnF, des travaux récents ont donné lieu à une traduction du thésaurus RAMEAU (que nous utilisons dans nos expérimentations) en SKOS [IB09]. Pour formaliser l'ontologie résultante, nous utilisons le langage OWL-Lite. Il apporte les compléments nécessaires pour expliciter des éléments caractéristiques d'un domaine cible. De plus, OWL est très utilisé en ingénierie ontologique pour la représentation sémantique d'ontologies et il apparaït comme un format d'échange intéressant, que ce soit en interne dans notre laboratoire ou

avec des partenaires. Pour visualiser et analyser les résultats que nous obtenons, nous utilisons l'éditeur Protégé [NFM00] qui donne la possibilité d'ouvrir et d'éditer différents types de fichiers (OWL, RDF, SKOS, etc.). Très modulaire, il donne la possibilité d'ajouter des plugins et nous utilisons notamment OntoGraph qui facilite la lecture de la structure sémantique que nous produisons.

Dans une démarche de description de fonds documentaires dans les centres documentaires de types bibliothèques et médiathèques, nous proposons de construire, sur la base du travail d'annotations de documents réalisé par des experts, une couche conceptuelle de type ontologie d'un domaine cible. Cette structure sémantique, synthétisant le travail d'annotation, offre aux experts bibliothécaires un support d'aide à l'indexation. De plus, l'ontologie obtenue offre une représentation sémantique facilitant la découverte d'un domaine cible. Une caractéristique importante de ces fonds documentaires est qu'ils contiennent d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire. Il est primordial pour ces centres de valoriser ces spécificités territoriales pour répondre à des objectifs d'information et d'éducation. Nous allons maintenant présenter le domaine d'application dans lequel nous positionnons nos travaux, en l'occurence le **Territoire**.

Chapitre 4

Définition du domaine cible : le territoire

Sommaire

4.1 Introduction 69
4.2 Le territoire 70
4.2.1 Le territoire des géographes 70
4.2.2 Le territoire comme outil de la géographie sociale 71
4.2.3 L'architecture du territoire 72
4.2.4 Les composantes d'un territoire 73
4.2.5 Implications dans nos travaux 74
4.2.6 Le territoire en géomatique 75
4.3 L'information géographique dans les textes 76
4.3.1 Définitions et modélisation 76
4.3.2 Le TALN pour l'extraction d'informations géographiques . 77
4.4 Construction d'ontologies géographiques 80
4.4.1 Les différents types d'ontologies géographiques 80
4.4.2 Le « cas » Territoire 81
4.5 Discussion 82

4.1 Introduction

La notion de **territoire** intéresse de nombreux chercheurs dans différents domaines ayant des actions liées à l'implantation de l'Homme tels que la géographie, l'économie, la médecine, la géologie, la sociologie, le droit, etc. Dans ce chapitre, nous proposons une synthèse succincte des travaux visant à définir cette notion complexe qu'est le territoire et nous tentons ensuite de tirer parti de ces définitions dans notre domaine pour identifier les approches permettant d'identifier, d'extraire et de structurer un ensemble d'informations caractéristiques d'un territoire à partir de documents textes.

L'encyclopédie Universalis définit le mot territoire de façon générale (valable pour l'ensemble de ces domaines) comme l'« *étendue de la Terre sur laquelle vit un groupe humain* ». Dans nos travaux, nous cherchons à identifier les éléments communs aux différentes définitions proposées dans l'ensemble des domaines pour proposer une vue territorialisée du fonds documentaire annoté. Pour ce faire, nous nous appuyons plus particulièrement sur les travaux des géographes pour qui le territoire est un concept faisant partie intégrante de leur problématique scientifique. Aussi, les travaux des géographes ont à minima un point de vue ouvert vers les autres disciplines et proposant une vue synthétisant celles de toutes les disciplines. Une définition conventionnelle en géographie au niveau international existe, décrivant le territoire comme un espace sur lequel s'exerce une autorité limitée par des frontières politiques et administratives. Cependant, sa définition est encore sujet à discussion comme le montrent ces différents travaux [Pio92, Sch94, LL03, GV04, DMB05, Gui07] et elle évolue en fonction des domaines d'activités.

Notre objectif est d'identifier un ensemble d'éléments qui vont nous permettre de définir un modèle suffisamment formel d'un territoire pour être utilisé dans nos travaux. Pour cela, nous présenterons dans une première section des travaux proposant de définir la notion de territoire en géographie, définitions qui mettent en avant l'importance de la composante spatiale ainsi que les relations avec les composantes thématiques et temporelles. Dans une deuxième section, nous décrivons les travaux qui visent à spécifier ces définitions dans le domaine de la géomatique. La géomatique est un domaine scientifique à cheval entre la géographie et l'informatique qui tente d'apporter des éléments de réponses aux problèmes de stockage, de traitement et de diffusion des informations géographiques générées, toujours plus nombreuses. Nous exposons ensuite dans cette section les différents travaux visant à extraire des documents les entités qui décrivent un territoire. Nous verrons que ces entités sont forcément caractérisées par une composante spatiale.

Nous souhaitons structurer l'ensemble de ces informations, qui caractérisent un territoire, via une couche conceptuelle de type ontologique afin d'obtenir une représentation structurée d'un territoire. Pour cela, nous faisons ensuite un état des travaux en géomatique qui cherchent à modéliser un ensemble d'informations géographiques extraites de textes sous forme d'une ontologie. Il est à noter que nous avons restreint nos recherches dans ce chapitre à la manipulation de données textuelles non-structurées. Nous faisons enfin la synthèse de ce chapitre en soulignant les travaux les plus pertinents dans le cadre de notre problématique.

4.2 Le territoire

4.2.1 Le territoire des géographes

Jusqu'aux années 80, le terme territoire a une définition administrative ou politique et il est vu comme un espace limité par des frontières. Cette définition, conventionnelle, semble plutôt être influencée par la géographie anglo-saxonne qui s'est concentrée sur le

concept de lieu (place).

X. Piolle montre que le territoire s'impose comme un concept central dans la géographie française à partir des années 1980 [Pio92]. J. Levy et M. Lussault appuient ces propos en précisant : « dans la production francophone, on peut en repérer [le territoire] l'entrée officielle avec l'édition de 1982 des rencontres Géopoint, « *Les territoires de la vie quotidienne* ». » [LL03]. Dès lors, diverses définitions sont proposées. Le territoire n'est plus seulement un espace délimité par des frontières politiques et administratives. Il devient un système régi selon un espace, mettant en relation une multitude d'agents et d'objets matériels et immatériels [Bes06].

Malgré ce succès dans le domaine de la géographie et plus généralement dans le langage courant, le mot territoire soulève encore bien des interrogations. Il est souvent synonyme de lieu, d'espace, d'espace socialisé, d'espace géographique, de territoire éthologique ou d'espace approprié. Nous verrons que définir le concept de territoire n'est pas simple et que sa définition diverge en fonction des domaines d'activités.

[Gui07] exprime simplement la notion de **territoire** comme la superposition d'un espace et de pratiques sociales. Cette définition se rapproche de celle donnée dans le dictionnaire de géographie [GV04] : un espace géographique qualifié par une appartenance juridique (on parle ainsi de « *territoire national* ») ; ou par une spécificité naturelle ou culturelle : territoire montagneux, territoire linguistique. La deuxième définition précise les caractéristiques sociales du territoire et les divise en deux catégories que sont le politique et le culturel. Dans ce dernier cas, le terme d'aire (« *aire linguistique* ») pourrait lui être préféré. Quelle que soit sa nature, un territoire implique l'existence de frontières ou de limites. Ces deux derniers termes sont utilisés en fonction du type de territoire dont ils forment le périmètre. Un territoire politique, ou *subdivision administrative*, est délimité par une frontière ; un territoire naturel est circonscrit par une limite, terme moins juridique. Dans [BFT94], le territoire est conçu comme un concept défini par l'Homme et pour l'Homme. Sans l'imaginaire humain qui lui confère tout son sens, il n'existe point de territoire. Il permet ainsi de faire le lien entre des pratiques sociales humaines et un espace, ce qui empêche une représentation unique, objective et impartiale. Comme l'affirme [Gui08], les représentations territoriales, pouvant être définies individuellement ou collectivement, sont accordées à des pratiques sociales, qui peuvent être réelles, rapportées ou supposées. On parle alors d'images ou de vues d'un territoire qui peuvent être plus ou moins proches de la réalité en fonction des besoins et des informations que l'on a à disposition.

4.2.2 Le territoire comme outil de la géographie sociale

Il constitue le cadre méthodologique permettant d'évaluer la nature des rapports sociaux dans leur contexte de spatialisation. Le territoire suppose une appropriation de l'espace. Cet aspect renvoie aux origines éthologiques du concept qui tendent à borner le territoire de façon stricte et poussent à le défendre contre toute agression extérieure.

A travers sa dimension politique, le territoire constitue un outil de gouvernement, un mode de découpage et de contrôle de l'espace fondé sur une attitude volontariste et intentionnelle. Les dimensions juridico-administratives associées assurent des formes de

régulation, implicites ou explicites, imposées ou consenties. Cependant, une des limites majeures de la vision politique du territoire réside dans sa rigidité qui incite souvent à ne raisonner qu'en termes de frontières, de dedans et de dehors.

[Sch94], cité plus tard par [Eli02], montre bien que cette vision est trop réductrice et définit alors le territoire comme une « *notion concrète qui renvoie à une terre et non à un espace géométrique. [...] Le territoire a une localisation, une dimension, une forme, des caractéristiques physiques, des propriétés, des contraintes et des aptitudes. [...] Il y a un processus historique unique de formation d'une société et de son territoire. Le fonctionnement territorial d'une société ne peut être appréhendé hors de son rapport à sa propre histoire.* ». Nous retenons dans cette définition qu'un territoire est entre autres décrit par :

- un espace : définie ici par *une localisation et une dimension* ;
- la présence/intervention de l'homme : définie ici par *la formation d'une société* et *le fonctionnement territorial d'une société* ;
- dans une période : définie ici par un *processus historique unique*, et une description d'une société selon son *histoire*.

Nous allons montrer maintenant que ces spécificités permettent de définir le territoire en nous arrêtant plus précisément sur ces notions de localisation et de lien entre la formation d'une société et de son territoire dans le temps.

4.2.3 L'architecture du territoire

[DMB05] montre que parler du territoire au sens de la géographie sociale revient à affirmer, par hypothèse, que quelle que soit la mobilité des individus, quelle que soit la singularité de leur territorialité, il existe toujours entre eux une connivence, un accord implicite intervenant à un niveau d'échelle particulier de l'espace géographique. Cet accord porte en particulier sur l'identification commune de lieux. Pour que ces lieux, associés ou non, deviennent territoire, il est nécessaire qu'agents et acteurs les signifient conjointement. Dans notre cas, l'agent peut être assimilé aux documents contenant un ensemble de lieux, et l'acteur est le documentaliste qui réalise le travail d'annotation de ces documents sous forme de notices descriptives. Dans la mesure du possible, les notices descriptives contiennent une ou plusieurs références à des lieux qui relatent de façon globale l'espace géographique décrit dans le document. C'est le nombre, la fréquence de tels accords, l'intensité de leur conviction partagée qui déterminent la solidité ou la fragilité, la lisibilité plus ou moins nette et la stabilité plus ou moins affirmée d'une construction territoriale.

Une grande lisibilité se traduit théoriquement par l'instauration de frontières délimitant clairement le dedans et le dehors, l'intérieur et l'extérieur, ce qui relève du territoire et de ce qui lui échappe. Le territoire apparaît alors comme un espace vécu dans le temps, doté d'une cohésion sociétaire ancrée dans un espace géographique doté de ressources (matières premières, actifs, relations). Il permet ainsi de mettre en évidence les lieux selon que leur représentation correspond à des points (lieux-dits, communes, etc.), à des lignes (rivières, routes, etc.) ou à des surfaces (régions, etc.).

4.2.4 Les composantes d'un territoire

Nous avons vu que le territoire peut être vu comme un espace vécu dans le temps, doté d'une cohésion sociétale ancré sur un espace géographique composé de lieux. Dans l'espace géographique, les lieux se distinguent beaucoup plus aisément que les territoires. En général, leurs limites se perçoivent sans beaucoup d'ambiguïté. On entre et l'on sort d'un lieu pour une raison bien précise [DMB05]. Les lieux sont des espaces ou des édifices bien circonscrits définis par la contiguïté des points qui le composent. Ils abolissent la distance pour remplir une fonction de proximité. Leur réalité sensible et palpable surgit de leur clôture. Alors que le territoire se laisse difficilement borner dans nos représentations, le lieu tire de sa fermeture le plus clair de son identité.

Nous allons maintenant définir ces différents composants permettant de définir une représentation d'un territoire que sont le lieu, le sujet (ou thème) et le temps.

4.2.4.1 Le concept de lieu

Le mot **lieu** en français vient du latin locus, qui sert à traduire le mot grec topos et signifie place, endroit.

Le lieu n'est pas une notion récente en géographie ; son élaboration est liée à l'histoire de l'étude des relations de l'homme à l'espace. Longtemps abordée sous l'angle de la géographie régionale, la notion de lieu, à partir des années 1960, se transforme sous l'influence de l'approche analytique quantitative et spatiale et ne se réduit pas à la désignation d'une simple localisation.

Le territoire regroupe et associe les lieux [DMB05]. Il leur confère un sens collectif plus clair et plus affirmé que celui qui découle de leur stricte pratique. Dans ces conditions, territorialiser un espace consiste, pour des groupes sociaux, à multiplier les lieux. Le territoire souvent abstrait, idéal, vécu et ressenti plus que visuellement repéré et circonscrit, englobe des lieux qui se singularisent par leur valeur d'usage, par leur saisissante réalité. Dans ce contexte, la notion d'échelle spatiale est importante et fait référence à deux mesures fondamentales que sont l'étendue et la résolution [TAAM+10]. J. Charre définit l'étendue comme la taille de l'espace étudié (surface et contours de ce qui est représenté) et la résolution correspond quant à elle à la densité de l'information [Cha10]. [ODWA86] montre que la notion d'échelle fait également référence aux niveaux d'organisation sociale.

Nous remarquons ici que la notion de lieu est centrale dans la définition d'un territoire. Ainsi, un ensemble de lieux renseigne à la fois sur un espace (ou étendue) et une résolution (plus le nombre de lieux est important, plus la représentation de l'espace étudié sera précise).

Une vision plus vaste de ce concept de lieu est exposée notamment par [Ent96] qui indique que « *le concept lieu ne doit pas être interprété dans le sens de localisation, mais plutôt, dans un sens plus large, comme un ensemble d'éléments naturels, sociaux et culturels formant un tout sous l'action du sujet ou du moi* ». La notion de lieu fait émerger celle du sujet en tant que principal élément signifiant de l'action socio-spatiale.

C'est la perception même du lieu par le sujet qui lui donne sens. On ne parle plus alors du lieu comme un point ou une aire mais plutôt d'un ensemble de sentiments et perceptions. V. Berdoulay [BE98] affirme que « *le sujet et le lieu fonctionnent comme deux primitives de l'expérience humaine* ».

Le sujet, ou thème, apparaît ici comme une caractéristique importante du territoire car il permet de relier des lieux entre eux. Nous allons donc nous attarder brièvement sur cette notion de sujet et sur la relation qu'il a avec la notion de lieu.

4.2.4.2 Le lieu et le sujet

Le lieu et le sujet apparaissent alors comme les supports de l'expérience humaine du fait qu'ils sont étroitement imbriqués [DMB05]. Ceci est d'autant plus marqué quand l'identité territoriale est forte. Le sentiment d'appartenance est tel qu'on note une fusion entre le lieu et le sujet en tant que conscience de soi. Différents exemples, comme le nationalisme basque, breton ou encore corse, qui intensifie l'identité du groupe et le lien au lieu, peuvent être cités. Le sujet apparaît donc comme une composante qu'il est important de prendre en compte lorsque l'on souhaite représenter un territoire. Cependant, cette composante est complexe à identifier car elle couvre un champ très vaste.

Une troisième composante, le temps, prend une place importante lors de la définition du territoire. Comme le montre [Sch94] (*cf.* section 4.2.2 page 72), « *il y a un processus historique unique de formation d'une société et de son territoire* », indiquant l'importance du rôle du temps dans la constitution d'un territoire.

4.2.4.3 Le lieu et le temps

Fréquemment citée, la phrase de M. Marie [Mar82] illustre l'importance du temps long dans la construction symbolique du territoire : « *l'espace a besoin de l'épaisseur du temps, de répétitions silencieuses, de maturations lentes, du travail de l'imaginaire social et de la norme pour exister comme territoire. La mémoire joue un rôle majeur dans la construction territoriale, se révélant source de lien social et de cohésion. Le territoire devient alors une marque temporelle de la conscience d'être ensemble* ».

Il apparaît alors que les lieux peuvent se définir par rapport à nos expériences qui s'inscrivent dans le temps et l'espace. Le fait d'intégrer l'expérience à la signification d'un lieu entraîne l'incorporation du temps et de la durée à l'espace.

4.2.5 Implications dans nos travaux

D'après ce que nous avons vu précédemment, le territoire permet de représenter le contact vécu de l'Homme avec le milieu. Il associe trois éléments essentiels : d'abord une relation primaire et existentielle à la terre, ensuite le réseau des lieux pratiqués et vécus et enfin des référentiels représentés à des échelles multiples. De façon générale, ces trois niveaux sont présents dans les documents (images de scènes de foule, d'évènements,

de paysages, textes de récits de voyage, hebdomadaires, etc.). Pour notre part, nous cherchons à identifier et extraire des éléments caractérisants cette relation à la terre ainsi que le réseau de lieux défini implicitement lors de la constitution de fonds de documents. Nous avons vu que le territoire permet d'appréhender ce phénomène qu'est le contact vécu de l'homme avec le milieu. Dans ce cadre, le contact vécu de l'Homme fait référence à la composante sujet ou thématique et également à la composante temporelle. [Ent96, BE98] mettent en avant la relation entre les composantes sujet et temporelle avec la composante spatiale ; la composante spatiale étant vue alors comme un ensemble de lieux formant un espace défini selon la composante sujet. En somme, un territoire peut être définis comme un ensemble de lieux que l'on peut mettre en relation selon la composante sujet en fonction d'une periode donnée.

Le territoire peut alors être défini comme un concept offrant une vue spatialisée.

4.2.6 Le territoire en géomatique

Dans cette section, nous présentons brièvement dans un premier temps la géomatique en faisant un état des groupes de recherche travaillant dans le domaine. Nous avons vu que la notion de territoire en géographie (*cf.* section 4.2 page 70) correspond à un bâti social défini par l'humain dans un espace géographique donné, et nous verrons qu'il n'existe que peu de travaux qui cherchent à le représenter comme tel [LMP01]. De nombreux travaux s'appliquent à traiter la notion d'information géographique en vue de proposer des représentations administratives ou politiques d'espaces géographiques. Or, nous avons vu que le concept de territoire est constitué à minima par des éléments se rapportant à un espace, d'autres à un ou plusieurs sujet(s) (ou thème(s)) et un troisième groupe d'éléments à une periode. Dans un second temps, nous exposons un ensemble de travaux permettant d'identifier et d'extraire ce type d'informations dans des documents textes et nous faisons ensuite un état des travaux proposant de structurer cette connaissance en une ontologie géographique. Nous verrons qu'il existe peu de travaux à notre connaissance qui cherchent à extraire de textes des informations afin d'en construire une vue territorialisée sous forme d'une ontologie.

[Ber93] définit la **géomatique** comme l'ensemble des outils et des méthodes permettant de représenter, d'analyser et d'intégrer des données géographiques. [Jol04] complète cette définition en indiquant que *la géomatique est une discipline ayant pour objet la gestion des données à référence spatiale par l'intégration au moyen de l'informatique des savoirs et des technologies reliées à leur acquisition, leur stockage, leur traitement et leur diffusion, et principalement : la topométrie, la cartographie, la géodésie, la photogrammétrie et la télédétection.* [Jol04], p. 432.

La géomatique fait appel à des techniques informatiques spécifiques, notamment les Systèmes d'Informations Géographiques (SIG) pour l'acquisition des données, leur stockage, leur traitement et leur propagation. Depuis plusieurs années, un certain nombre de pays ont mis en place des structures de recherche pour fédérer les travaux dans le

domaine de l'information géographique : NCGIA [59] aux Etats-Unis, RRL [60] au Royaume Uni, Nexpri aux Pays-Bas, Réseau Géoïde au Canada, le CRC-SI [61] en Australie, etc. En France, le Groupement De Recherche Magis [62], dont le but est de valoriser la recherche française dans le domaine de la géomatique en France et à l'étranger (conférences SDH, SAGEO, etc.), en est un bon exemple. Ce GDR a également pour mission de favoriser la synergie de la recherche dans le domaine de l'Information Géographique en permettant à des équipes d'informaticiens et de géographes de collaborer. En géomatique, l'objet d'étude central est l'information géographique dans toutes ses formes (données brutes de capteurs automatiques, bases de données, cartes, Systèmes d'Informations Géographiques, enquêtes sur le terrain, commentaires textuels, etc.). L'intégration ou l'observation de cette information dans un territoire particulier n'est toutefois pas toujours explicite.

Nous nous appuyons sur le présupposé communément admis en géomatique que quelque soit le type d'information géographique traitée, elle apparaît sous forme d'Entités Géographiques (EGs), chacune étant composée d'une entité thématique (ou phénomène), d'une entité spatiale(ES) et d'une entité temporelle (ET) pouvant être implicite. Précisons que dans le cadre de nos travaux, nous nous concentrons sur l'analyse automatisée de l'information géographique dans des documents textes (notices descriptives attachées aux documents de type image, son vidéo et texte et documents textes complets).

4.3 L'information géographique dans les textes

4.3.1 Définitions et modélisation

Nous avons vu que l'expression de l'information géographique est abordée en abondance dans différents domaines et les travaux en géomatique, travaillant sur le contenu de documents textes, s'appuient sur les avancées des linguistes. Parmi ces derniers, divers travaux [Van86, Bor98, Tal00] centrent leur approche sur l'identification de relations spatiales exprimées par des marqueurs (prépositions, verbes, etc.) mettant en évidence un lien entre une entité à localiser et une entité de référence. D'autres travaux tels que [Par70, Sto02, AHV05] s'intéressent plus particulièrement à une catégorisation sous-jacente de l'information géographique en prenant comme point d'ancrage la composante spatiale. Nous retenons pour nos travaux la définition de A. Borillo indiquant que lorsqu'il est invoqué dans un texte, un lieu est une portion de l'espace matériel dans lequel nous nous situons et nous évoluons et dans le cas précis de lieux géographiques, il peut être rattaché à une catégorie (montagnes, lacs, etc.). A partie des travaux de [Van86] qui définit le concept cible/site, [Bor98] met en avant le fait qu'une référence à un lieu dans un texte correspond à une relation entre une entité concrète (l'objet cible décrit dans le texte) et une localisation (localisation de l'objet cible).

59. National Center for Geographical Information and Analysis
60. Research Regional laboratories
61. Cooperative Research Centre for Spatial Information
62. Méthodes et Applications pour la Géomatique et l'Information Spatiale

L'ensemble des travaux cités ici met en avant l'importance de la composante spatiale pour identifier et définir une information géographique. [Les07] s'appuie sur les travaux précédemment cités [Van86,Bor98] provenant du monde des linguistes pour restreindre la définition d'information géographique en géomatique : « *les EGs, auxquelles il s'intéresse dans le corpus, possèdent forcément une composante spatiale (ES) explicite. Celle-ci consiste en une ou plusieurs entités spatiales (ES) (« Pau », « la Maladetta »)* ». Ces travaux définissent un modèle cognitif dit *pivot* (*cf.* figure 4.1) mettant en avant l'entité spatiale telle qu'elle peut être présente dans la molécule géographique décrite ci-dessous.

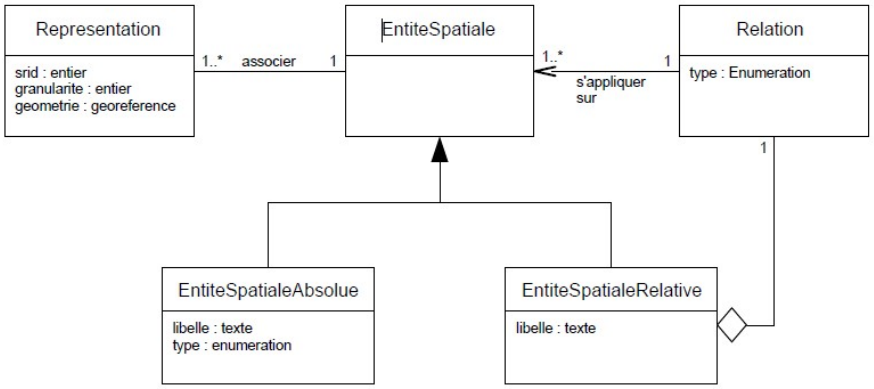


FIGURE 4.1 – Définition du Modèle Pivot [Les07]

Dans ce modèle, la composante spatiale exprimée dans un texte est constituée d'au moins une entité nommée et d'un nombre variable d'indicateurs spatiaux, précisant sa localisation. Nous nous accordons avec [LMR93] qui précise cette définition en indiquant que « *la propriété caractéristique de l'information géographique est d'ancrer ce qui peut être désigné comme phénomène (répondant à la question du 'quoi') dans une localisation spatiale (le 'où') et, régulièrement, temporelle (le 'quand')* ». Cette définition met à la fois en avant l'importance de la composante spatiale et ajoute que l'entité temporelle n'est donc pas toujours explicite et plus difficile à identifier. Cette propriété est présente dans les documents textes et plus particulièrement dans les documents mis à notre disposition comme en témoigne l'ensemble des extraits présentés dans ce mémoire.

Dans la section suivante, nous présentons les techniques existantes pour extraire l'information géographique dans des documents textes et nous dressons ensuite un état des travaux visant à structurer ces informations en une ontologie géographique. Nous verrons que peu de travaux cherchent explicitement à identifier et modéliser des informations propres à un territoire. Nous plaçons notre étude dans le domaine plus large qu'est la géographie.

4.3.2 Le TALN pour l'extraction d'informations géographiques

Comme indiqué précédemment, l'information géographique explicite la relation pouvant exister entre les composantes spatiale, temporelle, et thématique et nous avons vu

dans le chapitre 4.2 (*cf.* page 70) que ces composantes font partie intégrante de la notion plus complexe de territoire que nous souhaitons modéliser. Nous souhaitons donc nous appuyer sur les travaux de recherche visant à identifier et extraire l'information géographique pour proposer une représentation d'un territoire. Dans ce cadre, nous faisons maintenant un rapide état des travaux s'appuyant sur des techniques provenant du TALN pour identifier dans des textes ce type d'informations de façon automatique et de les modéliser pour des traitements ultérieurs. Nous en profitons pour présenter les outils qui sont régulièrement utilisés dans les chaînes de traitement linguistique pour valider les informations identifiées.

4.3.2.1 Processus d'extraction spécialisé

La section 3.2.2.1 (*cf.* page 51) énumère différents travaux provenant du TALN parmi lesquels nous pouvons citer [JKT97, BL02, AFG03, Bil06, Les07, LGN07] qui font émerger une chaîne de traitement linguistique standardisée s'appliquant sur des documents textes permettant d'identifier des concepts et relations. Rappelons que cette chaîne de traitement se décompose en quatre étapes principales que sont la lemmatisation, l'analyse lexicale et morphologique, l'analyse syntaxique et l'analyse sémantique (généralement liée à un domaine d'étude). En géomatique, divers travaux traitent de l'extraction de l'information géographique dans une problématique plus vaste de recherche d'information. Généralement, ces travaux cherchent à identifier l'information spatiale pour faire émerger l'information géographique englobante. Dans ce cadre, les approches statistiques sont alliées à d'autres techniques (notamment linguistiques et conceptuelles) afin de répondre à des problématiques spécifiques à ce type d'information [WP94, SVV01, JPR+02, FJA05, LF04, BE05, GSE+08, LNG08]. Nous détaillons maintenant les approches visant à identifier et annoter l'information spatiale et temporelle en vue d'extraire l'information géographique correspondante.

Méthodes d'extraction d'informations spatiales

Globalement, la méthologie mise en place s'appuie sur des patrons définis pour décrire l'information spatiale. Un patron est défini ici sous la forme de critères ou règles qui vont permettre de sélectionner un ensemble de termes dans les documents traités. Nous nous intéressons plus particulièrement aux travaux de [Les07, SBLG07]. La particularité de la chaîne proposée se situe au niveau de l'analyse morpho-syntaxique qui permet d'extraire les entités nommées spatiales [Lei04, Pas04, Les07, SM03], puis au niveau de l'analyse sémantique qui permet d'identifier des relations spatiales entre entités à partir de l'analyse de différents marqueurs (préposition, verbes, etc.) [Van87, Lau91, Bor98, Aur08]. La phase d'extraction des entités nommées s'appuie sur une ressource de noms de lieux (de type Gazetteer, ontologique ou encore SIG), et intègre généralement une étape de désambiguïsation en prenant en compte le contexte dans lequel est exprimée l'information. Dans cette optique, une première approche se base sur des grammaires définies par apprentissage (Machine Learning) [Gai02] ou définies manuellement [Lou08]. [LGN07] utilise des marqueurs respectant la construction de type [verbe de déplacement, préposition ?, syntagme ?, toponyme], aussi notée (V,P ?,_,E) pour définir la grammaire

à appliquer. Cette construction, intégrant des verbes de déplacement, permet dans une certaine mesure de lever une grande partie des problèmes d'ambiguïté que l'on peut trouver dans des propositions comme « quitter son mari », « traverser une mauvaise période », etc. Une seconde approche utilise des ontologies pour identifier une entité dans un contexte particulier [EFF08].

Rappelons que nous cherchons dans notre cas à identifier la composante spatiale à partir de laquelle nous souhaitons identifier les composantes thématiques et temporelles (lorsqu'un lien explicite existe au niveau de la phrase) pour définir l'information géographique correspondante. [Lei04] met en avant l'ambiguïté référentielle liée à l'identification et à l'extraction d'index géographiques. Aussi, pour lever cette ambiguïté, nous nous intéressons ici à la problématique du marquage des noms toponymiques selon un couple *Nom propre* et expressions antéposées décrivant des relations spatiales ou « *indirections* ». Par exemple, dans la phrase, « nous nous situons au centre ville de Pau », l'objectif est dans un premier temps d'identifier et d'extraire l'information géographique correspondante à l'expression « au centre ville de Pau » à partir du nom toponymique **Pau** afin d'être le plus précis possible. Dans un second temps, nous souhaitons relier l'information spatiale identifiée au qualifiant « centre ville » correspondant à la composante thématique pour construire une information géographique. En accord avec [EJ06], nous décomposons l'ensemble des travaux visant à identifier et annoter les entités nommées constituant une information spatiale en trois catégories : ceux dont le but est de désambiguïser les EN, ceux cherchant à construire une ressource spécifique pour le traitement des EN, et enfin ceux combinant les deux précédents, c'est à dire cherchant à faire de la désambiguïsation tout en exploitant une ressource spécifique. Parmi les travaux proposant de traiter finement les EN en s'appuyant sur des ressources spécifiques, [BP06] présentent une approche de désambiguïsation d'EN qui exploite la ressource encyclopédique Wikipédia. [Pas04] cherche à construire une ressource à partir de corpus pour annoter finement les EN. Cependant, ici, la ressource n'est pas construite pour regrouper des frontières administratives (villes, régions, pays, etc.) mais a pour objectif de structurer l'ensemble des thèmes liés à l'espace pour construire une représentation d'un territoire (objets physiques mais également des activités sportives ou autres, des évènements, etc.).

Méthodes d'extraction d'informations temporelles

Comme nous l'avons vu, la composante temporelle n'est pas toujours explicite dans les évocations d'informations géographiques dans les textes. Nous souhaitons prendre en compte cette information lorsque nous sommes en mesure d'identifier au niveau de la phrase un lien explicite avec les composantes spatiales et thématiques. De nombreux travaux depuis l'introduction des cadres de discours de [Cha97] se focalisent sur l'analyse de la temporalité au sein des textes. En géomatique notamment, l'une des perspectives majeures offertes par les techniques de TALN est de pouvoir identifier et extraire la dimension temporelle pour l'intégrer dans des systèmes d'information géographique. Les travaux de [MST03, MV07] par exemple s'appuient sur un SIG intégrant la composante temporelle pour proposer une application permettant aux utilisateurs de voyager dans le temps en s'appuyant sur l'espace (les cartes de Cassini liées au territoire français).

[BMRS06] décompose les travaux du TALN traitant la composante temporelle en deux groupes : ceux cherchant à s'appuyer sur un système calendaire pour identifier des expressions temporelles [MUC98, SH01, LPLGS07] et ceux s'appliquant à calculer l'ordonnancement dans un texte à travers les évènements idéntifiés dans le document [MW00, MT04, FGM+05].

Nous allons maintenant présenter les travaux visant à construire une ontologie géographique.

4.4 Construction d'ontologies géographiques

4.4.1 Les différents types d'ontologies géographiques

Les ontologies de domaine s'attachent à décrire le vocabulaire particulier du domaine concerné. L'information géographique n'échappe pas à cette particularité. De manière classique, les ontologies géographiques peuvent être utilisées pour l'exploration, ou encore pour l'extraction d'informations, et au-delà pour l'interopération de SIG [Kok06, HIC07, AM10]. La multi dimensionnalité des objets géographiques confère des spécificités qui, en accord avec [CPSV03], nous permettent de définir trois sortes d'ontologies géographiques :

1. les **ontologies cartographiques** (au sens géographique) plus spécifiquement dédiées à la description des concepts qui caractérisent l'espace comme le point, la ligne, etc. Ces ontologies sont typiquement élaborées par de grands organismes de normalisation, par exemple, l'OpenGIS par le biais du langage GML (Geography Markup Language).
2. les **ontologies de domaines géographiques** comme une ontologie modélisant les concepts des données hydrauliques, ou encore les données des réseaux électriques, etc. Ce sont des ontologies développées par une communauté d'utilisateurs du domaine concerné. Divers travaux ont été réalisés dans le cadre particulier des ontologies dans le domaine géographique [Min08]. [Uit01, Bro04] mettent en avant la nécessité de ces ontologies. Nous pouvons notamment citer les travaux de l'Ordnance Survey qui proposent une ontologie construite manuellement décrivant les entités administratives de façon générale et décrivant de façon précise le domaine de l'hydrologie, ceux du GCMD [63] proposant une ontologie relatant des Sciences de la Terre, ou encore les travaux intégrant une description détaillée de la géographie dans une problématique globale de représentation de la connaissance. Parmi ces travaux, nous pouvons notamment citer l'IEEE Standard Upper Working Group, qui dans le cadre de la construction d'une ontologie de haut niveau SUMO [64], ont construit une ontologie géographique détaillée (avec notamment plus de 400 classes), l'UNESCO [65], qui dans le cadre da la construction d'un thésaurus gé-

63. GCMD : Global Change Master Directory ; http ://gcmd.nasa.gov/index.html
64. http ://www.ontologyportal.org/
65. http ://www2.ulcc.ac.uk/unesco/index.htm

néral multilingue décrit le domaine de la géographie ou encore EUROVOC [66] qui propose un thésaurus multilingue couvrant les subdibvisions administratives et des éléments politiques liés à l'Europe parmi l'ensemble des domaines traités liés à l'activité de l'Union Européenne et enfin au niveau national l'ontologie géographique de l'INSEE avec le projet SchemaWeb [67] qui décrit les subdivisions géographiques de la France, le thésaurus RAMEAU [68] qui intègre également une structuration administrative du pays, ou enfin le projet GEONTO [69] qui propose une ontologie géographique définie automatiquement à partir de spécifications de bases de données de l'IGN [70] [AM10].

En majorité, les travaux autour de la construction d'une ontologie de domaines géographiques cherchent à décrire des sous-domaines, comme Towntology [71] [RLB+04] ou FoDoMuSt [72] [BBG+07] dans le domaine de l'urbanisme, le thésaurus multilingue AGROVOC [73] ou encore le projet français GIEA [74] [DAP06] dans le domaine de l'agriculture.

3. les **ontologies spatialisées** (ou spatio-temporelles), qui sont des ontologies dont les concepts sont localisés dans l'espace. Nous pouvons notamment citer l'outil ONTOAST [DMGVOM07] qui permet d'éditer et d'interroger des ontologies spatio-temporelles. Cependant, les frontières entre ces domaines ne sont pas nettement délimitées. Nous en voulons pour preuve les travaux menés dans le cadre de la directive européenne INSPIRE [Ins07] qui vise la mise en place d'une infrastructure d'information géographique au niveau européen.

Comme nous pouvons le voir, les travaux réalisés autour de la construction d'ontologies sont nombreux et récents en géomatique. Mais ils sont pour la majorité des cas orientés vers une représentation administrative ou politique d'un espace comme le projet SchemaWeb, l'ontologie proposée par l'Ordnance Survey ou encore le projet EUROVOC.

4.4.2 Le « cas » Territoire

Si des travaux en géomatique proposent des ontologies géographiques ou des méthodes pour les construire, nous n'avons pas identifié d'ontologie existante proposant une description adaptée à notre problématique. Si les dimensions administratives et/ou politiques sont prises en compte dans les travaux étudiés, la composante sujet (liée à l'intervention de l'Homme) n'est que très peu abordée et n'est pas intégrée dans les ontologies

66. http ://eurovoc.europa.eu/
67. http ://www.schemaweb.info/schema/SchemaDetails.aspx ?id=283
68. http ://rameau.bnf.fr/
69. http ://geonto.lri.fr/
70. Institut National de Géographie
71. http ://www.towntology.net/
72. http ://fodomust.u-strasbg.fr/ ; projet orienté traitement d'images
73. Agricultural Information Management (AIMS) Web site. visité le 15 février 2011, from http ://aims.fao.org/website/AGROVOC/sub
74. http ://www.projetgiea.fr

géographiques. Dans le cadre du projet CAVALA [75], des travaux en cours [COL09, Sal09] s'appliquent actuellement à construire à partir de textes une ontologie du territoire, intégrant des notions sociales et culturelles, dans le domaine du développement économique territorial. L'objectif de ces travaux est de pouvoir construire un modèle sémantique permettant d'analyser les changements à l'échelle de la région, le cas d'application étant la région Midi-Pyrénées. Ces travaux, bien qu'intéressants, s'appliquent à modéliser des éléments spécifiques liés à l'économie et ne correspondent donc pas réellement à nos besoins de modéliser de façon précise, un territoire à la fois administratif, politique et culturel à partir de documents textuels. Cependant, ces travaux, s'ajoutant aux besoins de centres documentaires tels que la MIDR de Pau de valoriser leurs fonds documentaires, nous confortent dans l'intérêt de construire une représentation d'un territoire sous forme d'ontologie.

4.5 Discussion

Nous avons pu remarquer que la notion de territoire est une notion complexe pour laquelle plusieurs définitions sont proposées et cela dans de nombreux domaines tels que la géographie, la sociologie, etc. L'encyclopédie Universalis définit le mot territoire de façon générale comme l'« étendue de la Terre sur laquelle vit un groupe humain ». Nous nous appuyons tout d'abord sur les travaux des géographes, pour qui le territoire est une notion centrale, afin de suggérer une définition, reconnue quel que soit le domaine, mettant en avant un ensemble d'éléments que l'on puisse modéliser. [Gui07] met en avant la relation entre l'Homme et la Terre en définissant la notion de **territoire** comme la superposition d'un espace et de pratiques sociales. [DMB05] indique que le territoire regroupe et associe des lieux. Dans le cadre de nos travaux, le territoire décrit dans un fonds documentaire correspond alors à un réseau défini par des lieux et par les relations identifiées entre ces lieux.

[DMB05] indique également que le territoire permet d'appréhender le phénomène qu'est le contact vécu de l'homme avec le milieu. Dans ce cadre, le contact vécu de l'Homme fait référence à la composante thématique et également à la composante temporelle. Le milieu fait lui référence à la composante spatiale et correspond à un ensemble de lieux qui forment un espace. [Ent96, BE98] mettent également en avant la relation entre les composantes sujet et temporelle avec la composante spatiale ; la composante spatiale étant vue alors comme un ensemble de lieux formant un espace défini selon la composante sujet. En somme, nous pouvons définir de façon simple un territoire comme un ensemble de lieux que l'on peut mettre en relation selon la composante sujet en fonction d'une periode donnée.

Concernant la modélisation de la connaissance en géomatique, beaucoup de travaux parmi lesquels nous pouvons citer [AM10, RLB+04, BBG+07] s'appliquent à constuire une ontologie géographique de domaines cibles. Cependant, il ne semble pas exister d'ap-

75. CAVALA, méthode Coopérative de suivi et d'éVAluation des poLitiques régionAles de développement économique, Région Midi-Pyrénées/CCRRDT, appel d'offres SHS 2007, Action-clé 10 « Evaluation des nouvelles politiques »

proche permettant de construire une ontologie du territoire à partir d'une ressource documentaire préexistante. En effet, parmi l'ensemble des ressources structurées (thésaurus ou ontologies géographiques), beaucoup sont structurées de façon administrative, restent très générales et ne permettent pas de représenter de façon suffisament précise le territoire décrit dans des documents. Nous nous rapprochons des travaux de [COL09, Sal09] qui s'appliquent actuellement à construire à partir de textes une ontologie du territoire mais ces travaux cherchent à représenter les évolutions économiques d'une région et donc un sous ensemble de ce qu'est un territoire.

Cependant, en géomatique, de nombreux travaux récents proposent d'identifier et de structurer des informations géographiques. Ces informations, également appelées informations géographiques, sont définies comme des molécules formées d'une composante spatiale, d'une composante temporelle et d'une composante thématique ou phénomène [Gal01, UTC04, PSA07].

Dans notre approche visant à construire automatiquement une représentation d'un territoire à partir de documents textes, une première étape consiste donc à identifier et extraire l'ensemble des informations géographiques présentes dans les documents textes traités. Parmi les travaux cherchant à identifier et extraire ces informations dans des textes, nous souhaitons nous appuyer sur les travaux réalisés au sein de notre laboratoire dans le cadre du projet PIV [SBLG07] qui proposent une chaine de traitement linguistique pour le traitement d'informations spacialisées. La chaine proposée intègre une étape d'identification et d'annotation des entités nommées constituant une information spatiale avec une phase intérmédiaire de désambiguïsation d'EN à partir d'un lexique externe. Nous nous rapprochons des travaux de [BP06] qui présentent également une approche de désambiguïsation d'EN mais le lexique utilisé, en l'occurrence la ressource encyclopédique Wikipédia, est très générale et n'est pas réellement représentative du domaine cible. Dans nos travaux, nous tentons de traiter finement les EN en nous appuyant sur un lexique structuré représentant le domaine cible pour faciliter le marquage de thèmes. L'utilisation d'un lexique structuré doit également nous permettre d'identifier plus précisément les relations entre la composante thématique et les composantes spatiales et temporelles formant l'information géographique.

En accord avec [Bac00], nous pensons que la construction d'ontologies dépend du domaine cible et des applications pour lesquelles elles sont définies. Dans notre cas, l'ontologie de territoire doit être construite pour représenter un territoire donné décrit dans un ensemble de documents. Notre contribution ne correspond pas uniquement à instancier une ontologie pour un contexte documentaire donné afin de faire émerger le territoire mais s'attache également à proposer une méthodologie complète automatisée permettant de construire une ontologie d'un territoire pour un fonds documentaire. Dans notre proposition, nous nous appuyons sur le fonds documentaire annoté par des experts ainsi que sur le vocabulaire contrôlé utilisé par les experts pour réaliser le travail d'annotation. Nous faisons l'hypothèse que l'extraction d'informations locales dans les documents et les notices descriptives, avec un filtre géographique, fait émerger une représentation territorialisée de la base documentaire traitée. Nous nous positionnons à cheval entre les ontologies de domaine géographique et les ontologies spatialisées. En effet, nous souhaitons

pouvoir construire une ontologie d'un territoire intégrant des informations spatialisées. L'annotation visant à identifier ces dernières s'appuie sur des ressources géographiques diverses (BD de l'IGN, gazetteers contributives) pour le typage et la validation d'EN spatiales candidates puis le calcul de géométrie correspondante.

Nous allons maintenant proposer une conclusion générale de l'état de l'art offrant une synthèse des travaux visant à identifier et structurer des connaissances pour un domaine cible à partir de documents textes annotés. En nous appuyant sur l'ensemble des définitions de la notion de territoire, nous proposons un noyau formel d'une définition de la notion de territoire afin d'en proposer ensuite une vue « grand public ».

Chapitre 5

Synthèse et proposition d'un modèle du territoire

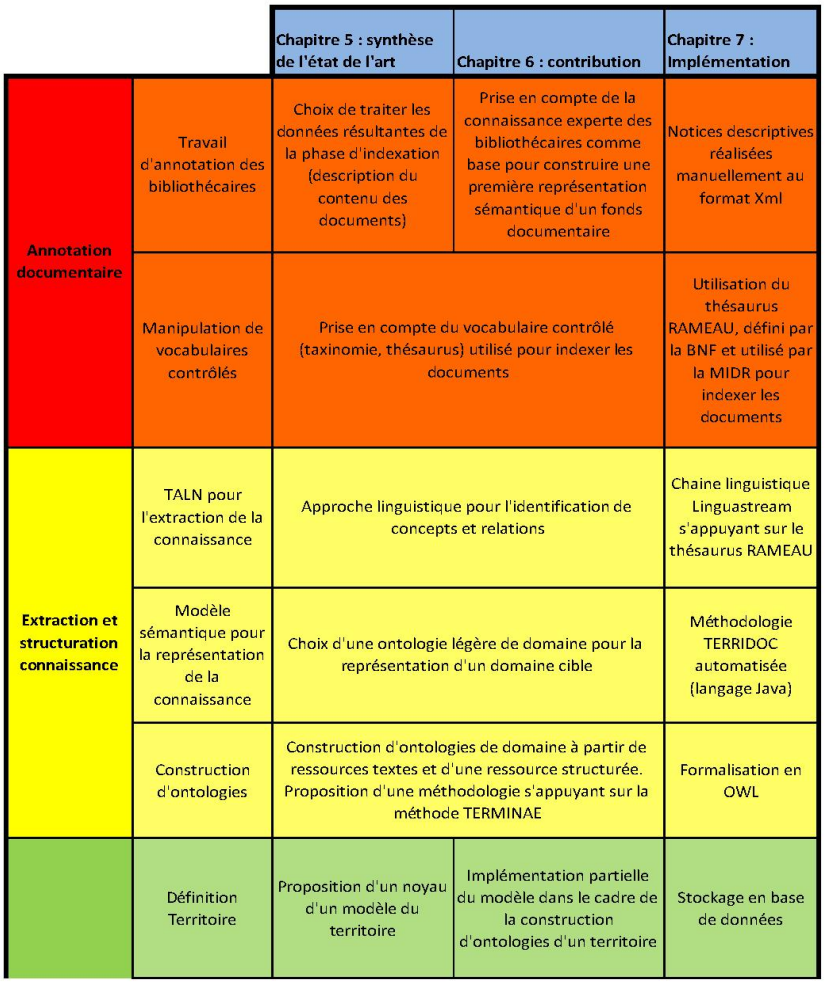
Sommaire

5.1 Synthèse de l'état de l'art 85
5.1.1 Un fonds documentaire annoté comme base de travail . . . 87
5.1.2 Choix de l'ontologie pour la représentation de la connaissance 87
5.2 Le territoire dans un espace documentaire 88
5.2.1 Définitions 88
5.2.2 Noyau de modèle du territoire 88
5.2.3 La composante spatiale 90
5.2.4 La composante temporelle 91
5.2.5 La composante thématique 92
5.2.6 Modèle pour la représentation d'un territoire 93

5.1 Synthèse de l'état de l'art

Nous concluons cette deuxième partie en présentant sous forme de tableau notre positionnement au sein de l'état de l'art présenté (*cf.* figure 5.1). Nous situons notre approche dans l'extraction et la structuration de la connaissance que nous appliquons dans le domaine de la géomatique en nous appuyant notamment sur des techniques provenant du TALN.

Dans une démarche de description de fonds documentaires dans les centres documentaires de types bibliothèques et médiathèques, nous proposons de construire, sur la base du travail d'annotations de documents réalisé par des experts, une couche conceptuelle de type ontologie d'un domaine cible. Rappelons qu'une caractéristique importante des fonds documentaires mis à disposition dans ce type de centres documentaires, comme celui mis à disposition par la MIDR de Pau sur lequel nous travaillons, est qu'ils contiennent d'abondantes références au territoire, et nous souhaitons utiliser ces spécificités ainsi que



		Chapitre 5 : synthèse de l'état de l'art	Chapitre 6 : contribution	Chapitre 7 : Implémentation
Annotation documentaire	Travail d'annotation des bibliothécaires	Choix de traiter les données résultantes de la phase d'indexation (description du contenu des documents)	Prise en compte de la connaissance experte des bibliothécaires comme base pour construire une première représentation sémantique d'un fonds documentaire	Notices descriptives réalisées manuellement au format Xml
	Manipulation de vocabulaires contrôlés	Prise en compte du vocabulaire contrôlé (taxinomie, thésaurus) utilisé pour indexer les documents		Utilisation du thésaurus RAMEAU, défini par la BNF et utilisé par la MIDR pour indexer les documents
Extraction et structuration connaissance	TALN pour l'extraction de la connaissance	Approche linguistique pour l'identification de concepts et relations		Chaine linguistique Linguastream s'appuyant sur le thésaurus RAMEAU
	Modèle sémantique pour la représentation de la connaissance	Choix d'une ontologie légère de domaine pour la représentation d'un domaine cible		Méthodologie TERRIDOC automatisée (langage Java)
	Construction d'ontologies	Construction d'ontologies de domaine à partir de ressources textes et d'une ressource structurée. Proposition d'une méthodologie s'appuyant sur la méthode TERMINAE		Formalisation en OWL
	Définition Territoire	Proposition d'un noyau d'un modèle du territoire	Implémentation partielle du modèle dans le cadre de la construction d'ontologies d'un territoire	Stockage en base de données

FIGURE 5.1 – Positionnement de nos travaux au sein de l'état de l'art présenté

les connaissances expertes des bibliothécaires pour construire une représentation du territoire implicitement décrit dans un fonds documentaire.

5.1.1 Un fonds documentaire annoté comme base de travail

Nous nous appuyons sur un corpus documentaire annoté par des experts bibliothécaires ainsi que sur le vocabulaire contrôlé utilisé pour réaliser ce travail d'annotation. Dans ce travail d'annotation, nous nous intéressons plus particulièrement à la phase d'« indexation » du document qui consiste à décrire un document à l'aide de représentations des concepts contenus dans ce document. Dans ce contexte, ces concepts sont choisis dans un vocabulaire contrôlé défini au préalable. Le travail d'indexation réalisé par les experts a pour objectif de décrire le contenu des documents mais n'a pas pour objectif explicite de mettre en valeur un territoire. Cependant, nous faisons l'hypothèse que les spécificités liées aux fonds documentaires dits territorialisés nous permettent d'exploiter ce travail de description pour construire une première représentation d'un territoire.

Parmi les vocabulaires contrôlés, les thésaurus sont régulièrement utilisés dans les centres documentaires pour modéliser la connaissance dans un domaine car ils comportent des propriétés potentiellement exploitables. Les règles sémantiques, terminologiques et syntaxiques caractérisant ce type de structure sont autant d'outils qui permettent de modéliser et structurer la connaissance dans un domaine cible. Le fait que les informations de description choisies par les experts bibliothécaires pour décrire le contenu des documents soient extraites d'un vocabulaire contrôlé, régulièrement structuré sous forme d'un thésaurus, nous permet d'exploiter le travail d'expert notamment sur le choix de la sémantique, de la terminologie et la syntaxe des termes sélectionnés pour décrire des documents.

5.1.2 Choix de l'ontologie pour la représentation de la connaissance

A partir d'un fonds documentaire annoté et du vocabulaire contrôlé utilisé durant la phase d'indexation, nous souhaitons mettre en place une méthodologie pour extraire cet ensemble de connaissances structurées et les « projeter » le plus automatiquement possible dans une structure de type ontologie. L'ontologie propose les outils pour répondre aux limites identifiées dans les autres vocabulaires contrôlés : manque de sémantique dans la définition des propriétés des concepts ainsi que dans la définition des relations. Nous positionnons nos travaux dans la construction d'une ontologie légère. Nous souhaitons construire l'ontologie de façon automatique et l'intégrer ensuite dans un système de recherche d'information destiné à tout type d'utilisateurs et nous faisons le choix de ne pas complexifier la structure de l'ontologie par des d'axiomes ou des restrictions. Nous nous rapprochons des travaux de [CHGM06] qui, sur la base de la méthodologie TERMINAE [BS99, AGBS00, BAGC04], propose de transformer un thésaurus du domaine cible en une ontologie.

En accord avec [Bac00] précisant que l'usage prévu de l'ontologie contraint et encadre sa construction pour un domaine cible, nous cherchons à construire une ontologie légère de domaine. Notons tout de même que cette conceptualisation est souvent qualifiée de partielle car il semble présomptueux de croire pouvoir formaliser dans une même structure sémantique toute la complexité d'un domaine.

Dans notre cas, l'ontologie de territoire doit être construite pour représenter un ter-

ritoire donné décrit dans un ensemble de documents. Notre contribution ne correspond pas uniquement à instancier une ontologie pour un contexte documentaire donné afin de faire émerger le territoire mais s'attache également à proposer une méthodologie complète automatisée permettant de construire une ontologie d'un territoire pour un fonds documentaire. Nous faisons l'hypothèse que l'extraction d'informations locales dans les documents et les notices descriptives, avec un filtre géographique, fait émerger une représentation territorialisée de la base documentaire traitée. Nous nous positionnons à la fois dans construction d'ontologies de domaine géographique et la construction d'ontologies spatialisées. En effet, nous souhaitons pouvoir construire une ontologie d'un territoire intégrant des informations spatialisées.

5.2 Le territoire dans un espace documentaire

5.2.1 Définitions

Le chapitre 4 (*cf.* page 69) met en avant la complexité de la notion de territoire pour laquelle plusieurs définitions sont proposées et cela dans de nombreux domaines tels que la géographie, la sociologie, etc. Le mot territoire est défini de façon générale par l'encyclopédie Universalis comme l'« étendue de la Terre sur laquelle vit un groupe humain ». Pour définir ce que nous ententons par territoire, nous nous appuyons sur les travaux des géographes pour qui le territoire est un concept faisant partie intégrante de leur problématique scientifique, parmi lesquels nous pouvons citer [Pio92, Sch94, LL03, GV04, DMB05, Gui07].

Définition 1 Sur la base des travaux de [DMB05] et [Gui07] qui s'attachent à mettre en valeur la notion de lieux au sein d'un territoire, nous pouvons affirmer tout d'abord que : *le territoire décrit dans un fonds documentaire correspond à un réseau défini par des lieux et par les relations identifiées entre ces lieux.*

Cette première définition fait intervenir la composante spatiale, à travers des lieux, comme un élément important.

Définition 2 Sur la base des travaux de [Ent96, BE98], [DMB05] décrit le territoire comme *le contact vécu de l'homme avec le milieu* et intègre les composantes thématiques (ce qui attrait à l'Homme) et temporelle (le contact vécu instaure une notion de temps plus ou moins long) dans la définition d'un territoire.

En somme, nous pouvons définir de façon simple un territoire comme un ensemble de lieux que l'on peut mettre en relation selon la composante thématique en fonction d'une periode donnée.

5.2.2 Noyau de modèle du territoire

Afin de définir un premier modèle général de ce que l'on entend par territoire, nous nous appuyons sur les définitions que nous venons de proposer et notamment sur la définition 2 précisant qu'un territoire peut être défini comme un ensemble de lieux (selon

la composante spatiale donc) mis en relation selon un ensemble de sujets (la composante thématique) et cela dans une periode donnée (la composante temps). Nous avons vu chapitre 4.2.6 (*cf.* page 75) que l'information géographique met également ces trois composantes en relations sous forme d'une entité géographique (EG), composée d'une entité thématique (ou phénomène), d'une entité spatiale(ES) et d'une entité temporelle (ET) pouvant être implicite. Nous retenons la définition proposée par [LMR93] qui précise que « la propriété caractéristique de l'information géographique est d'ancrer ce qui peut être désigné comme phénomène (répondant à la question du 'quoi') dans une localisation spatiale (le 'où') et, régulièrement, temporelle (le 'quand') ». Cette définition met à la fois en avant l'importance de la composante spatiale et ajoute que l'entité temporelle n'est donc pas toujours explicite et plus difficile à identifer. La composante spatiale est alors au cœur de l'information géographique et doit permettre de caractériser la composante thématique. Cette propriété est présente dans les documents textes et plus particulièrement dans les documents annotés mis à notre disposition.

En nous appuyant sur la définition 2 (*cf.* 5.2.1) ainsi que sur la notion d'information géographique, nous spécifions la notion de territoire comme un ensemble d'informations géographiques (*cf.* figure 5.2) dans lesquelles seules les composantes spatiales et thématiques sont nécessaires.

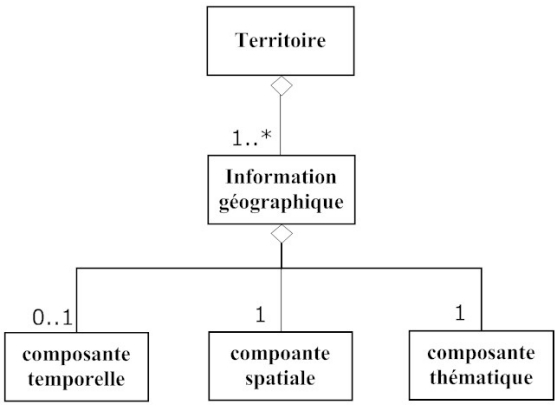


FIGURE 5.2 – Noyau de modèle du territoire

Nous allons maintenant détailler chacune des composantes de notre modèle et pour bien comprendre le champ d'application de nos travaux, nous nous appuyons tout au long de ce chapitre sur six extraits de phrases intégrant une entité géographique.

1. « C'est un document décrivant Auriac au 12 juin 1876 » ;
2. « Nous sommes au nord du pic d'Ossau » ;
3. « Ce document décrit le massif de la Maladetta au début des années 60 » ;
4. « Cette photo du mariage est prise à la mairie de Jurançon au XIXème siècle » ;
5. « La vallée d'Aspe est magnifique » ;
6. « Autour de Pau, la région est montagneuse » ;
7. « La commune de Pau durant la révolution française ».

5.2.3 La composante spatiale

Nous nous accordons avec des travaux réalisés au sein de notre équipe [Les07] qui proposent une définition plus restreinte de l'entité géographique que celle communément acceptée maintenant définie par [Gal01, UTC04, PSA07] entre autres, et dans notre équipe [Gai01] : « les entités géographiques (EG), auxquelles il s'intéresse dans le corpus, possèdent forcément une composante spatiale (ES) explicite. Celle-ci consiste en une ou plusieurs entités nommées de lieux (« la vallée d'Aspe », « Auriac ») ». [Les07] définit le modèle Pivot (*cf.* figure 4.1 page 77) dans lequel une entité spatiale est soit absolue (ESA : référence qui peut être directement localisées géographiquement) (exemples 1, 3, 4, 5 et 7), soit relative (ESR : composée d'entités spatiales absolues et de relations spatiales) (exemples 2 et 6). Dans le cadre de nos travaux, nous nous intéressons uniquement aux ESA définies comme des entités spatiales simples, en l'occurrence des noms toponymiques (*cf.* figure 5.3).

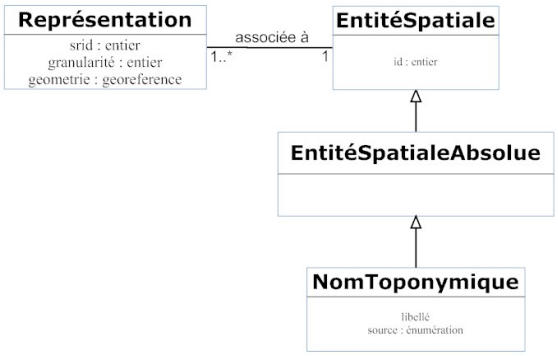


FIGURE 5.3 – Définition de la composante spatiale

Dans ce modèle, le nom toponymique peut être défini comme un nom d'un lieu. Le nom toponymique constitue à lui seul une entité spatiale absolue. Dans nos travaux, il est caractérisé par le nom de la ressource d'où provient l'information validant que c'est bien un lieu (par exemple, la base de données BDTopo de l'IGN) ainsi que par un libellé correspondant au nom du lieu. Si nous prenons les exemples d'entités géographiques présentées précédemment, les entités géographiques retenues sont les entités 1, 3, 4, 5 et 7 les noms toponymiques ici étant respectivement « Auriac », « Maladetta », « Jurançon », « Aspe » et « Pau ». Les entités 2 (« au nord du pic d'Ossau ») et 6 « autour de Pau » étant relatives, nous ne les retenons pas.

Au cœur de l'information géographique, nous nous appuyons sur la composante spatiale dans les traitements linguistiques que nous réalisons sur les documents textes et sur les notices attachées pour identifier les composantes temporelle (si elle existe) et thématique.

5.2.4 La composante temporelle

Comme nous l'avons vu précédemment, l'entité temporelle « peut être implicite, c'est-à-dire qu'elle n'est pas mentionnée directement dans le texte mais découle d'informations annexes » [Les07, LPLGS07]. L'ET peut être éloignée des autres composantes formant l'EG (par exemple dans le cas d'un journal de bord, la date est marquée en début de paragraphe et le reste de l'unité de texte décrit des phénomènes se passant à cette date). L'ET peut donc aussi être associée à plusieurs EG, quand elle recouvre un paragraphe entier par exemple. Cependant, nous souhaitons prendre en compte cette composante dans nos travaux pour tenter d'identifier une période globale décrite par l'ensemble des documents. Et, dans le cas où nous ne sommes pas en mesure de renseigner une ET pour les EGs présentes dans un document, nous proposons d'exploiter les informations présentes dans la notice descriptive associée pour définir cette ET. [LPLGS07] s'appuie sur le modèle Pivot [Les07] pour proposer un modèle de l'entité temporelle équivalent au modèle de l'entité spatiale. L'entité temporelle est définie comme une date, un intervalle de dates ou une période. Nous restreignons la notion de composante temporelle aux données calendaires absolues et nous ne prenons donc pas en compte les entités nommées tels que les évènements dans l'histoire (*cf.* exemple 7 contenant l'ET « La révolution française ») qui nécessite une base de connaissances que nous ne possédons pas. Sur la base du premier modèle proposé figure 5.3 et des travaux de [LPLGS07], nous proposons le modèle présenté figure 5.4.

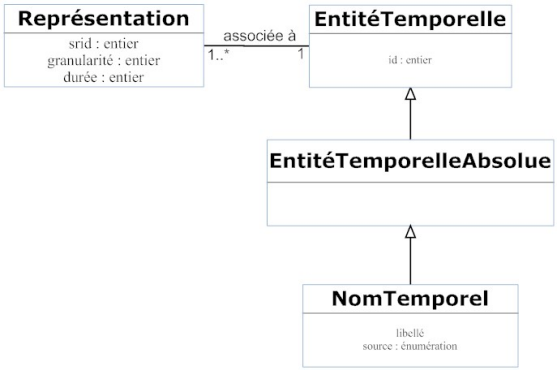


FIGURE 5.4 – Définition de la composante temporelle

A partir des septs exemples donnés précédemment, nous retenons les expressions 1 (« Auriac au 12 juin 1876 »), 3 (« le massif de la Maladetta au début des années 60 ») et 4 (« la mairie de Jurançon au XIXième siècle ») comme des informations contenant des entités temporelles. Les entités temporelles sont ici « 12 juin 1876 », « au début des années 60 » et « XIXième siècle ». Le nom temporel peut constituer à lui seul une entité temporelle absolue. Il est caractérisé par la durée en jours (le jour étant le grain le plus fin dans notre modèle) utilisé lorsque le nom temporel est une période ainsi que par un libellé correspondant au nommage de la date ou de la période.

Les définitions de la notion de territoire ainsi que les modèles proposés mettent en

avant l'importance de la composante spatiale et la possibilité d'exploiter la composante temporelle dans l'identification d'un territoire. Dans ce cadre, l'identification de la composante thématique se fait en s'appuyant sur les composantes spatiales et temporelles et nous faisons l'hypothèse que la structure sémantique obtenue à partir des relations que nous identifions entre ces composantes fait émerger une représentation d'un territoire.

5.2.5 La composante thématique

En accord avec [Les07] le thème (phénomène), ou composante thématique, correspond *a priori* à tout ce qui n'est pas spatial ou temporel dans le texte. C'est le thème dont il est question à un lieu et un instant donné (botanique, architectures, etc.). Nous proposons d'apporter des éléments de réponse concernant la composante thématique lorsque nous pouvons établir un lien avec les composantes spatiale et temporelle. Le thème apparait alors comme une précision, ou un qualifiant, d'un lieu ou d'une période pour former une entité géographique. Dans ce sens, le terme **qualifiant**, lorsqu'il existe au sein d'une information géographique identifiée dans un document texte, est défini comme un mot ou un groupe de mots (renseignant la composante thématique) décrivant le lieu qu'il désigne.

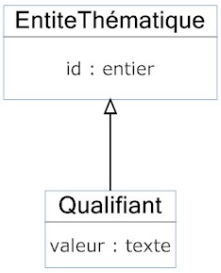


FIGURE 5.5 – Définition de la composante thématique

A partir des exemples donnés (de 1 à 7), les exemples 3 (« le massif de la Maladetta au début des années 60 »), 4 (« la mairie de Jurançon au XIXième siècle »), 5 (« la vallée d'Aspe ») et 7 « La commune de Pau » sont des entités géographiques intégrant un lien entre une composante thématique et une composante spatiale, voire une composante temporelle. Dans l'exemple 3, l'entité spatiale est « la Maladetta » et l'entité temporelle « au début des années 60 » et nous nous appuyons sur ces entités pour identifier le terme qualifiant, en l'occurrence ici « massif ». Dans l'exemple 4, l'entité spatiale est « Jurançon », l'entité temporelle « XIXième siècle » et l'entité thématique « mairie ». Dans l'exemple 5, le qualifiant est « vallée » et est ancré sur l'entité spatiale « Aspe ». Dans l'exemple 7 enfin, le qualifiant est « commune » et est ancré sur l'entité spatiale « Pau ».

Notons que dans une ressource toponymique (gazetteers, SIG [76], etc.), un nom toponymique peut avoir plusieurs représentations spatiales, ce qui peut poser problème

76. Système d'Information Géographique

lorsque l'on souhaite identifier l'espace pointée par ce nom. L'identification d'un qualifiant, qui précise la sémantique de l'entité spatiale attachée, doit permettre de désambiguiser de façon complète ou partielle le choix de la représentation spatiale adéquate. L'ensemble des entités géographiques extraites des documents sont autant d'informations candidates à enrichir la représentation d'un territoire obtenue à partir du travail d'indexation.

Nous avons défini les trois composantes constituant l'information géographique. Nous proposons maintenant un modèle détaillé de ce que nous entendons par territoire.

5.2.6 Modèle pour la représentation d'un territoire

Après avoir présenté un premier noyau de modèle ainsi que ses constituants, nous proposons maintenant un modèle détaillé de ce que nous entendons par territoire (c.f. figure 5.6).

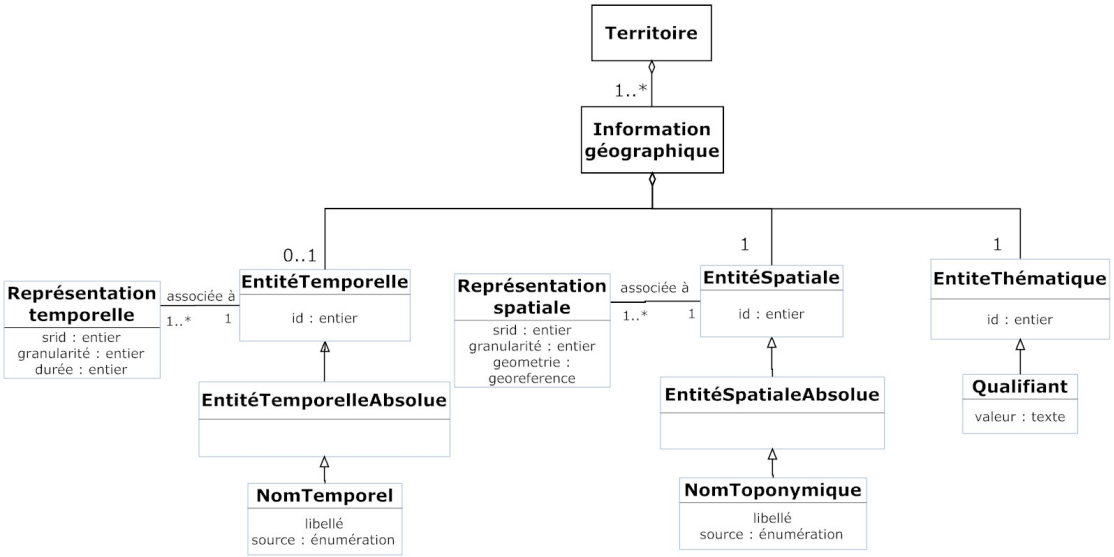


FIGURE 5.6 – Modèle du territoire

Dans ce modèle, nous reprenons le fait qu'un territoire est formé par une ou plusieurs informations géographiques. Dans le cadre de nos travaux, nous nous limitons à identifier et extraire les entités spatiales absolues de type nom toponymique sur lesquelles nous nous appuyons ensuite pour identifier les entités temporelles absolues (entités calendaires tels que les dates, periodes, etc.) et les entités thématiques.

5.2.6.1 Le modèle du territoire vis à vis du thésaurus

Nous attachons tout d'abord la notion d'entité thématique à l'ensemble des termes qui sont utilisés dans les notices descriptives par les experts bibliothécaires pour décrire

le contenu des documents. Ces termes, qu'ils soient des lieux, des dates ou des thèmes, sont choisis sur la base d'un vocabulaire contrôlé défini en amont pour faciliter le travail d'indexation. Le fait qu'un vocabulaire contrôlé de type thésaurus (ou taxonomie) soit utilisé par les experts dans le choix des termes pour décrire un ensemble de documents nous permet de proposer une représentation sémantique synthétique de ce travail d'indexation. Sur la base d'un modèle représentant un thésaurus, une entité spatiale, temporelle ou thématique est, lorsqu'elle est présente dans le thésaurus, une vedette ou un terme « employé pour » dans le vocabulaire contrôlé.

L'utilisation d'une ressource de type vocabulaire contrôlé implique que la liste de termes que nous obtenons intègre le contrôle de la synonymie, de l'homonymie et de la polysémie ainsi que des règles terminologiques et syntaxiques. Ces contrôles nous permettent de valider les termes choisis par les experts, en tant que thèmes. Chaque thème ayant un sens et un seul et ce thème est le seul à avoir ce sens. Il peut cependant être relié à un ensemble de termes dits « termes employés pour » qui ont le même sens dans le contexte du vocabulaire contrôlé. Par exemple, un terme marqué dans un texte comme une composante d'une information géographique (par un traitement TALN notamment) et faisant partie du thésaurus, peut être une vedette ou un terme « employé pour ». De plus, nous nous appuyons sur le fait que le vocabulaire contrôlé utilisé soit un thésaurus, constitué de termes reliés entre eux par des relations hiérarchiques et associatives respectant la norme AFNOR [77], pour décrire le lien entre notre modèle du territoire et une représentation de ce que nous entendons par vocabulaire contrôlé (figure 5.7). Dans ce cadre, un thème, dans le vocabulaire contrôlé, peut être attaché à un ensemble de notes pour le définir et décrire la façon de l'utiliser pour un travail d'indexation.

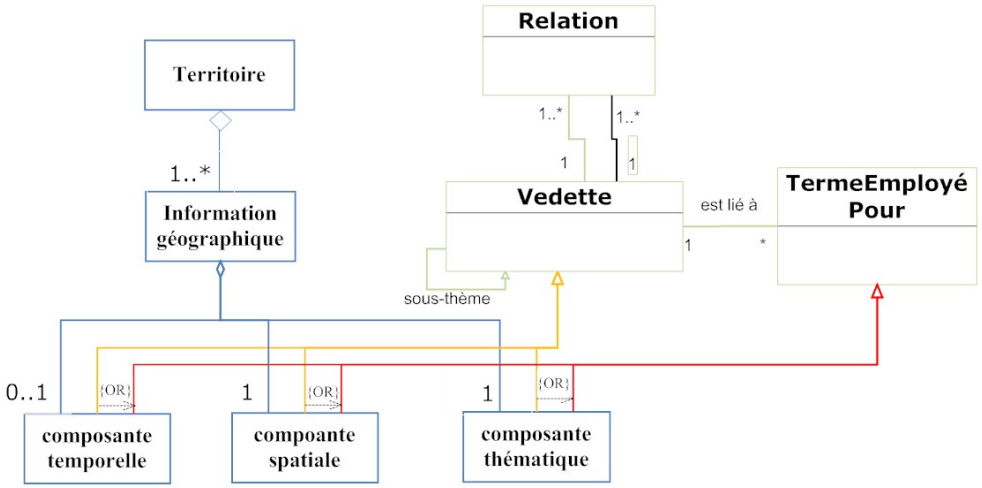


FIGURE 5.7 – Définition d'un territoire vis à vis du thésaurus

Ce modèle, bien que succinct, est un premier élément de réponse dans la représentation sémantique structurée d'un territoire en apportant des informations sur les

77. AFNOR. Documentation : règles d'établissement des thésaurus monolingues. NF Z47-100, 1981.

spécificités culturelles telles que les activités, les types de constructions, etc. Cependant, nous verrons qu'à lui seul, il ne décrit pas un territoire car il ne contient pas d'éléments explicites (propriétés ou autre) permettant de différencier une entité spatiale ou temporelle d'une entité thématique.

5.2.6.2 Le modèle du territoire vis à vis de l'ontologie

Nous faisons le choix d'utiliser le vocabulaire contrôlé de type ontologie pour représenter un territoire. En effet, l'ontologie permet d'expliciter une sémantique plus large qu'un thésaurus, et notamment par le biais d'éléments tels que les propriétés que l'on peut attacher aux concepts pour les décrire plus précisément ou encore de relations entre concepts que l'on peut expliciter. Nous exploitons notamment l'élément « propriété » pour indiquer une représentation spatiale ou temporelle lorsque cela est possible. Nous souhaitons également pouvoir redéfinir des relations hiérarchiques en relation d'instance lorsque cela nous semble nécessaire. Enfin, nous faisons l'hypothèse de pouvoir expliciter des relations associatives en appliquant le point de vue territoire sur l'ontologie.

La figure 5.8 met en avant les liens que nous souhaitons définir entre les éléments du modèle territoire que nous proposons (*cf.* figure 5.6) et le vocabulaire contrôlé de type ontologie. Nous reprenons dans ce schéma les principaux constituant de l'ontologie : ensemble de concepts, regroupant un ou plusieurs labels et pouvant être caractérisés par des propriétés, structurés par le biais de relations diverses (hiérarchiques ou avec une sémantique définie lors de la conception de l'ontologie). Dans le cadre de notre travail, l'ensemble des informations géographiques idéntifiées, que ce soit dans les notices descriptives ou dans les documents, sont ajoutées à l'ontologie sous forme d'instances, auxquelles nous pouvons ajouter une représentation spatiale et temporelle, lorsque l'on est en mesure d'établir un lien entre un concept et l'information géographique traitée. Ce lien est défini par le fait que la composante thématique (le qualifiant dans l'information géographique) est un label d'un concept dans l'ontologie et ce n'est que lorsque cette condition est vérifiée que l'on peut instancier un concept par une information géographique.

En exploitant les exemples donnés en début de section (de 1 à 7), nous schématisons figure 5.9 une représentation du territoire sous forme d'une ontologie dans laquelle nous décomposons la couche conceptuelle (les concepts de l'ontologie) de la couche Données (les informations géographiques).

Lors de la phase d'instanciation, une grande partie des informations géographiques peuvent en effet être reliées à l'ontologie par le lien entre le qualifiant et un label d'un concept. Seule l'information géographique « Auriac au 12 juin 1876 » ne peut être reliée à l'ontologie car elle ne possède pas de qualifiant explicite apparaissant en tant que label d'un concept de l'ontologie.

Dans ce chapitre, nous avons rappelé les éléments importants de l'état de l'art afin de positionner notre approche visant à proposer une méthodologie automatisée permettant de construire une représentation d'un territoire sous forme d'une ontologie légère de domaine. Tout d'abord, l'ontologie possède les éléments nécessaires pour représenter le travail d'indexation des experts bibliothécaires. Ensuite, l'application du point de vue

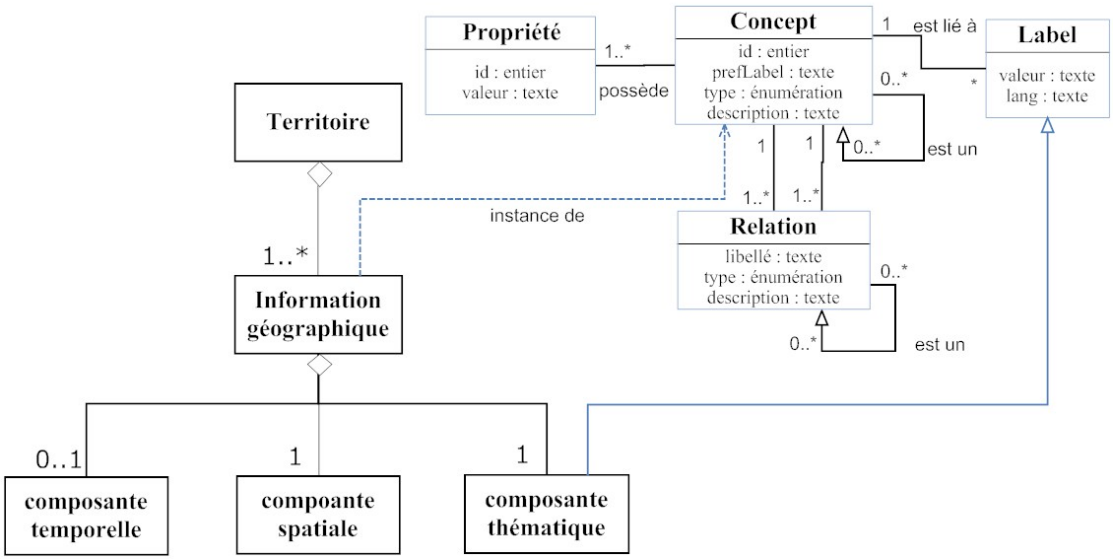


FIGURE 5.8 – Définition d'un territoire vis à vis de l'ontologie

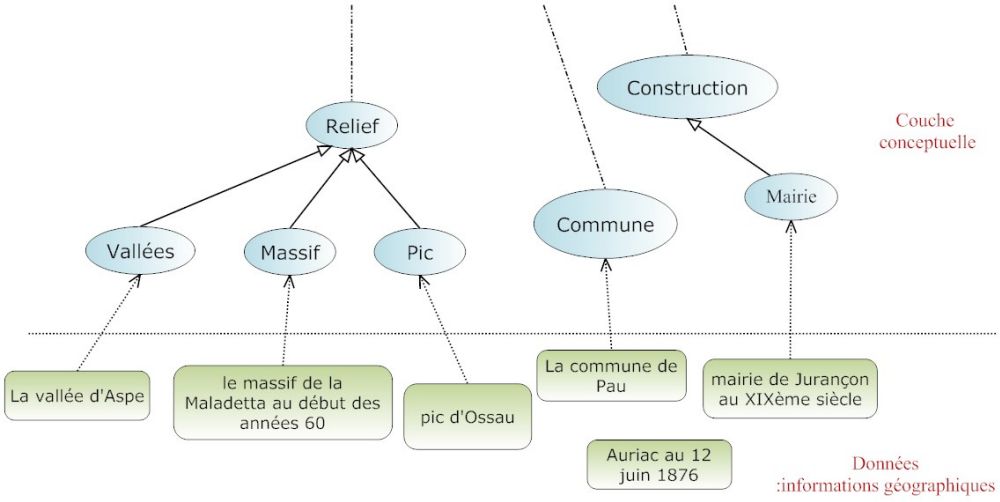


FIGURE 5.9 – Exemple schématisé d'une ontologie d'un territoire

territoire sur cette ontologie, doit nous permettre de différencier les éléments de type informations géographiques, que nous définissons sous forme d'instances, des concepts. L'ontologie donne donc la possibilité de définir des instances de concepts auxquelles nous pouvons ajouter des caractéristiques telles que leur représentation spatiale et temporelle. Nous faisons l'hypothèse que l'instanciation de l'ontologie par ces informations géographiques fait émerger un territoire. Afin de clarifier la notion de territoire, nous en

avons ensuite proposé un premier modèle adapté à notre besoin de représenter un ensemble d'informations constituant un territoire. Nous allons maintenant présenter notre méthodologie Terridoc.

Troisième partie

Contribution

Chapitre 6

Méthodologie opérationnalisée pour l'émergence d'une ontologie d'un territoire

Sommaire

6.1 Une première ontologie construite à partir de la connaissance des bibliothécaires . . . 103
6.1.1 Analyse des besoins . . . 104
6.1.2 Constitution du corpus . . . 105
6.1.3 Analyse de la connaissance experte . . . 106
6.1.4 Normalisation en réseau sémantique . . . 108
6.1.5 Formalisation du réseau sémantique . . . 120
6.1.6 Tests et Bilans . . . 121
6.2 TALN pour la représentation d'un territoire . . . 123
6.2.1 Chaîne de TAL pour l'indexation d'entités géographiques . 124
6.2.2 Application sur le contenu des notices descriptives . . . 125
6.2.3 Application sur le contenu des documents . . . 127
6.3 Enrichissement de l'ontologie à partir du contenu des documents . . . 129
6.3.1 Enrichissement de l'ontologie par des concepts . . . 130
6.4 Discussion . . . 131

Un état des travaux présenté chapitre 4.4 (*cf.* page 80) montre qu'il n'existe pas à notre connaissance une méthodologie complète et automatisée permettant de construire un modèle formel offrant une représentation sémantique d'un domaine cible, tel que le Territoire, à partir d'un ensemble de documents. Cela s'explique notamment par le fait que le territoire est une notion floue, difficile à définir et à modéliser à partir de documents textes. Nous proposons d'opérationnaliser une méthodologie, que nous nommons **TERRIDOC**, afin de construire une ontologie légère offrant une représentation

sémantique d'un territoire implicitement décrit par un fonds documentaire indexé manuellement par des experts sur la base d'un vocabulaire contrôlé. La méthodologie que nous proposons de mettre en place (*cf.* figure 6.1), s'appuie sur les travaux de [CHGM06] qui, sur la base de la méthodologie TERMINAE, permet de transformer un thésaurus d'un domaine cible en une ontologie. Le thésaurus utilisé doit permettre d'identifier un ensemble de concepts et de relations entre ces concepts.

La méthodologie que nous proposons se veut générique et applicable sur différents types de fonds documentaires indexés. Nous la décomposons en quatre étapes principales que nous présentons ci-dessous.

1. Construction d'une première ontologie d'un territoire à partir de la connaissance experte : la connaissance experte ici correspond au travail d'indexation (description du contenu de documents) réalisé par des experts bibliothécaires sur la base d'un vocabulaire contrôlé (*cf.* chapitre 2.1.2 page 28). Dans cette étape, nous appliquons les principales phases de la méthologie TERMINAE en allant de l'analyse des besoins jusqu'à la formalisation. Cependant, dans un premier temps, l'étape d'analyse linguistique préconisée dans TERMINAE est remplacée par une étape d'analyse lexicale que l'on applique sur les notices pour identifier le vocabulaire sélectionné par les experts pour décrire les documents lors de leur travail d'indexation ;
2. Chaîne de traitement linguistique pour l'identification d'entités géographiques à partir du travail d'indexation d'experts ;
3. Application de la chaîne de traitement linguistique pour l'identification d'entités géographiques à partir du contenu des documents ;
4. Enrichissement de l'ontologie minimale à partir du contenu des documents.

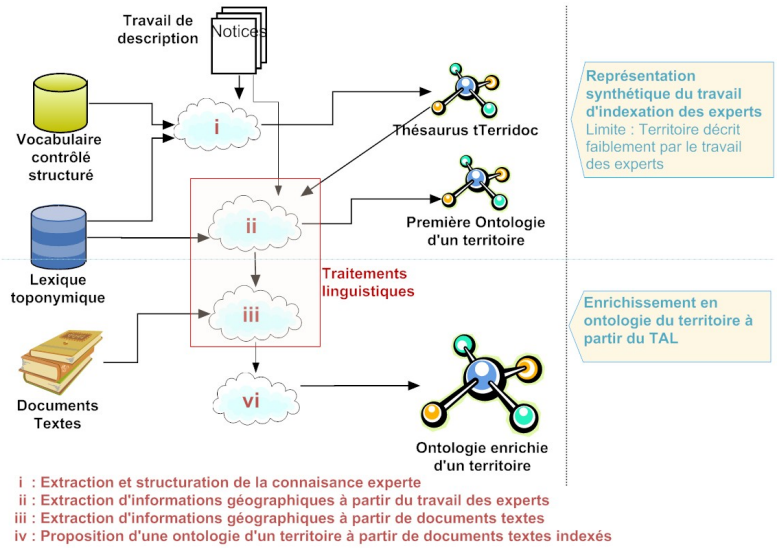


FIGURE 6.1 – Rappel de la méthodologie générale TERRIDOC pour l'émergence d'une ontologie d'un territoire

L'intérêt est de pouvoir proposer de façon incrémentale une représentation sémantique d'un territoire après chacune de ces étapes. Les experts sont ainsi à même de choisir parmi les différentes représentations celle qui correspond le mieux à leurs attentes. Après chacune des étapes, nous faisons le rapprochement avec le modèle du territoire que nous avons proposé chapitre 5.2.6 (*cf.* page 93) en mettant en avant les éléments du modèle directement concernés.

Nous allons maintenant présenter notre approche en détaillant les quatre étapes listées ci-dessus.

6.1 Une première ontologie construite à partir de la connaissance des bibliothécaires

Nous décrivons le processus de création de l'ontologie en appliquant les cinq étapes principales préconisées dans TERMINAE (*cf.* section 3.2.2.1 page 50) : (i) analyse des besoins (*cf.* section 6.1.1 page 104) ; (ii) constitution du corpus (*cf.* section 6.1.2 page 105) ; (iii) analyse linguistique (*cf.* section 6.1.3 page 106) ; (iv) normalisation en réseau sémantique (*cf.* section 6.1.4 page 108) ; (v) formalisation du réseau sémantique (*cf.* section 6.1.5 page 120).

A la différence de ce qui est présenté dans les travaux de recherche visant à construire une ontologie à partir d'une ressource structurée de type thésaurus [SLL+04, SGD04, Her05, CHGM06], nous définissons le vocabulaire contrôlé de départ sur lequel nous nous appuyons ensuite pour créer une ontologie de domaine. Rappelons que nous avons à disposition l'ensemble des notices descriptives correspondant au travail d'indexation (description du contenu des documents par des experts bibliothécaires) ainsi que le vocabulaire contrôlé utilisé pour réaliser ce travail d'indexation. L'étape d'**analyse de la connaissance experte** nous permet d'identifier et d'extraire un ensemble de termes à partir du travail d'indexation réalisé par des experts bibliothécaires sur un fonds documentaire. Ces termes sont ensuite structurés sous forme de vocabulaire contrôlé lors de l'étape de **normalisation**. Dans cette étape, les relations proviennent du vocabulaire contrôlé utilisé par les bibliothécaires pour indexer les documents. Nous obtenons alors un extrait du vocabulaire contrôlé utilisé par les bibliothécaires offrant une première représentation sémantique générale du fonds documentaire traité. Dans l'étape de **formalisation** décrite en détails dans l'implémentation, nous préconisons tout d'abord d'utiliser le langage SKOS. Il est adapté pour formaliser les structures sémantiques de type vocabulaire contrôlé et peut permettre d'intégrer le thésaurus obtenu dans un système de recherche d'informations ou encore pour échanger la connaissance avec d'autres centres documentaires.

Le domaine cible est le territoire et nous souhaitons dans nos travaux définir une structure sémantique capable d'intégrer et mettre en relation des informations afin de faire émerger une représentation d'un territoire décrit dans des documents. L'objectif est de pouvoir enrichir les termes constituant le premier thésaurus obtenu par des informations géographiques sous forme de propriétés, mais aussi d'expliciter plus précisément les relations entre ces termes par le point de vue territoire. Le thésaurus ne permettant

pas de réaliser ces enrichissements, nous proposons de transformer le thésaurus obtenu en une première ontologie légère de domaine. Cette opération est décrite dans l'étape de **normalisation**. L'ontologie offre un niveau conceptuel et des outils intéressants qui permettent notamment de prendre en compte ces propriétés géographiques et d'expliciter les relations de type associative. Nous préconisons ensuite d'utiliser le langage OWL pour formaliser l'ontologie de domaine car contrairement à SKOS, il permet de décrire plus précisément les relations entre concepts et il intègre la notion de propriétés que nous utilisons pour enrichir l'ontologie avec des informations caractéristiques d'un territoire, en l'occurrence des informations spatiales et/ou temporelles dans notre cas. En nous appuyant sur les travaux de [CHGM06], nous décomposons la transformation d'un thésaurus en une ontologie de domaine en trois actions distinctes : regroupement des termes en concepts ; structuration via les relations hiérarchiques ; structuration via les relations associatives. Cette transformation est réalisée dans l'étape de normalisation (*cf.* section 6.1.4.2 page 114) après avoir construit le thésaurus structurant le travail d'indexation des experts.

Exposons maintenant le processus de création d'une première ontologie d'un territoire en décrivant chacune des grandes étapes définies selon TERMINAE.

6.1.1 Analyse des besoins

Cette première étape vise à spécifier les besoins auxquels doit répondre l'ontologie. Les fonds documentaires que nous proposons de traiter ont la particularité d'être accompagnés de notices descriptives réalisées par des experts du domaine et contenant notamment une liste de termes choisis pour décrire le contenu des documents. Comme nous l'avons vu, ce travail dit d'indexation est réalisé sur la base d'un vocabulaire contrôlé, intégrant les contrôles de la synonymie, de l'homonymie et de la polysémie (*cf.* 2.1.2). Dans ce contexte, nous proposons de nous appuyer sur ces différentes ressources, formant un ensemble structuré de connaissances pour construire un premier thésaurus synthétisant le travail d'indexation des experts. Nous nous appuyons ensuite sur ce thésaurus que nous produisons pour construire une ontologie légère. Pour réaliser ces deux traitements, les besoins que nous identifions sont les suivants :

- *Analyse et spécification* des termes du domaine et de leurs variantes lexicales afin de les détecter dans les notices descriptives. Cela implique une analyse et des traitements particuliers liés à la structure des notices descriptives et au vocabulaire contrôlé utilisé. Dans le processus de création d'ontologie proposée, cette analyse est réalisée lors de l'analyse linguistique (*cf.* section 6.1.3 page 106) ;
- *Regroupement de ces termes* pour la création de concepts afin de déterminer les objets et notions référencés dans les documents. Ces traitements sont réalisés lors de l'étape de normalisation en réseau sémantique (*cf.* étape 4.1 section 6.1.4 page 108). Ici, seuls les termes identifiés dans les notices pour lesquels une correspondance existe dans le vocabulaire contrôlé sont sélectionnés ;
- *Structuration des concepts* en exploitant les relations présentes dans le vocabulaire contrôlé. Nous travaillons sur la base d'un thésaurus complet permettant de traiter les relations taxonomiques, associatives et termes rejetés permettant de proposer

une représentation sémantique structurée. La structuration est réalisée lors de la phase de normalisation (*cf.* sections 6.1.4.1 et 6.1.4.2). ;
– *Formalisation de l'ontologie* dans un langage interprétable par le système afin qu'il soit capable de la manipuler. Ce processus de formalisation correspond à la dernière étape de la méthodologie TERMINAE (*cf.* chapitre 6.1.5 page 120).

Les besoins étant identifiés, nous allons maintenant décrire l'étape visant à sélectionner un ensemble de ressources pour construire notre ontologie.

6.1.2 Constitution du corpus

Nous nous accordons avec [Bac00, WC03] pour dire que le choix du corpus est une étape primordiale dans la construction de l'ontologie. De façon générale (*cf.* section 3.2.2.1, page 49), un corpus représentatif du domaine ou de l'application cible facilite la construction d'une ontologie pour ce domaine ou cette application. Il doit décrire les éléments de connaissance qui seront intégrés automatiquement dans l'ontologie. Dans le cadre de nos travaux, le type de fonds documentaire que nous proposons de traiter est constitué de documents hybrides indexés. Rappelons qu'un fonds documentaire hybride est constitué de documents de type différent (image, son, texte et vidéo). Les notices descriptives attachées aux documents sont le produit d'un travail d'indexation réalisé par des experts en s'appuyant sur un vocabulaire contrôlé structuré (*cf.* chapitre 2.1, page 26). L'ensemble des notices nous permet d'identifier une liste de termes qui seront utilisés pour créer des concepts et la ressource vocabulaire contrôlé est ensuite utilisée pour structurer ces concepts.

Les figures 6.2 et 6.3 sont des extraits de notices descriptives que nous utiliserons pour expliciter les différentes étapes de notre approche visant à construire une première ontologie minimale d'un territoire.

```
Termes : Stations climatiques, thermales, etc. <sep> Barèges(Hautes-Pyrénées) <sep>
Eaux minérales <sep> Pyrénées (France) <sep> 18e siècle

Titre : Observation sur les eaux minérales de Barèges, des Eaux Chaudes et les
autres eaux minérales de Bigorre et du Béarn

Légende : Médecin du 18e siècle, T. de Bourdeu décrit le thermalisme pyrénéen

Date : 2007-04-16
```

FIGURE 6.2 – Extrait de notice descriptive 1

Dans ces extraits de notices, les champs *Titre, légende et Date* correspondent au travail de description et le champ *Termes* est le résultat de la phase d'indexation. Nous remarquons dans les éléments de description de nombreuses références à un territoire telles que « eaux minérales de Barèges, des Eaux Chaudes », « eaux minérales de Bigorre et du Béarn », ou encore « eaux minérales de Bagnères de Bigorre ». Nous remarquons également des références à un territoire dans les termes sélectionnés par les experts lors

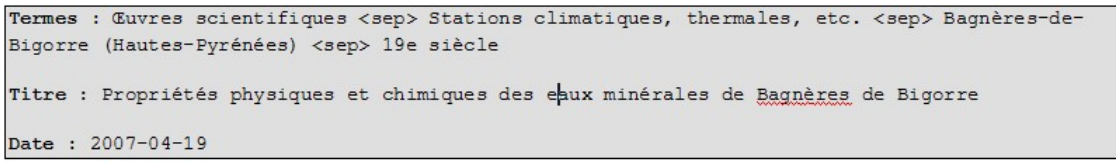


FIGURE 6.3 – Extrait de notice descriptive 2

de la phase d'indexation : « Barèges (Hautes-Pyrénées) », « Pyrénées (France) », « 18e siècle », « Bagnères de Bigorre (Hautes-Pyrénées) » et « 19e siècle ».

De façon générale, les bibliothécaires n'ont pas pour objectif d'indexer un fonds documentaire en mettant directement en avant un territoire mais dans le cas où les documents constituants le corpus ont une forte connotation géographique, nous faisons l'hypothèse que ce travail d'indexation accompagné de la ressource vocabulaire contrôlé peuvent être appréhendés seuls pour construire une première ontologie légère d'un territoire. Cependant, le corpus de documents apparait comme un complément intéressant qui peut véhiculer une quantité importante d'informations caractéristiques du domaine cible et sur lequel nous souhaitons nous appuyer pour obtenir une représentation la plus précise possible d'un territoire.

Bien que la méthodologie soit applicable à des ensembles de documents provenant de divers domaines, nous préconisons dans notre approche d'utiliser des fonds documentaires à connotation géographique.

6.1.3 Analyse de la connaissance experte

De façon générale, le contenu d'un document peut être structuré physiquement et logiquement ; dans le premier cas en utilisant les objets textuels et les marques typo-dispositionelles (gras, italique, énumérations, etc.), et dans le second cas, en annotant les unités textuelles.

De la même façon, les notices descriptives sont constituées d'un ensemble d'annotations, et leur imbrication, lorsqu'elle existe, traduit des relations hiérarchiques entre les différentes unités textuelles annotées. Dans le cas des centres documentaires de type bibliothèques ou médiathèques, la structure des notices est généralement pauvre, voire inexistante avec une liste d'annotations (*titre, auteur, legende, liste de termes descripteurs, etc.*) qui sont au premier niveau de la notice. L'analyse de la structure des notices descriptives doit nous permettre d'identifier un vocabulaire de termes qui servira de base à la construction d'un premier thésaurus. Lors de la phase d'indexation, les termes sont sélectionnés par les bibliothécaires dans un vocabulaire contrôlé, et utilisés dans les notices sous forme de liste. Dans les extraits de notices présentés figures 6.2 et 6.3, cette liste est annotée par le champ *termes* et chacun des termes est séparés par l'élément « <sep> ».

Afin d'extraire automatiquement ces termes, nous nous appuyons sur la notion de *patron structurel (PS)* mise en avant dans [KAG09] dont le rôle est de caractériser la

sémantique portée par des balises suivant leur signification, et de produire dans notre cas les fondements d'un vocabulaire contrôlé. Nous présentons dans la figure 6.4 le PS permettant d'identifier les vedettes.



FIGURE 6.4 – Forme d'un Patron Structurel

Lorsqu'une annotation *Termes* est présente dans une notice et qu'elle contient plusieurs unités textuelles notées *UTs* (de *UT1* à *UTn*, avec *n* le nombre de termes dans l'annotation *termes*) séparées par l'élément « <sep> », alors nous appliquons la fonction *EnrichListeAutorites* qui permet d'enrichir une liste d'unités textuelles donnée en paramètre par les unités textuelles présentes dans l'annotation termes d'une notice. La fonction *UTs* renvoie sous forme de liste les unités textuelles marquées par la balise fournie en paramètre et un élément séparateur, l'élément « <sep> » dans le cas présent. La définition des patrons structurels est ici dépendante du jeu d'annotations et de leur sémantique. La définition de tels patrons structurels nécessite en effet une analyse (identification de la sémantique des annotations et de la nature des relations lorsqu'il est possible d'en identifier) des documents de la part de l'expert/ontologue.

Nous obtenons en résultat de ce premier traitement un ensemble de termes que nous nommons **VocTerridoc**. Si nous reprenons les extraits de notices descriptives présentées figure 6.2 et 6.3 (*cf.* page 105), nous obtenons notamment les termes présentés figure 6.5.



FIGURE 6.5 – Exemples de concepts extraits des notices descriptives

Pour chaque unité textuelle *UT* extraite d'une notice, nous attachons dans la structure définie un lien vers le document correspondant. Les termes représentent alors le niveau conceptuel et les documents le niveau physique. Le document décrit par la notice descriptive figure 6.3 sera relié aux termes *œuvres scientifiques*, *Stations climatiques, thermales, etc.*, *Bagnères-de-Bigorre (Hautes-Pyrénées)* et *19e siècle*.

Au regard du modèle proposé figure 5.7 (*cf.* page 94) qui présente les liens entre notre modèle du territoire et le modèle simplifié d'un vocabulaire contrôlé de type thésaurus, VocTerridoc est constitué d'entités spatiales, temporelles ou thématiques qui sont éga-

lement des termes « employé pour » ou des vedettes dans le vocabulaire contrôlé utilisé lors de l'indexation. A cette étape, tous les termes de VocTerridoc sont considérés en tant qu'entités thématiques.

Cette liste de termes obtenus est une première étape vers la définition d'une structure sémantique représentant le travail des bibliothécaires ; il reste maintenant à identifier l'ensemble des relations entre ces termes pour structurer la connaissance.

6.1.4 Normalisation en réseau sémantique

L'étape suivante consiste à structurer le vocabulaire VocTerridoc en un thésaurus en nous appuyant sur la structure du vocabulaire contrôlé utilisé par les experts bibliothécaires pour réaliser leur travail d'indexation. Vis à vis du modèle présenté figure 5.7 (*cf.* page 94), nous cherchons ici à identifier les relations hiérarchiques et associatives entre vedettes. Nous nous appuyons ensuite sur ce travail de normalisation sous forme de thésaurus intégrant la connaissance experte pour construire une première ontologie d'un territoire.

6.1.4.1 Première normalisation en thésaurus

Pour faciliter la lecture, nous nommons le thésaurus que nous cherchons à construire **tTerridoc**. Parmi les différentes structures de type vocabulaire contrôlé (*cf.* chapitre 2.2 page 29), les thésaurus sont de loin les plus utilisés dans les centres documentaire pour indexer les documents. Il arrive cependant que ce soit des taxonomies mais le fait que la définition d'un thésaurus englobe celle d'une taxonomie, nous permet de proposer un traitement qui s'adapte aux deux types de ressources. Notre choix se porte donc sur le thésaurus car étant un vocabulaire contrôlé structuré, il fait partie de la même famille que la ressource utilisée par les experts pour indexer les documents et permet ainsi de représenter l'ensemble de la connaissance provenant des notices descriptives. Notons également que les ontologies sont encore peu, voire pas utilisées à l'heure actuelle dans ce domaine d'activité pour assister le travail des experts et que nous ne proposons pas de patrons structurels pour analyser les éléments caractéristiques des ontologies comme les propriétés. Enfin, notre choix est conforté par le fait que le thésaurus respecte les règles syntaxiques, terminologiques et sémantiques propres aux vocabulaires contrôlés (*cf.* chapitre 2.1.1 page 26) qui faciliteront ensuite le passage en une ontologie.

Nous présentons figure 6.6 un extrait d'une ressource thésaurus respectant cette définition pour le terme « *Stations climatiques, thermales, etc.* », extrait que nous utiliserons par la suite pour exemplifier les différentes étapes.

Cette exemple respecte la définition globale d'un thésaurus (*cf.* figure 5.7 page 94) dans lequel des vedettes, attachées à un ou plusieurs terme(s) employé(s) pour, sont reliés entre eux par des relations hiérarchiques et associatives (élément relation du modèle).

L'interprétation des termes marqués et des relations donne lieu à des fragments de thésaurus. Dans l'exemple ci-dessus, « *Stations climatiques, thermales, etc.* » est une vedette dans la ressource thésaurus, qui a notamment pour terme générique « *Lieux de villégiature* », pour terme spécifique « *Stations d'été* », et pour terme employé pour



FIGURE 6.6 – Extrait de thésaurus avec le terme « Stations climatiques, thermales, etc. »

« *villes thermales* ». Les différentes règles d'interprétation applicables sur un thésaurus donné, quel que que soit le formalisme utilisé (XML, TXT, etc.), sont formalisées à l'aide de patrons structurels (PS). Nous présentons ci-dessous les cinq patrons définis dans le processus de structuration du vocabulaire VocTerridoc en thésaurus tTerridoc : (1) gestion des termes vedettes ; (2) gestion des termes rejetés ; (3) gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc ; (4) gestion des relations associatives entre termes du vocabulaire vocTerridoc ; (5) Structuration du thésaurus via les relations génériques.

1. **Gestion des termes vedettes** : Le patron spécifié dans le tableau 6.1 indique que pour chaque terme du vocabulaire vocTerridoc, si *t1* est une vedette dans la ressource thésaurus, une vedette *a1* est créée dans le thésaurus tTerridoc. *a1* a pour nom *t1* et l'ensemble des informations attachées à la vedette *a1* (définition, etc.) dans la ressource thésaurus sont attachées à *a1* dans tTerridoc.

Conditions	- t1 terme vedette
Actions	- Création vedette a1

TABLE 6.1 – Gestion des termes vedettes

Si nous reprenons l'extrait de la ressource thésaurus présenté figure 6.6, le terme « *Stations climatiques, thermales, etc.* » présent dans le vocabulaire vocTerridoc est une vedette dans la ressource thésaurus, ce qui nous permet de créer une vedette de même nom dans le thésaurus tTerridoc. Les vedettes correspondent à l'élément « vedette » dans le modèle présenté 5.7 (*cf.* page 94). A cette étape, les entités spatiales, temporelles et thématiques sont traitées et modélisées de la même façon.

2. **Gestion des termes rejetés** : Le patron qui suit (*cf.* tableau 6.2) permet de traiter les termes du vocabulaire vocTerridoc qui ne sont pas vedettes dans la ressource thésaurus. Si *t1* n'est pas une vedette, alors, si *t1* est relié à une vedette *a1* dans la ressource thésaurus par la relation « employé pour », la vedette *a1* est créée dans le thésaurus tTerridoc si elle n'existe pas déjà. Le terme *t1* est alors ajouté tel quel à la structure tTerridoc et relié par la relation « employé pour » à *a1*.

Conditions	- t1 est terme employé pour a1
Actions	- Création vedette a1
	- Création terme t1
	- Création relation « employé pour » liant t1 et a1

TABLE 6.2 – Gestion des termes rejetés

Prenons l'exemple du terme « *Stations de cures* » qui est un terme rejeté dans l'extrait de la ressource thésaurus présenté figure 6.6, si ce terme est présent dans le vocabulaire vocTerridoc, alors la vedette reliée à ce terme, en l'occurrence « *Stations climatiques, thermales, etc.* » est créée dans le thésaurus tTerridoc si elle n'existe pas déjà. Le terme « *Stations de cures* » est alors ajouté tel quel et une relation de type « employé pour » est ajouté entre la vedette « *Stations climatiques, thermales, etc.* » et le terme « *Stations de cures* ». Qu'ils correspondent à des entités spatiales, temporelles ou thématiques, les termes rejetés identifiés dans cette étape correspondent à l'élément « terme employé pour » décrivant des thèmes (vedettes) dans le modèle présenté figure 5.7 (*cf.* page 94).

3. **Gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc** : Ce patron (*cf.* tableau 6.3) permet d'enrichir la structure avec les relations hiérarchiques lorsque deux termes du vocabulaire vocTerridoc existent dans la ressource thésaurus en tant que vedette et qu'ils sont liés par une relation hiérarchique.

   Soit *a2* un terme du vocabulaire vocTerridoc qui est vedette dans la ressource thésaurus, une vedette *a2* est créée dans tTerridoc si elle n'existe pas déjà. Pour chaque terme *a1* de tTerridoc, si *a2* est générique à *a1* dans la ressource thesaurus, alors une relation de type « générique » est créée entre *a2* et *a1* dans tTerridoc signifiant qu'*a2 est générique à a1*. De la même façon, si *a2* est spécifique à *a1* dans la ressource thésaurus, alors une relation de type « spécifique » est créée entre *a2* et *a1* dans tTerridoc signifiant qu'*a2 est spécifique à a1*.

   Prenons l'exemple des termes « *Barèges (Hautes-Pyrénées)* » et « *Stations climatiques, thermales, etc.* », chacun des termes étant une vedette dans la ressource thésaurus, deux vedettes sont créées dans tTerridoc si elles n'existent pas déjà. Dans la ressource thesaurus, une relation hiérarchique est identifiée entre ces deux vedettes indiquant que « *Barèges (Hautes-Pyrénées)* » est un terme spécifique à « *Stations climatiques, thermales, etc.* », ce qui nous permet de créer une relation hiérarchique dans tTerridoc indiquant que « *Barèges (Hautes-Pyrénées)* » est un

Conditions	- a1 et a2 deux termes vedettes distincts
	- a2 terme générique/spécifique à a1
Actions	- Création vedette a2
	- Création relation hiérarchique liant a1 et a2

TABLE 6.3 – Gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc

terme spécifique à « *Stations climatiques, thermales, etc.* ». les relations hiérarchiques correspondent à l'élément « sous-thème » dans le modèle présenté figure 5.7 (*cf.* page 94).

4. **Gestion des relations associatives entre termes du vocabulaire vocTerridoc** : Ce patron (*cf.* tableau 6.4) permet d'enrichir la structure avec les relations associatives lorsque deux termes du vocabulaires vocTerridoc existent dans la ressource thésaurus en tant que vedettes et qu'ils sont liés par une relation de type « terme associé ».

   Soit *t1* et *t2* deux termes du vocabulaire vocTerridoc qui sont vedettes dans la ressource thesaurus, deux vedettes *a1* et *a2* sont créées dans tTerridoc si elles n'existent pas déjà. Si *t1* est relié à *t2* dans la ressource thesaurus par la relation de type associative, alors une relation « terme associé » est créée entre *a1* et *a2* dans tTerridoc.

Conditions	- t1, t2 deux termes vedettes
	- t1 terme associé à t2
Actions	- Création vedettes a1 et a2
	- Création relation « terme associé » entre a1 et a2

TABLE 6.4 – Gestion des relations associatives entre termes du vocabulaire vocTerridoc

Prenons l'exemple des termes « *Eaux minérales* » et « *Stations climatiques, thermales, etc.* », chacun des termes étant une vedette dans la ressource thésaurus, deux vedettes sont créées dans tTerridoc si elles n'existent pas déjà. Dans la ressource thesaurus, une relation associative est identifiée entre ces deux vedettes indiquant que « *Eaux minérales* » est un terme associé à « *Stations climatiques, thermales, etc.* », ce qui nous permet de créer une relation associative dans tTerridoc indiquant que « *Eaux minérales* » est un terme associé à « *Stations climatiques, thermales, etc.* ». Ces relations associatives entre deux vedettes correspondent à l'élément « relation » dans le modèle présenté figure 5.7 page 94. Une vedette peut être reliée à une ou plusieurs autre vedette(s) et une relation est définie entre deux concepts uniquement.

En exploitant le vocabulaire vocTerridoc et la structure de la ressource thésaurus, nous définissons une première version du thésaurus tTerridoc intégrant des vedettes, des termes rejetés, des relations hiérarchiques et associatives entre les termes provenant de vocTerridoc lorsqu'elles sont identifiées dans la ressource thé-

saurus. La figure 6.7 schématise le résultat de l'étape de structuration à partir de l'extrait de termes identifié dans les notices présentées figures 6.2 et 6.3.

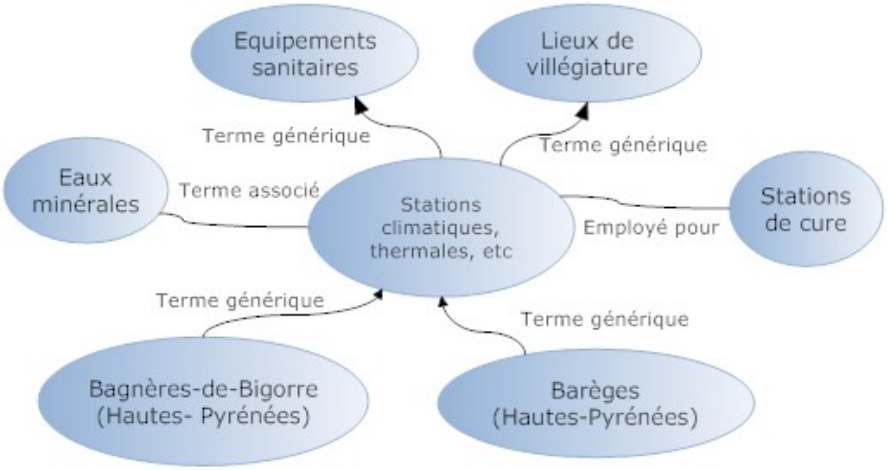


FIGURE 6.7 – Extrait de la structuration en thésaurus à titre d'exemple

Nous obtenons en appliquant l'ensemble des patrons présentés un ensemble de petites structures sémantiques constituées des termes utilisés par les experts pour décrire le contenu des documents. Les règles présentées sont appliquées sur les termes constituant le vocabulaire vocTerridoc qui peuvent décrire des thématiques très diverses et être distants (sémantiquement) dans la ressource thésaurus utilisée pour réaliser le travail d'indexation. Nous proposons d'enrichir cette première représentation par l'ensemble des termes présents dans la ressource thésaurus qui sont des termes génériques aux termes constituant le vocabulaire vocTerridoc.

5. **Structuration du thésaurus via les relations génériques** : Nous considérons chaque terme du vocabulaire comme concept de bas niveau car rattaché directement à des documents et nous enrichissons le thésaurus tTerridoc avec les concepts plus génériques de la ressource thésaurus en ajoutant dès que nécessaire les relations de type « générique » liées. Le but visé par l'enrichissement du thésaurus via ces termes génériques est de permettre le regroupement en une seule structure des termes extraits.

   Le PS ci-dessous (Figure 6.8) permet de construire de façon récursive une structure de plusieurs niveaux hiérarchiques en partant des termes constituant le premier thésaurus tTerridoc et en exploitant les relations de généricité entre vedettes lorsqu'elles existent dans la ressource thésaurus.

   Concernant ce PS, chaque vedette *a1* du thésaurus tTerridoc est recherchée dans la ressource thésaurus via la fonction *equals*. Lorsque la vedette est identifiée dans la ressource thésaurus, pour chacune des vedettes génériques *a2* qui lui sont associés par la relation générique dans la ressource thésaurus, la vedette *a2* est alors créée dans tTerridoc et la relation générique *r1* est ensuite ajoutée entre *a1* et *a2* dans

```
ressourceThesaurus : vocabulaire contrôlé utilisé pour l'indexation
tTerridoc : thésaurus résultant de l'approche
Début
    Pour chaque autorité aT de tTerridoc faire
        structurationHierarchique(aT, tTerridoc, ressourceThesaurus)
    Fin Pour
 Fin
Fonction structurationHierarchique(autorite, tTerridoc, ressourceThesaurus)
   Pour chaque autorité aVC de ressourceThesaurus faire
      Si equals (autorite, aVC) alors
          Pour chaque terme générique de aVC faire
                pere = terme générique courant
                creationAutorite(pere, tTerridoc)
                creationrGenerique(pere, autorite, tTerridoc)
                structurationHierarchique(pere, tTerridoc, ressourceThesaurus)
        Fin Pour
      Fin Si
   Fin Pour
Fin fonction
```

FIGURE 6.8 – Patron Structurel permettant d'enrichir la structure de tTerridoc à partir des relations génériques

le thésaurus tTerridoc.

Prenons l'exemple d'un document dont le thème principal annoté par l'expert est indiqué dans l'extrait d'une troisième notice présentée figure 6.9. Le traitement proposé ci-dessus permet d'obtenir une structure sémantique enrichie (Figure 6.10).

```
Termes : Ski-alpinisme <sep> Barèges (Hautes-Pyrénées) <sep> 18e siècle
```

FIGURE 6.9 – Extrait de notice descriptive 3

Le lien défini explicitement entre le terme *Ski-alpinisme* et le groupe de termes présenté figure 6.7 permet ici de les regrouper en une structure unique. A noter ici que de la richesse de la structure de la ressource thésaurus dépend la qualité et la complétude de la structure tTerridoc.

Dans cette étape, nous définissons un thésaurus tTerridoc en nous appuyant sur les notices descriptives attachées aux documents constituant le fonds documentaire ainsi que sur la ressource de type vocabulaire contrôlée utilisée pour indexer ces documents. tTerridoc respecte la structure d'un thésaurus définie figure 5.7 (*cf.* page 94) dans lequel des vedettes, correspondant à des entités spatiales, temporelles ou thématiques, sont structurées. Afin de pouvoir intégrer cette représentation sémantique dans un système à base de connaissances, il est nécessaire de passer par une étape de formalisation du thésaurus dans un langage compréhensible par ce système, étape que nous allons détailler maintenant.

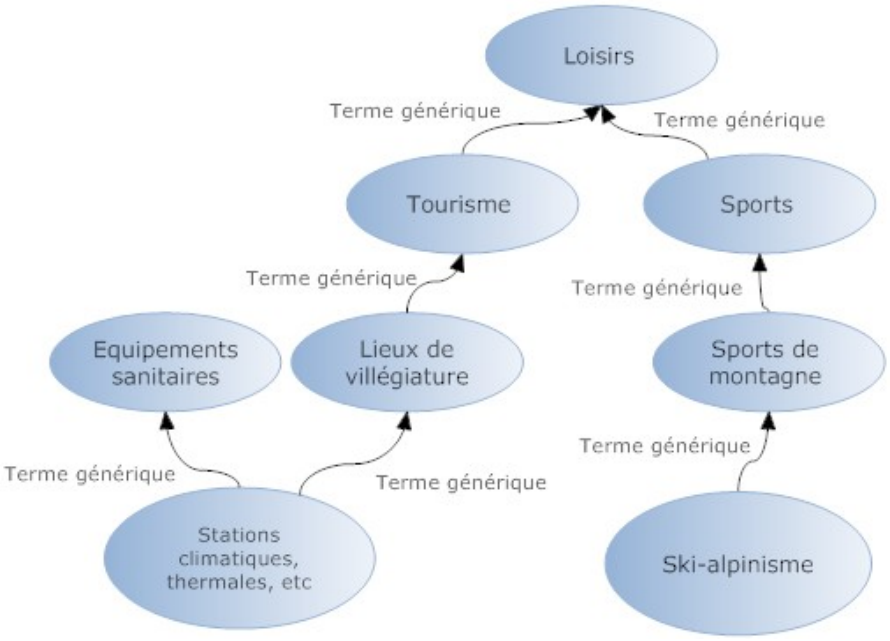


FIGURE 6.10 – Enrichissement du vocabulaire identifié par les termes « génériques »

6.1.4.2 Normalisation en une première ontologie d'un territoire

La construction d'une ontologie à partir d'une ressource de type vocabulaire contrôlé implique une phase de normalisation des éléments constituants que sont les concepts, les relations et les propriétés (*cf.* figure 5.8 page 96 définissant une ontologie de façon succincte et dans laquelle nous présentons les liens avec le modèle du territoire). Nous la déclinons en quatre actions : (1) regroupements des termes en concepts ; (2) structuration via les relations hiérarchiques ; (3) structuration via les relations associatives ; (4) instanciation de l'ontologie pour la description un territoire. Décrivons maintenant étape par étape la transformation du thésaurus tTerridoc en ontologie. Pour chacune de ces étapes, nous faisons le lien avec le modèle du territoire que nous avons proposé chapitre 5.6 (*cf.* page 5.6) en mettant en avant les éléments du modèle qui sont concernés dans l'étape de création de l'ontologie.

1. **Regroupement des termes en concepts** : En nous basant sur le thésaurus tTerridoc, nous proposons un mécanisme automatique de regroupement de différents termes, ayant le même sens dans le thésaurus tTerridoc, en un seul concept. Pour effectuer ce regroupement, nous traitons les termes reliés par la relation « employé pour » indiquant qu'un ensemble de termes portant le même sens dans le thésaurus tTerridoc est représenté par un terme vedette du thésaurus. La règle spécifiée dans le tableau 6.5 indique que pour chaque vedette *a1* du thésaurus tTerridoc, un concept *c1* est créé portant pour label *a1* et l'ensemble des informations attachées à la vedette *a1* (définition, etc.) sont formalisées sous forme de propriétés attachées

au concept *c1*. Pour chacun des termes *t2* reliés par la relation « employé pour » à *a1* dans tTerridoc, un label *t2* est ajouté au concept *c1*.

Conditions	- a1 employé pour t2
	- a2 terme vedette
Actions	- Création concept c1 avec label a1 et t2
	- Ajout propriétés a1 (définition, commentaires, etc.)

TABLE 6.5 – Regroupement des termes en concepts

Nous avons vu section 2.2.1.3 (*cf.* page 30) que le thésaurus fait partie de la famille des vocabulaires contrôlés et que par définition, il ne peut pas contenir de termes utilisés dans plusieurs groupes de termes décrivant chacun un sens spécifique. La polysémie est aussi traitée et lorsqu'un terme est polysémique, le sens souhaité doit être explicité en le spécifiant par exemple à chaque fois entre parenthèse (ex : *Grue (appareil)* et *Grue (animal)*). Il n'est donc pas nécessaire d'effectuer des traitements supplémentaires pour traiter la polysémie.

Aussi, les thésaurus dans leur définition font régulièrement la différence entre un même terme au singulier et au pluriel (le singulier représentant le sens du terme et le pluriel une catégorie d'éléments rapportant à ce terme). Dans les ontologies, un concept décrit un sens et doit être au singulier. Lors de l'identification et de la formalisation des concepts, nous intégrons une phase de lemmatisation qui va permettre de faire des regroupements par unités de sens. Lorsqu'un terme est présent au pluriel ainsi qu'au singulier dans le thésaurus, un concept unique est défini dans l'ontologie regroupant les caractéristiques et propriétés des deux termes provenant du thésaurus. Les termes décrivant un sens sont notamment accompagnés d'une définition succincte ainsi que d'une catégorie (Nom Commun, Auteur, Titre, etc.). Ces propriétés sont ajoutées au concept lors de sa création.

Le travail de regroupement des termes en concepts est réalisé en respectant un ensemble de critères reconnus pour construire une ongologie (*cf.* voir section 3.1 page 41). Le fait de nous appuyer sur un vocabulaire contrôlé utilisé dans les centres documentaires de type bibliothèque et médiathèque nous permet de garantir tout d'abord une certaine clarté dans la définition des concepts. Les termes du vocabulaire contrôlé sont définis par des experts et ils sont toujours accompagnés d'une définition et régulièrement de règles d'utilisation. Dans un vocabulaire contrôlé, la cohérence est garantie par le fait qu'un terme n'a qu'un seul sens et que ce terme est le seul à avoir ce sens (*a priori* pas de contradiction entre les définitions). Aussi, nous proposons de respecter autant que possible le critère de normalisation des noms en nous appuyant sur le vocabulaire contrôlé et en regroupant également un terme présent deux fois, au pluriel et au singulier, en un même concept. Enfin nous faisons l'hypothèse que le vocabulaire sélectionné par des experts, pour décrire le contenu des documents, est une base de travail adaptée pour réprésenter un territoire implicitement décrit dans les documents et nous garantie ainsi un engagement ontologique minimal.

Pour faire le lien avec le modèle présenté figure 5.8 (*cf.* page 96), les concepts se rapportent à l'élément « concept » du modèle, les termes « employé pour » sont des labels dans l'ontologie (se rapportant à l'élément « label »), et les propriétés à l'élément « property ».

D'après l'extrait du thésaurus tTerridoc figure 6.7 (*cf.* page 112), nous remarquons que le terme « *Stations climatiques, thermales, etc.* » est vedette et le terme « *stations de cure* » est relié à cette vedette par la relation « employé pour ». En appliquant la règle énoncée ci-dessus qui permet de regrouper des termes en concept, un concept est créé avec pour identifiant le nom du terme (au singulier) représentant le groupe de terme (ici « *Station_climatique,_thermale,_etc.* »). Le terme vedette « *Stations climatiques, thermales, etc.* » ainsi que le terme rejeté « *stations de cure* » sont définis en tant que label de ce concept (au singulier). Le fait que l'identifiant du concept porte le nom de la vedette permet de garder le lien entre le terme du thésaurus tTerridoc et l'ontologie pour de possibles actions futures.

La relation de type « employé pour » étant traitée pour regrouper les termes en concepts, nous allons maintenant prendre en compte les relations taxonomiques et associatives entres termes pour proposer une première structuration en ontologie.

2. **Structuration via les relations hiérarchiques** : En reprenant la structure du thésaurus tTerridoc, les relations hiérarchiques entre termes, devenus labels d'un concept dans l'ontologie lors de la phase de regroupement, sont retenues comme relations candidates pour représenter des relations « sous classes ». Nous spécifions tableau 6.6 la règle exposant que pour chaque vedette *a2* reliée à une autre vedette a1 par la relation « terme générique », un concept *c1*, s'il n'existe pas déjà, est créé portant pour label *a1* et l'ensemble des informations attachées à la vedette *a1* (définition, etc.) sont formalisées sous forme de propriétés attachées au concept *c1*. De la même façon, un concept *c2* est créé portant pour label *a2*. Une relation de type « sous classe » est alors créée entre le concept *c2* et *c1* pour indiquer que *c1 est « sous classe » de c2*.

Conditions	- a1 et a2 deux termes vedettes a2 terme générique à a1
Actions	- Création concept C1 ayant pour label a1 - Création concept C2 ayant pour label a2 - Ajout lien « sous classe » entre C1 et C2

TABLE 6.6 – Structuration via les relations hiérarchiques

D'après l'extrait du thésaurus tTerridoc figure 6.7 (*cf.* page 112), nous remarquons que les termes « *Stations climatiques, thermales, etc.* » et « *Lieux de villégiature* » sont vedettes de la ressource thésaurus et que ces deux vedettes sont liées par une relation de type hiérarchique indiquant que « *Lieux de villégiature* » est générique à « *Stations climatiques, thermales, etc.* ». La règle énoncée ci-dessus qui permet de structurer l'ontologie en exploitant les relations hiérarchiques du thésaurus implique de créer deux concepts s'ils n'existent pas au préalable

dans l'ontologie, l'un avec pour identifiant « *Station_climatique,_thermale,_etc.* » et pour label le terme « *Station climatique, thermale, etc.* » et l'autre avec pour identifiant « *Lieu_de_villégiature* » et pour label « *Lieu de villégiature* ». Une relation hiérarchique est ensuite créée dans l'ontologie pour indiquer que *« Station_climatique,_thermale,_etc. » est sous classe de « Lieu_de_villégiature »* .

De façon générale, les relations hiérarchiques dans les thésaurus peuvent engendrer des redondances car ces structures sémantiques sont limitées au niveau de la formalisation. La relation de généricité est une relation transitive et permet le type d'inférence suivant : si *C1 « est une sous classe de » C2* et *C2 « est une sous classe de » C3*, alors *C1 « est une sous classe de » C3*, *C1*, *C2* et *C3* étant des concepts. Nous exploitons cette propriété pour supprimer ces redondances présentes dans le thésaurus. En reprenant l'exemple précédent, si *C1 « est une sous classe de » C3*, cette relation sera supprimée car par inférence, *C1 « est une sous classe de » C3* en passant par *C2*. Les relations hiérarchiques correspondent à l'élément « est un » reliant deux concepts dans le modèle présenté figure 5.8 (c.f. page 96).

Aussi, nous avons vu que les relations hiérarchiques incluent la relation générique (genre-espèce), la relation partitive (tout-partie), la relation d'instance et les relations poly-hiérarchiques. La définition de ces relations « terme plus spécifique », « terme plus générique » est orientée par l'utilisation faite des thésaurus, c'est-à-dire l'aide au travail du documentaliste (indexation, recherche), et non par la formalisation de la connaissance du domaine. Nous proposons (section 4 page 118) une première méthode de désambiguïsation partielle en nous appuyant sur le domaine d'application qu'est le territoire.

Dans cette étape de structuration via les relations hiérarchiques, nous nous accordons avec [AGPTP98] qui préconise la diversification des hiérarchies lors de la création d'une ontologie mais nous ne proposons pas de traitement particulier pour garantir le critère de distance sémantique minimale.

Nous cherchons ensuite à identifier et ajouter des relations associatives (de type « associé à ») entre concepts de l'ontologie.

3. **Structuration via les relations associatives** : Cette étape consiste à définir dans l'ontologie les relations associatives provenant du thésaurus tTerridoc. Pour chaque relation associative entre deux termes du thésaurus, nous définissons une relation associative dans l'ontologie entre les concepts dont ils sont labels. Nous spécifions tableau 6.7 la règle exposant que pour chaque vedette a1 reliée à une autre vedette a2 par la relation « terme associé », un concept *c1*, s'il n'existe pas déjà, est créé portant pour label *a1* et l'ensemble des informations attachées à la vedette *a1* (définition, etc.) sont formalisées sous forme de propriétés attachées au concept *c1*. De la même façon, un concept c2 est créé portant pour label *a2*. Une relation de type « associé à » est alors créée entre le concept *c2* et *c1* pour indiquer que *c1* est lié sémantiquement à *c2* sans que l'on puisse pour le moment qualifier plus précisément ce lien.

   D'après l'extrait du thésaurus tTerridoc figure 6.7 (*cf.* page 112), nous remarquons que les termes « *Stations climatiques, thermales, etc.* » et « *Eaux minérales* » sont

Conditions	- a1 et a2 deux termes vedettes
	- a1 terme associé à a2
Actions	- Création concept C1 avec pour label a1
	- Création concept C2 avec pour label a2
	- Ajout relation « terme associé » entre C1 et C2

TABLE 6.7 – Structuration via les relations associatives

vedettes de la ressource thésaurus et que ces deux vedettes sont liées par une relation de type associative indiquant que « *Stations climatiques, thermales, etc.* » est lié sémantiquement à « *Eaux minérales* ». L'application de la règle énoncée ci-dessus permet de structurer l'ontologie en exploitant les relations associatives du thesaurus. Deux concepts sont tout d'abord créés s'ils n'existent pas au préalable dans l'ontologie, l'un avec pour identifiant « *Station_climatique,_thermale,_etc.* » et pour label le terme « *Station climatique, thermale, etc.* » et l'autre avec pour identifiant « *eau_minérale* » et pour label « *Eau minérale* ». Une relation associative est ensuite créée dans l'ontologie pour indiquer que « *Station_climatique,_thermale,_etc.* » *est associé à « Lieu_de_villégiature »* .

Dans cette étape, les relations associatives sont définies via l'élément « relation » et il est possible d'en préciser le sens (*cf.* figure 5.8 page 96).

Discussion et mise en relation avec le modèle du territoire Nous obtenons à cette étape une première ontologie légère offrant une représentation sémantique du fonds documentaire traité. Le thésaurus tTerridoc est utilisé pour créer des concepts, propriétés attachées à ces concepts, des labels, des relations de type « est un » ainsi que des relations associatives entre concepts *cf.* figure 5.8 page 96).

Parmi les critères reconnus pour construire une ontologie (*cf.* voir section 3.1 page 41), nous prenons en compte les critères de normalisation des noms, de clarté et de complétude dans la définition des concepts. Nous tenons compte également du critère d'engagement ontologique minimal et de la diversification des hiérarchies. Nous ne proposons pas de traitement particuliers pour garantir les critères de modularité.

L'ontologie obtenue à cette étape ne met pas en avant les spécificités territoriales présentes dans le fonds documentaire, que ce soit dans les documents ou dans les notices attachées. Nous décrivons dans le paragraphe suivant notre démarche automatisée permettant d'identifier des éléments décrivant un territoire et de les attacher ensuite à l'ontologie légère que nous venons de générer.

4. **Instanciation de l'ontologie pour la description un territoire** : Nos travaux visent à construire une ontologie d'un territoire dans laquelle nous souhaitons identifier et représenter un certain nombre de thèmes, normalisés sous forme de concepts, dans un espace géographique ainsi que dans un espace temporel. Dans ce modèle, les entités géographiques se rapportant à des lieux ou à des périodes qui sont reliées par une relation hiérarchique à un concept sont redéfinies comme des

instances de ce concept. En nous appuyant sur une ressource de type gazetteers, un premier travail consiste à identifier dans l'ontologie obtenue dans l'étape précédente (*cf.* section 3 page 117) l'ensemble des concepts qui sont des entités spatiales. La ressource gazetteers apporte des informations sur les entités spatiales identifiées et nous permet notamment de connaître leur(s) représentation(s) spatiale(s) que nous pouvons stocker dans notre modèle. Cette étape permet par la même occasion de désambiguïser une partie des relations hiérarchiques provenant du thésaurus.

Nous spécifions tableau 6.8 la règle exposant que pour chaque concept *C1* avec pour label *l1* qui correspond à un nom_toponymique, le concept *C1* est supprimé. Pour chaque concept *C2* relié à *C1* par la relation hiérarchique indiquant que *C1 est « sous classe » de C2*, une instance *I1* est créée et la relation de type « sous classe de » dans laquelle le concept *C1* intervenait est redéfinie en relation de type « instance de » indiquant que *C1 est une « instance de » C2*. L'instanciation réalisée met en avant une information géographique dans laquelle un label de C2 représente la composante thématique et l'instance décrit la composante spatiale. Une propriété *geometrie* est ajoutée au concept *C2* afin d'indiquer la représentation spatiale pour chacune de ses instances. Les autres concepts restent inchangés. En

Conditions	- l1 label de concept C1
	- l1 validé comme un nom_toponymique
	- C1 « sous classe » de C2
Actions	- Suppression C1
	- Création I1 instance de C2
	- Ajout propriétés spatiales à I1

TABLE 6.8 – Instanciation de l'ontologie pour la description un territoire

reprenant l'extrait présenté figure 6.7 (*cf.* page 112), les concepts « *Bagnères-de-Bigorre_(Hautes-Pyrénées)* » et « *Barèges_(Hautes-Pyrénées)* », tous deux « sous classes » du concept « *Station_climatique,_thermale,_etc.* » sont présents dans la ressource gazetteers, ce qui nous permet de les identifier comme entités spatiales. Les concepts « *Bagnères-de-Bigorre_(Hautes-Pyrénées)* » et « Barèges_(Hautes-Pyrénées) » sont alors supprimés et remplacés par des instances du concept « *Station_climatique,_thermale,_etc.* », chacune des instances ayant pour identifiant le nom de l'information géographique concernée, soit le concept père concaténé à celui du concept correspondant. Pour « Barèges_(Hautes-Pyrénées) » par exemple, l'instance créée a pour identifiant « Station_climatique,_thermale,_etc._de-_Barèges_(Hautes-Pyrénées) ». Une propriété *geometrie* est ajoutée au concept « *Station_climatique,-_thermale,_etc.* » afin d'indiquer la représentation spatiale pour chacune de ses instances.

Le fait d'identifier ces deux éléments comme des entités spatiales nous permet de désambigüiser les relations de type générique avec les concepts « *eaux_minérales* » et « *stations_climatiques, _thermales,_etc.* » en la précisant en relation d'instance : « instance_de » dans notre première ontologie du territoire, signifiant ainsi que

« *Barèges_(Hautes-Pyrénées)* » par exemple est une instance de « *stations_climatiques,_thermales,_etc.* » (figure 6.11).

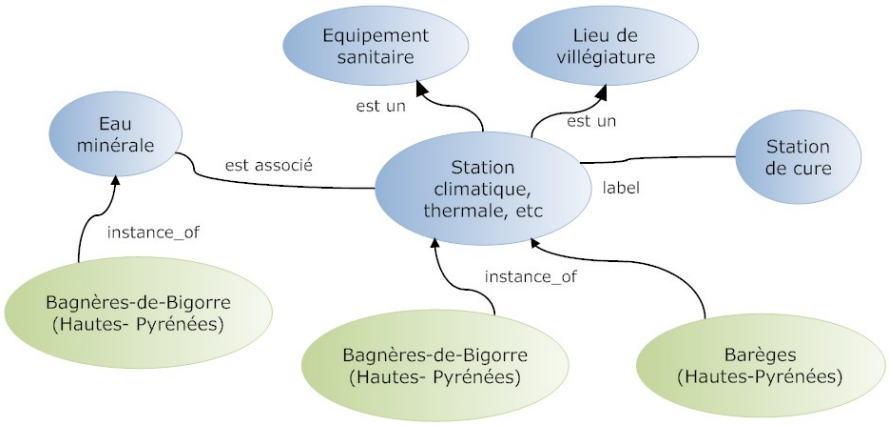


FIGURE 6.11 – Extrait de l'ontologie générée

Nous obtenons après cette étape d'instanciation une première ontologie d'un territoire à partir du travail d'indexation d'experts.

6.1.5 Formalisation du réseau sémantique

L'étape de formalisation nécessite de transcrire dans un langage formel les connaissances structurées obtenues pour les rendre interprétables par la machine. Comme le montre [Her05], la disponibilité de ressources du type vocabulaire contrôlé sous un format normalisé est un enjeu important dans le domaine de l'accès à l'information. Parmi les différents langages standards permettant de formaliser cette information pour en faciliter l'accès (XML, standards W3C RDF, OWL, SKOS, etc.), SKOS (Simple Knowledge Organisation System ou Système simple d'organisation des connaissances) est un langage de représentation de schémas de concepts réellement adapté pour formaliser la famille des vocabulaires contrôlés de type taxonomies et thésaurus. Nous préconisons dans un premier temps de formaliser le thésaurus tTerridoc en SKOS. Ce choix est le résultat de nombreuses discussions avec les experts bibliothécaires, qui au fur et à mesure des avancements de nos travaux, ont émis le besoin d'obtenir une vue globale de leur travail d'indexation. Et cette base de connaissances est le résultat d'une étape intermédiaire de notre méthodologie offrant une représentation sémantique synthétique et formalisée du travail d'indexation des experts. Le formalisme SKOS peut alors permettre d'intégrer tTerridoc à un SBC proposant des outils de recherche ou il peut aussi permettre des échanges entre centres documentaires.

6.1.5.1 Formalisation du thésaurus tTerridoc

Comme son nom l'indique, SKOS est destiné à proposer un système permettant d'exprimer et de gérer des modèles interprétables par des machines dans la perspective

du web sémantique et permet donc la représentation du thésaurus que nous avons défini. On le dit « simple » par opposition notamment au langage OWL, qui est plus à même de représenter des structures plus riches : les ontologies. Son formalisme de représentation repose sur les graphes RDF. Le concept constitue le centre du graphe auquel peuvent notamment être attachés en tant que propriétés RDF :

- les indications portant sur le concept lui même : des termes préférentiels (*skos :prefLabel*) ou alternatifs (*altLabel*), les équivalents dans d'autres langues ; des termes cachés ; et la représentation par une image ;
- les différents types de notes : notes de définition et d'application, exemples, notes historiques, etc. Nous utilisons l'élément *skos :scopeNote* défini dans Skos pour ajouter les notes ;
- les relations sémantiques : hiérarchie (*skos :broader*) et association (*skos :related*).

Nous obtenons donc une première représentation sémantique structurée sous forme de thésaurus tTerridoc que nous formalisons en SKOS. Cette représentation sémantique offre une vue globale du travail de description réalisé par les experts sur un fonds documentaire. Les documents sur lesquels nous nous appuyons renferment de nombreuses références au territoire, et le travail d'indexation est donc une source de connaissance intéressante que nous exploitons pour proposer une première représentation d'un territoire.

6.1.5.2 Formalisation de l'ontologie

Nous préconisons d'utiliser OWL (en version OWL-Lite) pour formaliser l'ontologie car il offre plus de vocabulaire pour décrire les propriétés et les classes. On peut citer à titre d'exemple : les relations entre classes (par exemple la disjonction), les cardinalités (par exemple « exactement un »), l'égalité, le typage plus riche des propriétés, les caractéristiques des propriétés (par exemple la symétrie) et les classes énumérées.

6.1.6 Tests et Bilans

6.1.6.1 Tests

Nous décrivons dans le chapitre implémentation nos expérimentations permettant d'appliquer notre méthodologie sur un extrait d'un fonds documentaire hybride constitué de 900 documents et notices descriptives attachées se rapportant au territoire des Pyrénées entre le XVIIIième et le XXième siècle. Le vocabulaire contrôlé utilisé dans nos expérimentations est le thésaurus RAMEAU que nous décrivons en détails section 7.2.1.1 page 140. Pour identifier les entités spatiales dans l'ontologie, nous nous appuyons sur les bases de données BDTopo et BDCarto mises à disposition par l'IGN.

L'ontologie obtenue fait ressortir un ensemble de thèmes importants dans un espace géographique donné qui caractérise le fonds documentaire traité. La première étape d'analyse du travail d'indexation permet de construire un premier vocabulaire vocTerridoc constitué de **232 termes**. L'étape de structuration de ce vocabulaire à partir du thésaurus RAMEAU nous permet de construire un thésaurus contenant **540 vedettes**

(formant un vocabulaire de 3502 termes/labels si l'on prend l'ensemble des termes rejetés qui leur sont attachés) reliées sur **32 niveaux hiérarchiques** pour former un thésaurus que nous nommons tTerridoc. Le traitement des relations entre termes vedettes nous permet de construire **80 relations associatives** et **467 relations hiérarchiques**.

L'étape suivante de transformation du thésaurus tTerridoc en ontologie nous permet d'obtenir une structure composée de **507 concepts**. Ces concepts sont reliés sur **32** niveaux par la relation hiérarchique « est un ». **63 relations associatives** sont créées pour enrichir la structure et relier des concepts de même niveau conceptuel.

La dernière étape d'instanciation de cette ontologie permet d'identifier **26** entités géographiques. L'ontologie résultante est constituée de **481 concepts** et **26 instances** qui forme une première représentation territorialisée du fonds documentaire.

6.1.6.2 Premiers bilans

L'application de notre méthodologie TERRIDOC sur un fonds documentaire territorialisé indexé nous permet d'obtenir une première ontologie légère d'un territoire implicitement décrit par le travail d'indexation des experts documentalistes. Le thésaurus tTerridoc obtenu dans une étape intermédiaire offre une représentation sémantique structurée synthétisant le travail d'indexation et il peut être utilisée tel quel pour analyser ou encore valider le travail réalisé. Nous présentons un cas d'application découlant de notre approche section 8.3 page 192.

Nous obtenons ensuite une ontologie légère offrant une vue globale d'un territoire décrit dans les documents du fonds traité. Notons que dans nos expérimentations, nous n'intégrons pas les résultats concernant la composante temporelle car ils nécessitent encore une phase d'analyse. Le territoire identifié est alors caractérisé par un ensemble de sujets décrits dans un espace géographique. En effet, si nous reprenons le modèle du territoire défini section 5.6 page 93, l'ensemble des concepts sont des *entités thématiques* et l'ensemble instances de l'ontologie sont des *entités géographiques (EGs)* formant un espace géographique.

Cette ontologie offre ainsi une première représentation d'un territoire, en prenant en compte les composantes spatiales et thématiques, qui, en fonction du besoin, peut apparaître comme un résultat décrivant un territoire de façon suffisamment précise. Cependant, si l'on souhaite proposer une représentation plus détaillée d'un territoire décrit dans l'ensemble du fonds documentaire (documents + notices), les chiffres montrent que cette représentation est limitée en instances et offre une vue très générale de l'espace géographique décrit dans le fonds documentaire. Cela s'explique par le fait que les spécificités spatiales se résument à une liste d'ES choisies par les bibliothécaires pour décrire de façon très générale le contenu des documents, ce qui reste relativement limité en informations.

Pour enrichir la première représentation du territoire obtenue et vérifier la validité au niveau sémantique de cette première ontologie, nous proposons d'instancier l'ontologie par des éléments caractérisant le territoire identifiés dans le fonds documentaire (notices descriptives et documents textes attachés) via une chaine de traitement linguistique. La chaîne de traitement que nous proposons doit permettre d'extraire des ressources

textes des entités géographiques lorsqu'il est possible d'identifier un lien explicite avec les concepts de l'ontologie. Dans un premier temps, nous appliquons la chaine de traitement sur les éléments contenant des phrases (titre, légende, etc.) constituant les notices descriptives. Ce traitement doit nous permettre d'obtenir une vue synthétique d'un territoire implicitement décrit par un fonds documentaire hybride. Dans un second temps, nous proposons d'appliquer cette même chaîne de traitement linguistique aux documents plein-texte liés aux notices pour enrichir la représentation spatiale du fonds documentaire traité. Nous faisons donc le choix d'écarter à partir de cette étape le traitement des documents image, son et vidéo qui nécessitent des connaissances et des expérimentations qui sortent du cadre de nos travaux. Nous nous restreignons aux documents textes, ce qui implique de travailler sur un thésaurus synthétisant le travail d'indexation des experts obtenu à partir des notices attachés à ces documents textes.

6.2 TALN pour la représentation d'un territoire

Une contribution importante de notre travail concerne les étapes permettant d'enrichir de façon incrémentale la première ontologie de territoire obtenue à partir de la méthodologie présentée chapitre 6.1 en s'appuyant sur les documents textes et notices descriptives attachées. Dans une première étape (cf. section 6.2.1), nous proposons une chaine de traitement linguistique automatisée qui permet de marquer dans des documents textes l'ensemble des informations géographiques (entité thématique + nom toponymique). La chaîne de traitement intègre l'ontologie construite sur la base du travail des experts afin d'identifier des liens entre les concepts de l'ontologie et des lieux, propres à un territoire. Lorsqu'un lien est identifié entre un concept et un lieu dans un texte, le lieu est ajouté à l'ontologie sous forme d'instance du concept concerné par la relation.

Dans une seconde étape (*cf.* section 6.2.2), nous appliquons la chaîne de traitement sur l'ensemble des notices descriptives et plus précisément aux éléments correspondant à la phase de description du document (titre, légende, etc.) dans le but d'enrichir la représentation du territoire par des entités spatiales. Nous décrivons brièvement ensuite nos expérimentations réalisées sur le jeu de notices descriptives utilisées pour créer la première ontologie afin de valider cette étape. Nous montrons que l'ontologie obtenue offre, par l'ajout de ces instances caractérisant un espace géographique, une représentation plus précise du territoire décrit dans les notices attachées à un ensemble de documents hybrides. Nous verrons que, malgré la prise en compte des différents constituants de la notice descriptive, la représentation que nous obtenons offre une vue encore générale du territoire décrit par le fonds traité. Le résultat peut s'avérer imprécis si l'objectif est de pouvoir faire ressortir du fonds documentaire un maximum d'entités spatiales. Pour tenter de répondre à ce type de besoin, nous proposons ensuite (*cf.* section 6.2.3) d'appliquer la chaîne de traitement sur le contenu des documents textes eux-mêmes. Les spécificités du fonds documentaire sur lequel nous réalisons nos expérimentations nous laissent penser que cette étape doit permettre d'identifier une quantité importante d'entités géographiques. Ces EGs sont ensuite destinées à intégrer l'ontologie sous forme d'instance et ainsi enrichir la représentation du territoire obtenue à partir des notices descriptives.

Les expérimentations présentées ensuite confirment l'intérêt de notre approche.

6.2.1 Chaîne de TAL pour l'indexation d'entités géographiques

Nous définissons dans cette section une chaîne de traitement complète (Figure 6.12) de construction d'index particulièrement adaptée à l'aspect géographique des contenus [BKG09].

6.2.1.1 Chaîne pour la détection d'entités nommées

En nous appuyant sur les travaux de [AFG03], nous définissons une première chaîne de traitement composée de quatre grandes phases :

1. la lemmatisation pour segmenter les mots ;
2. l'analyse lexicale et morphologique pour la reconnaissance des mots ;
3. l'analyse syntaxique, basée sur des grammaires, afin de trouver les relations entre les mots ;
4. enfin l'analyse sémantique pour réaliser une interprétation plus spécifique sur les syntagmes retenus.

Résultent de ce traitement linguistique un ensemble d'entités nommées et des relations existantes entre ces dernières.

L'étape (1) s'appuie sur une « tokenisation » classique. Nous adoptons ensuite une démarche qui consiste à marquer rapidement des EN candidates puis à appliquer les étapes suivantes de l'analyse à ces EN uniquement. Un marqueur de token d'entité nommée candidate (2) utilise des règles typographiques (majuscule en début de token). Puis, un analyseur morpho-syntaxique (3) associe un lemme et une nature à chaque token d'EN candidate (i.e. « *Association* », nom). Un analyseur sémantique (4) associé à des règles de grammaire qualifie des entités nommées absolues (ENA) et des entités nommées relatives (ENR). Une ENA est une entité nommée simple : « le président de France », et une ENR est une EN complexe définie à partir d'une autre EN : « le véhicule Peugeot du président de France ». Afin de repérer des relations sémantiques [AB08], nous utilisons des patrons lexico-syntaxiques. Un patron lexico-syntaxique représente une expression régulière, formée de mots, de catégories grammaticales ou sémantiques, et de symboles. Le traitements des entités nommées géographiques nécessite des traitements particuliers que nous allons maintenant présenter.

6.2.1.2 Cas particulier des entités géographiques

Au sein de la chaîne de traitement proposée, nous nous appuyons sur l'annotation automatique d'Entités Nommées (EN) conceptualisées particulières : les noms toponymiques décrivant un espace et les EN temporelles calendaires décrivant une période. Les entités géographiques nécessitent des adaptations de la chaîne permettant de détecter des entités nommées (figure 6.12).

L'étape (1) n'est pas modifiée par rapport à la première chaîne. Dans l'étape (2) permettant d'identifier des entités géographiques candidates, le marqueur de token spatial candidat utilise en plus des règles typographiques des règles lexicales (lexiques d'introducteurs d'entités nommées spatiales). Puis, l'analyseur morpho-syntaxique (3) associe un lemme et une nature à chaque token spatial candidat (i.e. « *Marais* », nom). Dans la phase d'analyse sémantique (4), les règles de grammaire sont enrichies pour qualifier des entités géographiques intégrant des entités spatiales absolues (ESA) et des entités spatiales relatives (ESR). Dans le cadre de nos travaux, nous nous intéressons uniquement aux EGs intégrant des ESA : « *le quartier du Marais* », « *le pic d'Ossau* », « *la vallée d'Ossau* », etc. Pour le traitement des entités nommées géographiques, nous intégrons à la chaîne linguistique une dernière étape (5) permettant de valider et de géolocaliser les ESA à l'aide de ressources externes ou de gazetteers contributives internes. Cette chaine de traitement est détaillée dans [KKS+09].

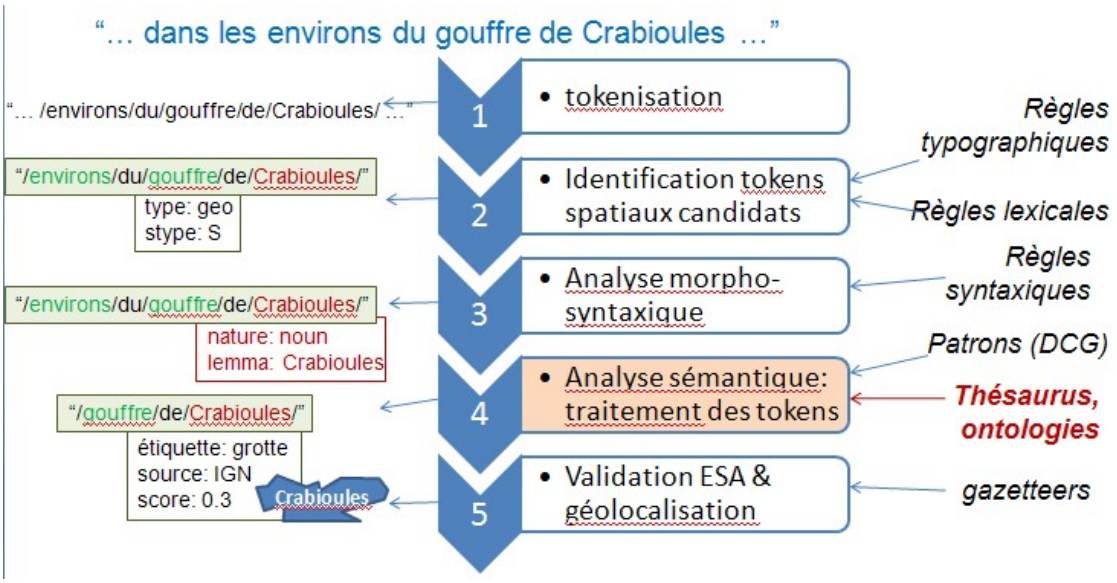


FIGURE 6.12 – Chaîne de traitement d'informations géographiques dans des documents textuels

La chaîne de traitement proposée permet de produire une liste d'EGs à partir de textes. Nous l'appliquons sur l'ensemble des notices descriptives utilisées pour construire la première ontologie dans l'objectif d'enrichir l'ontologie avec de nouvelles instances de concepts. Ces instances doivent apporter de nouvelles informations caractérisant le territoire décrit dans le fonds documentaire.

6.2.2 Application sur le contenu des notices descriptives

Nous cherchons à identifier les entités géographiques qui ont un lien sémantique avec les concepts présents dans l'ontologie générée sur la base du travail d'indexation.

Nous appliquons tout d'abord notre chaîne de traitement sur les notices descriptives attachées aux documents du fonds documentaire traité. Pour exemplifier notre approche, l'exécution de notre chaîne de traitement sur les extraits de notices présentées figure 6.2 et figure 6.3 (*cf.* page 105) nous permet d'obtenir la liste d'entités géographiques suivante :

- « *les autres eaux minérales de Bigorre et du Béarn* »
- « *des eaux minérales de Bagnères de Bigorre* »

A partir de cette liste, nous présentons maintenant l'étape d'instanciation de l'ontologie obtenue sur la base du travail d'indexation des experts.

6.2.2.1 Instanciation de l'ontologie à partir des notices descriptives

Cette étape consiste à instancier l'ontologie en exploitant la liste des entités géographiques résultante de notre chaîne de traitement appliquée sur les notices descriptives.

La règle définie pour instancier l'ontologie (*cf.* tableau 6.9) indique que, pour chacune des entités thématiques t1 constituant une EG *EN1*, une première étape permet de vérifier si ce dernier apparait en tant que label d'un concept dans l'ontologie. S'il est présent (label du concept C1), le nom_toponymique associé devient alors une instance I1 du concept C1 identifié dans l'ontologie que l'on relie par la relation de type « instance de ». La représentation spatiale est ajoutée à l'instance I1.

Conditions	- EN1 validé comme un nom_toponymique
	- t1 entité thématique de EN1 et label de concept C1
Actions	- Création I1 instance de C1
	- Ajout propriétés spatiales à I1

TABLE 6.9 – Instanciation de l'ontologie à partir des notices descriptives

Ainsi en reprenant la liste d'entités géographiques obtenues à partir des notices présentées figures 6.2 et 6.3 :« eaux minérales de Barèges », « eaux minérales des Eaux Chaudes », « eaux minérales de Bigorre », « eaux minérales du Béarn » et « des eaux minérales de Bagnères de Bigorre ». Trois EGs sont retenues après validation via l'appel au gazetteers : « eaux minérales de Barèges », « eaux minérales des Eaux Chaudes », et « des eaux minérales de Bagnères de Bigorre ». Pour chaque EG retenue, si elle existe déjà dans l'ontologie, on passe à l'EG suivante. Parmi les trois EGs identifiées, « *eaux minérales de Bagnères de Bigorre* » et « *eaux minérales de Barèges* » sont déjà présentes dans l'ontologie obtenue à partir du travail d'indexation « . Stations_climatiques,_thermales,_etc._de-_Bagnères-de-Bigorre_(Hautes-Pyrénées) », notamment, est une instance des concepts « *Stations-_climatiques,_thermales,_etc.* » et n'est donc pas traitée dans cette étape. L'entité géographique « *eaux minérales des Eaux Chaudes* » n'étant pas présentes dans l'ontologie, une phase de recherche permet d'indiquer que l'entité thématique de ces informations géographiques « eaux minérales » est le label du concept « eaux_minérales » dans notre ontologie. Une instance du concept « eaux_minérales » est alors créée avec pour idenfifiant « eaux_minérales_des_Eaux-_Chaudes ».

Nous présentons figure 6.13 une représentation visuelle de l'étape d'instanciation

réalisée à partir de l'extrait de l'ontologie présenté figure 6.13 (*cf.* page 127) et des deux notices exemples schématisée figures 6.2 et 6.3.

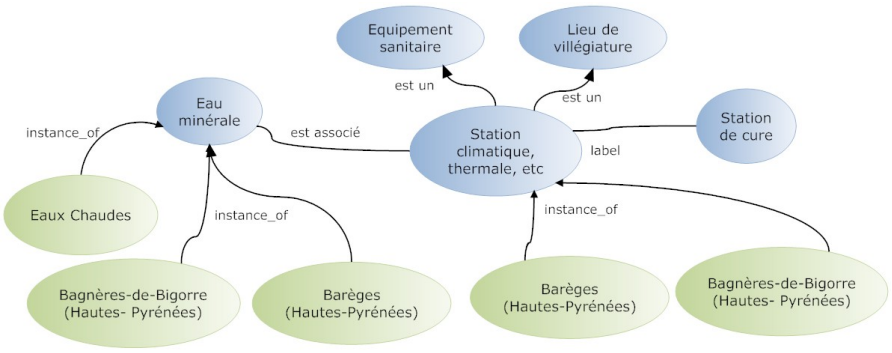


FIGURE 6.13 – Extrait de l'ontologie enrichie

6.2.2.2 Tests et conclusion

L'application de la chaîne de traitement linguistique sur un ensemble de notices descriptives attachées à des documents de type image, son, vidéo et texte nous permet d'identifier un ensemble d'entités géographiques formant un espace géographique. L'intégration de la première ontologie synthétisant le travail d'indexation des experts dans cette chaîne de traitement linguistique nous permet d'identifier des liens de type « instance_de » entre ces entités géographiques et les concepts de notre ontologie.

En appliquant notre chaîne de traitement linguistique sur les 900 notices descriptives consituant notre corpus (*cf.* voir le résultat des expérimentations figure 7.26 page 159), nous identifions 266 entités géographiques (126 termes distincts) parmi lesquelles **115** entités sont attachées à un label de concept présent dans l'ontologie, ce qui nous permet d'obtenir une ontologie constituée de **141** instances décrivant un territoire (pour 26 instances avant le traitement linguistique appliqué sur les notices descriptives). La quantité des entités géographiques ajoutées à notre ontologie minimale augmente donc de façon significative lorsque nous appliquons notre chaîne de traitement linguistique sur les notices descriptives.

Seules, ces ressources peuvent ne pas s'avérer suffisantes pour définir une structure suffisamment complète pour décrire un territoire. Nous proposons donc d'enrichir la représentation d'un territoire en utilisant comment point d'entrée le contenu de documents textes dits territorialisés. Rappelons que ce type de document se caractérise par une omniprésence des noms de lieux relatifs à un territoire particulier.

6.2.3 Application sur le contenu des documents

Nous proposons d'appliquer notre chaine de traitement linguistique sur les documents eux-mêmes afin d'identifier des entités géographiques dans le contenu des documents constituant le corpus. Nous faisons l'hypothèse en traitant directement les documents

de pouvoir obtenir une représentation plus précise du territoire décrit par le fonds documentaire indexé. Pour ce faire, nous limitons nos traitements aux documents textes car le traitement du contenu des documents de type image, son et vidéo impliquent des traitements particuliers qui vont bien au-delà du cadre de nos travaux de thèse.

6.2.3.1 Instanciation de l'ontologie à partir du contenu des documents textes

Nous cherchons ici à identifier par une analyse linguistique du corpus de référence des éléments caractérisant un espace. Le fonds documentaire de référence est composé de versions électroniques de documents texte. Ces documents contiennent de très nombreuses références au territoire pyrénéen, comme le montre l'extrait présenté figure 6.14. Nous faisons le choix de restreindre l'identification des EGs aux entités nommées constituées d'un concept présent dans l'ontologie minimale et ancré sur un nom toponymique.

Après avoir visité aussi le Vignemale et le Pic du Midi de Bigorre, je désirai ne point quitter les Pyrénées sans avoir fait du moins un effort en faveur de l'ascension de la Maladetta. Je partie en conséquence pour Bagnères de Luchon une seconde fois et, passant par le col du Tourmalet, la Mourquette d'Arreau et la belle vallée de Louron (que je trouve bien plus riante que celle de Campan - j'en demande pardon à sa réputation), j'arrivai à Luchon le 17 juillet.

Mes mesures furent immédiatement prises pour recueillir toutes les informations nécessaires à mon projet. Accompagné de mon fidèle et brave Pierre Sanic, guide de Luz, qui ne m'avait pas quitté depuis près de six semaines d'excursions continuelles, et que l'approche du danger seul pouvait faire sortir de son humeur pacifique, je parlai de mes intentions à plusieurs guides de Luchon, qui étaient censés connaître leurs montagnes. Il n'en était pas ainsi ; non seulement ils n'avaient aucun renseignement précis sur la Maladetta, mais encore paraissaient-ils fort peu disposés à s'en procurer, tant l'amour d'un gain obtenu sans peine, dans des promenades faciles, et la crainte des périls les avaient rendus indifférents pour le plus beau monument de leur pays.

FIGURE 6.14 – Extrait d'un fichier texte indexé par la chaîne de TAL

L'intégration de l'ontologie obtenue à partir du travail d'indexation nous permet d'identifier une quantité importante d'entités géographiques (*cf.* figure 7.27 page 160). Parmi ces entités géographiques, **2967** sont reliées à des concepts présents dans l'ontologie. La quantité des entités géographiques identifiées comme instances de notre ontologie minimale augmente considérablement lorsque nous appliquons notre chaîne de traitement

linguistique sur le contenu des documents textes. Nous montrons ici que le traitement du contenu des documents texte est un élément indispensable lorsque l'on souhaite obtenir une représentation la plus complète possible d'un territoire. Cependant, dans le cadre de nos travaux, le traitement du contenu des documents implique de restreindre le corpus aux documents texte uniquement.

6.2.3.2 Tests et conclusion

Nous remarquons d'après les statistiques présentées (*cf.* figure 7.27 page 160) que beaucoup d'entités géographiques, identifiées via notre chaîne de traitement, ne sont pas constituées d'entités thématiques correspondant à label d'un concept de notre ontologie. Nous proposons de définir un processus automatisé afin de prendre en compte ces informations afin d'enrichir notre ontologie minimale par de nouveaux concepts et relations. Nous nous appuyons pour cela sur la ressource RAMEAU, afin d'identifier des termes du thésaurus qui, bien qu'ils n'aient pas été utilisés par les experts bibliothécaires pour décrire le contenu des documents, semblent avoir une importance dans le contenu des documents.

6.3 Enrichissement de l'ontologie à partir du contenu des documents

Rappelons ici qu'à partir de l'analyse linguistique du corpus, nous souhaitons enrichir l'ontologie légère en y insérant les informations contenues dans les entités géographiques identifiées à partir du traitement linguistique présenté section 6.2.1 (*cf.* page 124). L'enrichissement se fait en deux étapes : ajout de nouveaux labels/concepts permettant d'élargir la couverture sémantique du domaine puis ajout de relations spatiales explicitant les liens entre concepts selon le point de vue territorial. Dans les travaux présentés ici, l'étape d'enrichissement de la structure de l'ontologie n'étant pas assez avancée, nous la positionnons en pespective.

La phase d'enrichissement, bien qu'automatique, est validée en fin de processus par les experts bibliothécaires du domaine. L'ontologie enrichie doit permettre de faciliter la découverte d'un territoire à travers les documents. L'approche présentée figure 6.15 [KKS+09] permet d'exploiter de façon automatisée la liste des entités géographiques pour enrichir l'ontologie légère par des éléments offrant une représentation plus riche d'un territoire. Pour chacune des entités thématiques constituant une entité géographique, une première étape permet de vérifier si ce dernier apparait en tant que label d'un concept dans l'ontologie. S'il est présent, le nom toponymique associé, une fois validé, devient alors une instance du concept identifié dans l'ontologie que l'on relie par la relation de type « instance de »
textitcf. section 6.2.1.2 page 124). Dans le cas où le terme n'est pas présent en tant que label dans l'ontologie, nous proposons d'enrichir le vocabulaire de l'ontologie.

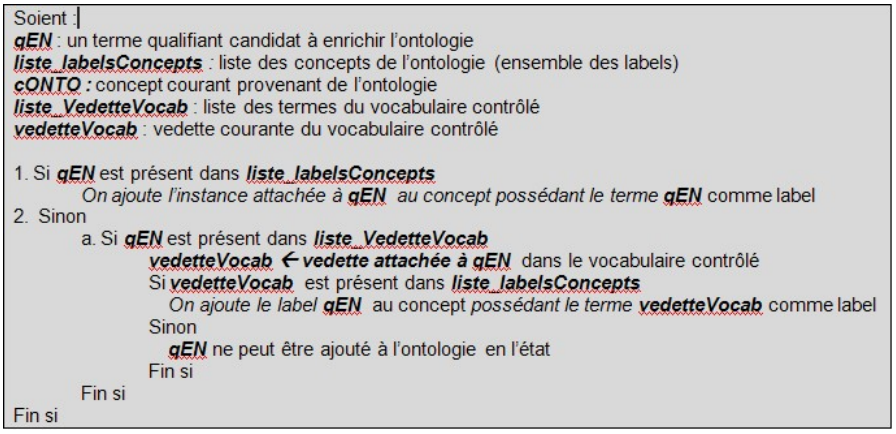


FIGURE 6.15 – Algorithme général d'enrichissement de l'ontologie

6.3.1 Enrichissement de l'ontologie par des concepts

Nous faisons appel à nouveau à la ressource thésaurus utilisée par les experts. Le patron présenté tableau 6.10 indique que si le terme existe dans le thésaurus, la vedette correspondante est comparée à l'ensemble des concepts (chacun avec ses différents labels) afin d'identifier d'éventuelles approximations de sens entre ce terme et les concepts de l'ontologie. Si une équivalence est trouvée, le terme devient alors un nouveau label de concept dans l'ontologie.

Conditions	- t1 validé comme un nom_toponymique
	- t3 entité thématique de t1 et synonyme de label t2 de concept C2
(dans le thésaurus)	
Actions	- Ajout t3 label de C2
	- Création I1 instance de C2
	- Ajout propriétés spatiales à I1

TABLE 6.10 – Enrichissement de l'ontologie par des concepts

Prenons l'exemple du syntagme « la ville thermale des Eaux-Chaudes », l'entité thématique « ville thermale » est présente dans le thésaurus en tant que terme « employé pour » de la vedette « Stations climatiques, thermales, etc. » (figure 6.6 page 109). Rappelons que dans l'exemple des stations climatiques, « Stations climatiques, thermales, etc. » existe dans l'ontologie en tant que label du concept « Station_climatique, _thermale,_etc. », ce qui nous permet de lui ajouter un nouveau label « ville thermale ». Le nom toponymique Eaux-Chaudes est alors ajouté en tant qu'instance du concept « Station_climatique,_thermale,_etc. » avec comme identifiant « Station_climatique,_thermale, _etc._des_Eaux-Chaudes ».

Cette étape nous permet de prévoir un enrichissement de l'ontologie de **940 concepts** (soit potentiellement 7283 nouvelles instances) comme le montre la figure (*cf.* figure 7.28

page 161). Ces propositions d'enrichissement doivent en fin de traitement être validées par les experts bibliothécaires.

La méthodologie proposée permet ainsi d'enrichir l'ontologie en ajoutant de nouveaux labels et de nouveaux concepts afin de proposer une couverture sémantique d'un territoire plus importante.

6.4 Discussion

Ce chapitre présente de façon détaillée une méthodologie opérationalisée qui permet de construire une représentation sémantique synthétique sous la forme d'une ontologie d'un territoire à partir d'un fonds documentaire annoté par des experts.

Le processus de création d'ontologie intègre dans une première étape (*cf.* section 6.1.4.1 page 108)l'identification, l'extraction et la structuration sous forme d'un thésaurus tTerridoc du vocabulaire expert utilisé pour décrire le contenu des documents. Le thésaurus produit offre une réprésentation sémantique synthétisant le travail d'indexation des experts et nous proposons de le formaliser en SKOS pour en faciliter l'accès. Des présentations aux experts de la MIDR ainsi qu'à la BNF [Ker10] montrent l'intérêt de ces premiers travaux et les experts y voient notamment des possibilités d'aide au travail d'indexation et à la maintenance d'un vocabulaire contrôlé. Au niveau scientifique, des publications résultent de ces premiers travaux [KBG08, KBGre].

Dans une seconde étape, nous décrivons le processus de transformation du thésaurus tTerridoc en une première ontologie d'un territoire (*cf.* section 6.1.4.2, page 114). Nous obtenons une ontologie légère offrant une vue globale d'un territoire décrit dans le vocabulaire formant le travail d'indexation. Le territoire identifié est alors caractérisé par un ensemble de sujets décrits dans un espace géographique. En effet, si nous reprenons le modèle du territoire défini figure 5.8 (*cf.* page 96), l'ensemble des instances de l'ontologie sont des *entités géographiques (EGs)*, formant un espace géographique, et les concepts instanciés apparaissant comme des *entités thématiques*. Nous n'intégrons pas les résultats concernant la composante temporelle car ils nécessitent encore une phase d'analyse.

Une contribution importante de notre travail concerne les étapes permettant d'enrichir de façon incrémentale la première ontologie de territoire obtenue à partie de la méthodologie présentée chapitre 6.1 en s'appuyant sur les documents textes et notices descriptives attachées. Une troisième étape permet d'enrichir cette première représentation d'un territoire en appliquant notre chaine de traitement linguistique sur les notices descriptives attachées aux documents du corpus traité (c.f. section 6.2.2 page 125). En effet, si l'on souhaite proposer une représentation détaillée d'un territoire décrit dans l'ensemble du fonds documentaire (documents + notices), les chiffres montrent que la première ontologie obtenue à partir du travail d'indexation des experts est limitée en instances et offre une vue très générale de l'espace géographique décrit dans le fonds documentaire. Cela s'explique par le fait que les spécificités spatiales se résument à une liste d'entités spatiales choisies par les bibliothécaires pour décrire de façon très générale le contenu des documents, ce qui reste relativement limité en informations. La chaîne de

traitement que nous proposons permet d'extraire de ressources texte des informations géographiques lorsqu'il est possible d'identifier un lien explicite entre une information spatiale (un nom toponymique) et un concept de l'ontologie. Nous montrons section 6.2.2 (c.f. page 125) que l'ontologie obtenue offre, par l'ajout de ces instances caractérisant un espace géographique, une représentation plus précise du territoire décrit dans les notices attachées à un ensemble de documents hybrides.

Cependant, malgré la prise en compte des différents constituants de la notice descriptive, la représentation que nous obtenons offre une vue encore générale du territoire décrit par le fonds traité. Le résultat peut s'avérer imprécis si l'objectif est de pouvoir faire ressortir du fonds documentaire un maximum d'entités spatiales. Pour tenter de répondre à ce type de besoin, nous proposons ensuite (*cf.* section 6.2.3) d'appliquer la chaîne de traitement sur le contenu des documents texte eux-mêmes. Les spécificités du fonds documentaire sur lequel nous réalisons nos expérimentations nous laissent penser que cette étape doit permettre d'identifier une quantité importante d'entités géographiques. Ces EGs intègrent l'ontologie sous forme d'instances et enrichissent la représentation du territoire obtenue à partir des notices descriptives. Les expérimentations présentées ensuite confirment l'intérêt de notre approche. Au niveau scientifique, des publications résultent de ces travaux [BKDB07, Ker08, BKG09, BGKS10, KBGre].

Nous travaillons actuellement à l'identification d'éléments candidats à enrichir l'ontologie d'un territoire. Nous proposons un processus permettant d'ajouter de nouveaux labels/concepts afin d'élargir la couverture sémantique du domaine. La phase d'enrichissement, bien qu'automatique, doit être validée en fin de processus par les experts bibliothécaires du domaine. L'ontologie enrichie doit permettre de faciliter la découverte d'un territoire à travers les documents. L'approche présentée figure 6.15 [KKS+09, KKB+re] permet d'exploiter de façon automatisée la liste des entités géographiques dans le cadre du projet Géonto pour enrichir une ontologie géographique par des éléments offrant une représentation plus précise d'un territoire.

Quatrième partie

Implémentations

Chapitre 7

Création d'une ontologie légère décrivant le territoire des Pyrénées

Sommaire

7.1 Chaîne de traitement TERRIDOC 137
7.1.1 Description de TERRRIDOC 137
7.1.2 Technologies et langages utilisés 138
7.2 Construction d'une première ontologie à partir de la connaissance structurée des bibliothécaires 139
7.2.1 Constitution du jeu d'essai 139
7.2.2 Module (1) d'extraction de la connaissance experte : création du vocabulaire vocTerridoc 144
7.2.3 Module (2) de structuration des termes : normalisation en thésaurus tTerridoc 145
7.2.4 Module (3) de transformation du thésaurus tTerridoc en une première ontologie légère 152
7.2.5 Module (4) d'instanciation de l'ontologie pour l'émergence d'un territoire 154
7.3 TALN pour la représentation d'un territoire 156
7.3.1 Modules (5 et 6) pour l'enrichissement de la représentation d'un territoire à partir du travail d'indexation d'experts . 158
7.3.2 Modules (7 et 8) pour l'enrichissement de la représentation d'un territoire à partir du contenu des documents 160
7.4 Enrichissement de l'ontologie à partir du contenu des documents 161

La méthodologie que nous proposons chapitre 6 page 101 permet de modéliser la phase de création d'une ontologie d'un territoire en quatre étapes principales : (i) Construction d'une première ontologie minimale à partir d'un vocabulaire contrôlé structuré ; (ii)

TALN pour l'identification d'entités géographiques à partir du travail d'indexation d'experts ; (iii) TALN pour l'identification d'entités géographiques à partir du contenu des documents ; (iv) Enrichissement de l'ontologie minimale à partir du contenu des documents. Nous reprenons figure 7.1 le processus de création d'ontologies dans lequel nous spécifions les ressources utilisées pour valider notre approche.

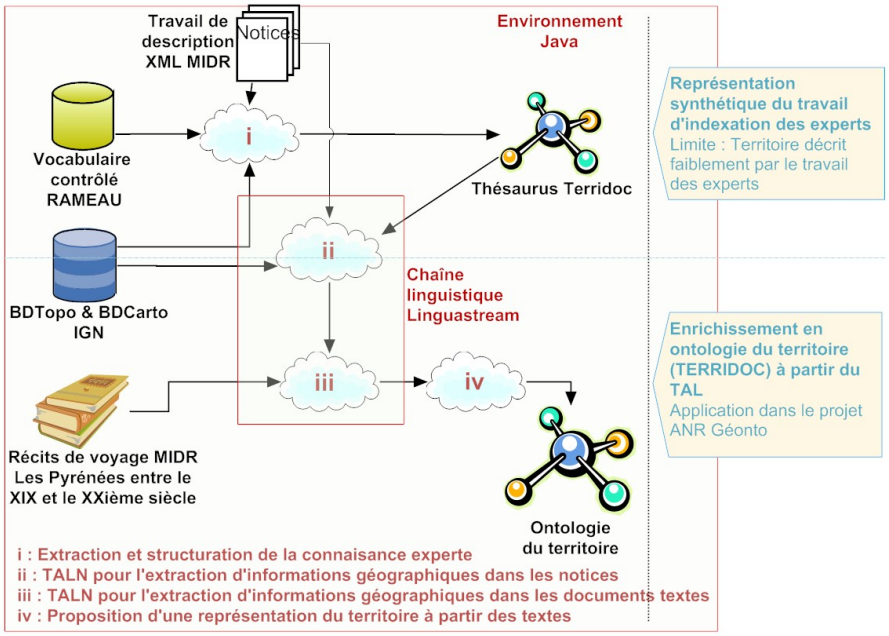


FIGURE 7.1 – Technologies utilisées pour l'implémentation de la méthodologie Terridoc

Dans un premier temps, nous expérimentons le processus de création d'une première ontologie d'un territoire réalisée sur la base :

- d'un fonds documentaire, constitué de documents et de notices descriptives Xml attachées, mis à disposition par la MIDR ;
- du thésaurus RAMEAU utilisé par les experts bibliothécaires pour indexer les documents du corpus traité ;
- des bases de données BDTopo et BDCarto mises à disposition par l'IGN, et servant de lexiques toponymiques pour valider les entités nommées identifiées dans les notices.

Avant de présenter nos expérimentations, nous détaillons l'architecture de la chaine Terridoc ainsi que les technologies utilisées pour développer l'ensemble des modules.

7.1 Chaîne de traitement TERRIDOC

7.1.1 Description de TERRRIDOC

Nous détaillons figure 7.2 l'architecture de TERRIDOC que nous décomposons en huit étapes disctinctes.

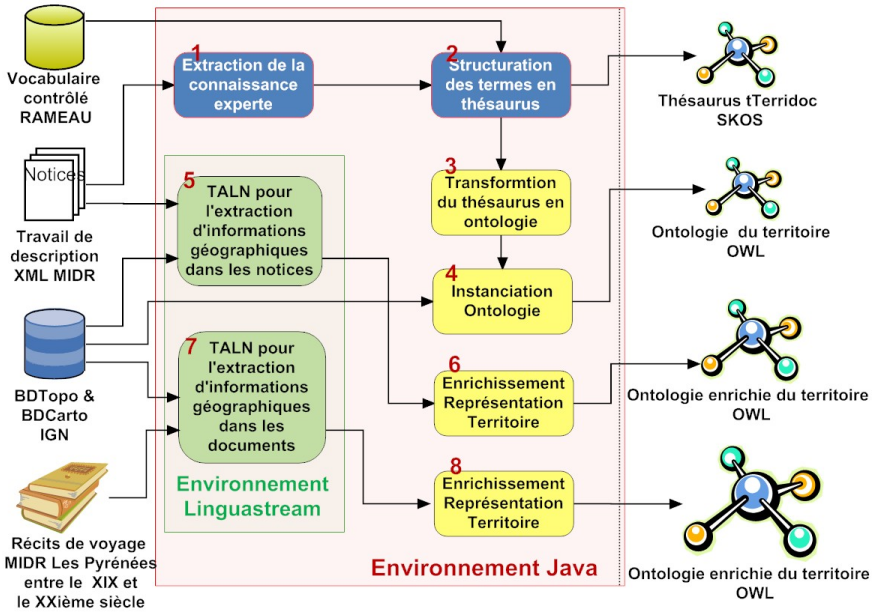


FIGURE 7.2 – Chaîne de traitement TERRIDOC

La chaîne TERRIDOC se compose de huit modules :

1. Extraction de la connaissance experte ;
2. Structuration des termes en thésaurus tTerridoc ;
3. Transformation du thésaurus tTerridoc en une première ontologie ;
4. Instanciation de l'ontologie pour l'émergence d'un territoire ;
5. TALN pour l'extraction d'informations géographiques dans les notices ;
6. Enrichissement de la représentation ontologique d'un territoire ;
7. TALN pour l'extraction d'informations géographiques dans les documents ;
8. Enrichissement de la représentation ontologique d'un territoire.

Les modules 1, 2, 3 et 4 permettent d'obtenir une première représentation sémantique d'un territoire à partir du travail d'indexation réalisé par les bibliothécaires de la MIDR sur la base du thésaurus RAMEAU défini par la BNF. A chaque étape de notre traitement, la validation des entités spatiales (constituant des entités géographiques), est réalisée en faisant appel aux bases de données BDTopo et BDCarto définis par l'IGN. Bien qu'intéressante pour les experts, des expérimentations mettent en avant le fait que

l'ontologie obtenue est limitée en informations géographiques car le vocabulaire correspondant à l'ensemble des termes utilisés lors du travail d'indexation n'en contient que trop peu.

Nous proposons d'appliquer la chaîne de traitement, définie section 6.2 page 123 à l'aide de la plateforme Linguastream [78] [BCEM03, Bil06], qui s'appuie sur des modules de traitement linguistique surfacique et utilisant le langage XML pour la structuration des résultats. Nous appliquons, dans le module 5 de TERRIDOC, la chaîne de traitement linguistique sur l'ensemble des champs des notices descriptives (légende, titre, etc.), et nous montrons que ce traitement permet d'enrichir, à l'aide du module 6, la représentation sémantique obtenue à partir du travail des experts (notamment par des entités géographiques).

Toutefois, la représentation du territoire est encore limitée et nous proposons alors d'appliquer, module 7 de Terridoc, notre chaîne de traitement linguistique non plus sur les notices mais sur les documents eux-mêmes. L'expérimentation sur un fonds documentaire de 900 documents fait ressortir une quantité importante d'éléments spatiaux et thématiques qui caractérisent d'un territoire, mais aussi des relations entre ces éléments. Le module 8 de Terridoc permet ensuite d'enrichir l'ontologie par ces entités géographiques.

7.1.2 Technologies et langages utilisés

Au niveau développement, JAVA est le langage choisi pour développer les différents modules car il est maitrisé à la fois par les différents membres de l'équipe de recherche T2I au sein du LIUPPA ainsi que par les membres de l'équipe à l'entreprise DIS. Aussi, JAVA est le langage sur lequel s'appuie Linguastream pour fonctionner. Enfin, ce langage apporte l'ensemble des fonctionnalités requises pour traiter les sources textes et notices Xml en entrée ainsi que des librairies facilitant la productions des sorties SKOS [79] et OWL [80]. L'environnement de développement choisi est Eclipse 3.3 WTP dans la version standard.

Au niveau de la chaîne de traitement linguistique, nous avons défini et testé la chaine complète en utilisant l'environnement d'édition visuel intégré à Linguastream. Une fois validée, nous avons intégré la chaine linguistique dans un module Java exécutable en ligne de commande pour éviter de passer par l'éditeur de Linguastream. Dans la chaine linguistique, la phase de marquage des entités géographiques implique de définir des patrons respectant la syntaxe DCG (Definite Clause Grammars) de Prolog [81]. L'extraction et la manipulation des entités géographiques nécessitent de définir des fichiers intérmédiaires que nous formalisons en XML en nous appuyant sur les langages XSLT [82] et Xquery [83].

78. http ://www.linguastream.org/
79. la librairie SkosApi : http ://skosapi.sourceforge.net/
80. La librairie Jena : http ://jena.sourceforge.net/ontology/
81. module exécutable Swi-Prolog
82. http ://xmlfr.org/w3c/TR/xslt/
83. http ://www.w3.org/TR/xquery/

Présentons maintenant en détails nos expérimentations en décomposant la chaîne Terridoc selon les modules présentés figure 7.2.

7.2 Construction d'une première ontologie à partir de la connaissance structurée des bibliothécaires

Avant de décrire plus en détails la chaîne de traitement, nous présentons tout d'abord le module permettant d'identifier et sélectionner un ensemble de documents indexés sur lesquels nous exécutons notre chaîne de traitement.

7.2.1 Constitution du jeu d'essai

Le fonds documentaire mis à disposition par la MIDR de Pau est constitué de 5000 documents. Nous proposons un premier module permettant de sélectionner un ensemble de documents pour créer l'ontologie de territoire.

7.2.1.1 Ressources disponibles

Le premier objectif est de constituer un fonds documentaire hybride conséquent regroupant textes, images, sons et vidéos. L'ensemble devra être accompagné des notices descriptives correspondantes au format Xml.

Fonds documentaire de la MIDR et notices descriptives

Le fonds documentaire que nous utilisons est constitué de documents décrivant la région des Pyrénées entre le XIXe et le XX siècle. A ces documents sont attachés des notices descriptives Xml réalisées manuellement par les experts bibliothécaires de la MIDR. Un guide est publié par la MIDR [Bar05], concernant l'indexation (basée sur le schéma Dublin Core) et le stockage des documents. Ce document permet de standardiser les informations décrivant les documents constituant le fonds documentaire de la Médiathèque. La démarche est intéressante et importante pour assurer la collaboration des différents intervenants (extérieurs pour la plupart). 26 types de documents, et principalement des images et des documents textes (journaux et livres) composent le fonds documentaire (figure 7.3).

A chaque document numérique ou ensemble de documents est liée une notice descriptive comportant une vingtaine de champs (figure 7.4). La notice descriptive est formalisée en XML, selon des règles définies par la MIDR [Bar05], respectant les préconisations de description des données définies pour la BNRP [84].

Dans leur démarche, les experts bibliothécaires doivent, lorsqu'ils le peuvent, utiliser des termes du thésaurus RAMEAU pour décrire le contenu des documents. Et, le fait que ces informations soient extraites d'un vocabulaire contrôlé nous permet d'exploiter le travail d'expert notamment sur le choix de la sémantique, de la terminologie et la syntaxe des termes sélectionnés pour décrire des documents.

84. Bibliothèque Numérique des Ressources Paloises

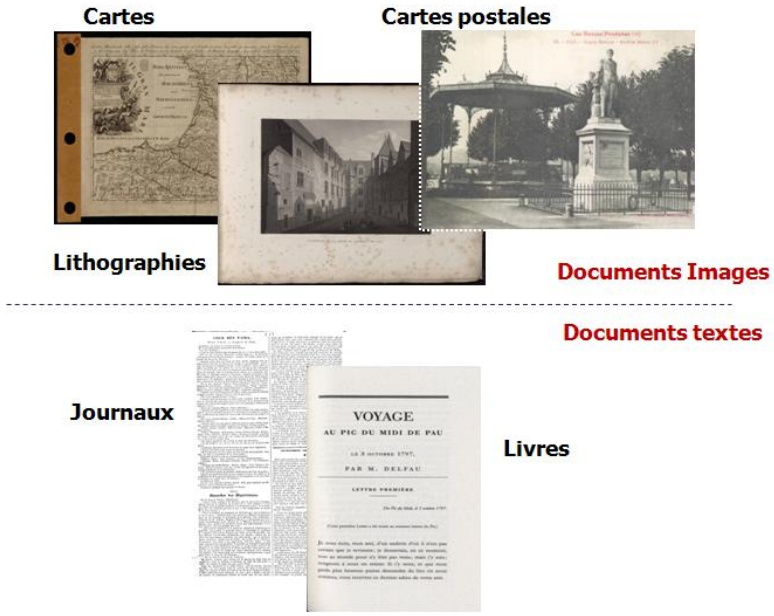


FIGURE 7.3 – Fonds documentaire de la MIDR

Nous avons décrit le travail d'annotation réalisé par les experts de la MIDR et formalisé en XML (cf figure 7.4) en faisant un état des outils utilisés pour le réaliser, et notamment le thésaurus RAMEAU. Afin de réaliser nos expérimentations, nous allons maintenant présenter le module permettant d'extraire un sous-ensemble représentatif du fonds documentaire.

Le thésaurus RAMEAU

Chaque partenaire utilise le vocabulaire d'indexation qui lui est propre. La MIDR, comme l'ensemble des médiathèques ainsi que la majorité des bibliothèques municipales et universitaires en France, utilise le langage d'indexation RAMEAU. Le langage d'indexation précoordonné RAMEAU se compose d'un vocabulaire de termes reliés entre eux et d'une syntaxe indiquant les règles de construction des vedettes-matière à l'indexation. RAMEAU est un thésaurus complet contenant des vedettes pouvant être reliées par des relations hiérarchiques (relations génériques et spécifiques), et des relations associatives. Les vedettes sont décrites sous la forme de notices au format Unimarc conforme à la norme ISO 2709. Une présentation graphique est donnée figure 7.5 avec l'exemple de la vedette « Stations climatiques, thermales, etc. ».

Dans l'exemple donné figure 7.5, la vedette « Stations climatiques, thermales, etc. » possède entre autres comme termes « employés pour » les termes « Stations thermales » et « Villes thermales ».

Dans l'exemple de notice descriptive présentée figure 7.4 page 141, les vedettes choisies par l'expert pour décrire le contenu du document sont listées dans la balise *doc_dee*.

RAMEAU respecte l'ensemble des règles propres aux vocabulaires contrôlés, et no-

```
<NOTICE>
<BNSA_GEOREF>Hautes-Pyrénées</BNSA_GEOREF>
<BNSA_GEOREF>Pyrénées-Atlantiques</BNSA_GEOREF>
<BNSA_GEOREF>Haute-Garonne</BNSA_GEOREF>
<DATM>2007-04-12</DATM>
<DOC_ANALYSE>0</DOC_ANALYSE>
<DOC_AUTEUR>Clausade, Gustave de</DOC_AUTEUR>
<DOC_AUTEURSEC>Malbos, E. de</DOC_AUTEURSEC>
<DOC_AUTMORAL>Ch. Gosselin (Paris)</DOC_AUTMORAL>
<DOC_COLLECTION>Legs Dussert</DOC_COLLECTION>
<DOC_COTE>99144R</DOC_COTE>
<DOC_DAT_CREAT>2007-04-12</DOC_DAT_CREAT>
<DOC_DEE>Pyrénées (France) -- Descriptions et voyages -- 19e siècle</DOC_DEE>
<DOC_LANGUE>Français</DOC_LANGUE>
<DOC_REF>PHO00019023</DOC_REF>
<DOC_TITRE>Voyage d'artiste : Guide dans les Pyrénées par E.E</DOC_TITRE>
<DOC_TYPE>Œuvre imprimée</DOC_TYPE>
<FT_DATE>2007-04-12</FT_DATE>
<FT_ORIGINAL_SIZE>144590</FT_ORIGINAL_SIZE>
<FT_SFNAME>MIDR_IMPR_99144R.TXT</FT_SFNAME>
<MIDR_SOURCE>Bibliothèque</MIDR_SOURCE>
<PHO_COPYRIGHT>Médiathèque intercommunale Pau-Pyrénées</PHO_COPYRIGHT>
<PHO_FONDS>Médiathèque intercommunale Pau-Pyrénées</PHO_FONDS>
<PHO_LEGEND>Auteurs identifiés par Labarère</PHO_LEGEND>
<USERM>Elisabeth Laulheret</USERM>
<FT_CID>33356</FT_CID>
</NOTICE>
```

FIGURE 7.4 – Notice descriptive exemple réalisée par la MIDR

tamment la polysémie, traitée en spécifiant à chaque fois entre parenthèses le sens que l'on souhaite donner au terme (ex : *Grue (appareil)* et *Grue (animal)*) ; le pluriel avec par exemple le terme *Château(x)*, le singulier *Château* défini le concept « *Château* » alors que son pluriel (*Châteaux*) regroupe un ensemble de châteaux.

Le journal des créations et des modifications [85] informe par ailleurs les utilisateurs, deux fois par an, des enrichissements et des évolutions du langage d'indexation.

RAMEAU est utilisé par les bibliothécaires de la médiathèque de Pau afin d'indexer, via plusieurs termes appelés vedettes, le contenu des documents texte, image, son et vidéo. Une vedette est ici définie par un ou plusieurs terme(s) pour représenter un concept. La vedette peut être de différents types :

- Nom commun RAMEAU (exemples : *Eaux minérales*, *Montagnes*, etc.) ;
- Nom géographique RAMEAU : noms géographiques, villes anciennes ou sites archéologiques (exemple : *Pyrénées (France)*) ;
- Subdivision chronologique RAMEAU (exemple : *19e siècle*) ;
- Nom de personne : personne physique, familles, dieu ou déesse, personnage mythologique, légendaire ou fictif (exemple : *Rousseau, Jean-Jacques (1712-1778)*) ;
- Collectivité : toute organisation ou groupe de personnes ou d'organisations iden-

85. http ://rameau.bnf.fr/utilisation/journal.htm

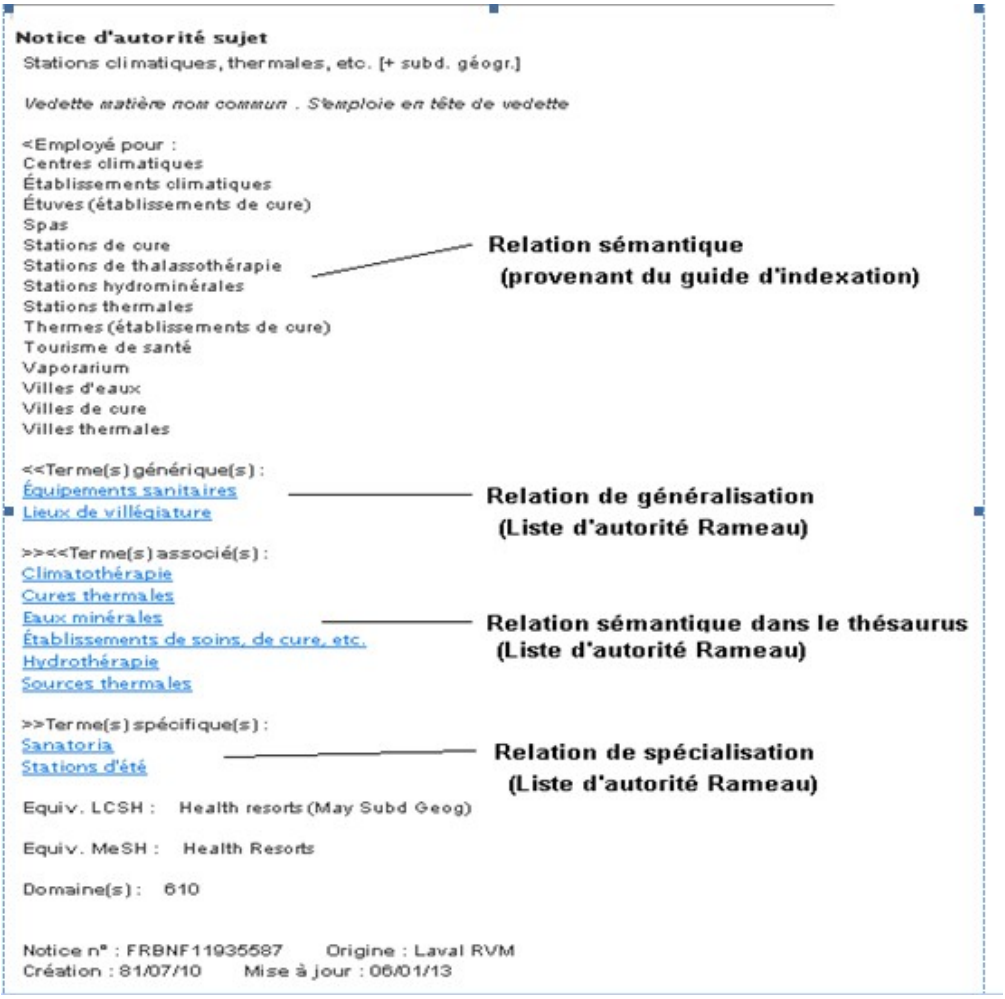


FIGURE 7.5 – Notices RAMEAU de la vedette « Stations climatiques, thermales, etc. »

tifiés par un nom particulier, y compris les groupes ou manifestations temporaires ayant un nom (norme AFNOR : NF Z 44-060) (exemple : *Académie d'agriculture de France*) ;

– Titre : titre d'œuvres, de livres liturgiques et sacrés, de publications en série (exemple : *la trilogie romanesque His dark materials traduite en français À la croisée des mondes français datant de 1995*).

Pour nos traitements, nous devons exploiter le thésaurus RAMEAU initialement formalisé au format UNIMARC [86].

Une étude systématique des balises, de leurs attributs et de leur organisation au sein du document de spécifications au format XML a permis de déterminer comment identifier les vedettes, leurs propriétés ainsi que les relations entre les vedettes. Cette

86. http ://www.bnf.fr/documents/UnimarcA.pdf

analyse ne présente pas de difficulté majeure car la documentation associée au format MARC est riche et véhicule l'ensemble des informations nécessaires à l'extraction de la sémantique et des relations du thésaurus. Une description détaillée est donnée dans la deuxième édition du manuel UNIMARC.

7.2.1.2 Module de constitution du corpus

L'outil Exlibris, qu'utilise la médiathèque pour indexer et stocker l'ensemble du fonds documentaire, a des fonctions d'extraction de contenu des documents très limitées. En effet pour récupérer un jeu d'essai avec les notices associées par exemple, l'outil propose d'extraire les documents répertoire par répertoire. Le problème est que l'on récupère tout d'abord les notices rangées par centaines dans des fichiers Xml et il est alors impossible d'extraire directement les documents image, texte, vidéo et sons correspondant. Après avoir obtenu de la part de la médiathèque un ensemble de notices descriptives ainsi que des répertoires complets de documents texte, image, vidéo et son, nous avons développé un module permettant de définir un jeu de test le plus complet possible contenant les notices Xml de départ et l'ensemble des documents pour lesquelles nous disposons de ces notices (voir figure 7.6).

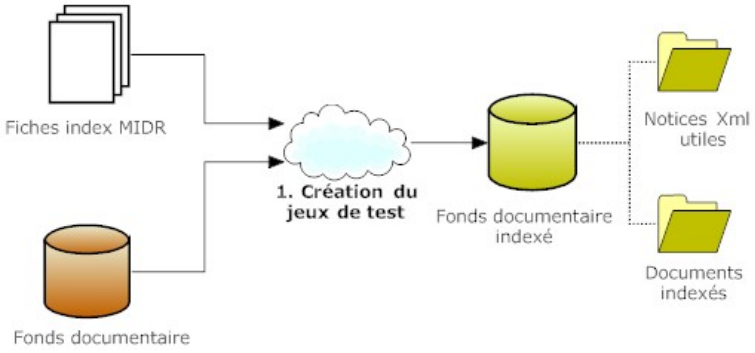


FIGURE 7.6 – Module de création du jeu d'essai

Spécifications

Nous obtenons ainsi en sortie de ce module un jeu de documents de tous types, pour lesquels il existe une notice descriptive, dans une arborescence correspondant à l'arborescence de départ (celle du répertoire de documents source). Les notices descriptives, lorsqu'elles correspondent à un document en notre possession, sont copiées dans un répertoire de notices, le nom de la notice étant le nom du document associé. Pour les 5000 documents mis à disposition par la MIDR, ce module nous permet de sélectionner un ensemble de 900 documents pour lesquels nous pouvons identifier un lien explicite avec une notice descriptive. Ce jeu de documents est constitué de 615 documents image, 50 documents vidéo, 35 documents sonores et 200 documents texte.

Titre du module	CreationFondsDocumentaireTest
Prére-quis	
Entrées	- Répertoire de documents source - Répertoire de documents destination - Répertoire de notices source - Répertoire de notices destination - Chemin vers le fichier contenant les logs du traitement
Sorties	- Répertoire de documents pour lesquels une notice Xml est identifiée - Répertoire de notices correspondantes aux documents sélectionnés
Fonctions	- Création d'un fonds de documents contenant des documents pour lesquels il existe une notice Xml associée - Décomposition des fichiers Xml de notices de départ en fichiers notices unitaires : un fichier –> une notice

7.2.2 Module (1) d'extraction de la connaissance experte : création du vocabulaire vocTerridoc

Les experts de la MIDR indexent les documents sans apporter d'éléments d'information sur la forme (typographie, etc.). De plus, la structure logique des notices est quasiment inexistente comme nous pouvons le voir figure 7.4 (cf. page 141). En effet, tous les champs donnant des informations sur le document décrit sont au premier niveau de la notice, ce qui ne permet pas d'identifier des relations hiérarchiques ou associatives. Le travail de description du contenu des documents est stocké dans les balises *doc_dee*.

A partir de la base de notices Xml, nous définissons le module schématisé figure 7.7. Le but est d'extraire l'ensemble des termes définis dans les balises *doc_dee* en s'appuyant sur le thésaurus RAMEAU afin de définir un premier vocabulaire décrivant le fonds documentaire que nous nommons vocTerridoc.

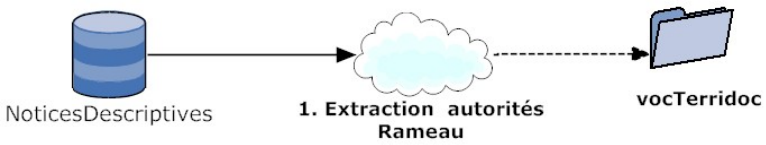


FIGURE 7.7 – Définition de la liste de termes

Spécifications

Nous reprenons les extraits de notices présentés chapitre 6.1 page 103 (cf. figures 6.2 et 6.3 page 105), pour décrire la phase d'implémentation des différentes étapes de notre méthodologie. Nous obtenons dans cette étape d'extraction de la connaissance experte

Titre du module	CreationFondsDocumentaireTest
Prére-quis	
Entrées	- Répertoire de notices source - Chemin vers le fichier contenant les logs du traitement
Sorties	- Vocabulaire de termes Rameau au format Skos
Fonctions	- Création d'un vocabulaire de termes caractérisant sémantiquement le fonds documentaire indexé par la MIDR ; - Liaison des documents aux termes les caractérisant ; - Regroupement des termes en fonction de la casse, de l'accentuation et du singulier/pluriel - Transposition de ce vocabulaire au format Skos pour les enrichissements futurs

un premier vocabulaire **vocTerridoc, consitué de 232 termes**, dont un extrait est présenté figure 7.8.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <vocTerridoc>
3  <terme>Barèges (Hautes-Pyrénées)</terme>
4  <terme>Station climatique, thermale, etc.</terme>
5  <terme>Eaux minérales</terme>
6  <terme>Bagnères-de-bigorre (Hautes-Pyrénées)</terme>
7  <terme>Pyrénées (France)</terme>
8  <terme>Oeuvres scientifiques</terme>
9  <terme>18e siècle</terme>
10 <terme>19e siècle</terme>
11 ....
12 </vocTerridoc>
```

FIGURE 7.8 – Extrait des termes provenants des notices descriptives

Ce vocabulaire est une première étape vers la définition de l'ontologie d'un territoire. A ce niveau, aucun traitement n'est effectué pour identifier les noms toponymiques ainsi que les entités temporelles. Les termes sont tous considérés de la même façon pour la structuration en thésaurus. Nous décrivons dans l'étape suivante l'implémentation du module permettant d'identifier l'ensemble des relations entre ces termes pour développer notre vocabulaire en thésaurus tTERRIDOC.

7.2.3 Module (2) de structuration des termes : normalisation en thésaurus tTerridoc

Le module présenté figure 7.9 nous permet de définir dans un premier temps les relations entre tous les termes provenant du vocabulaire VocTerridoc en s'appuyant sur

le thésaurus RAMEAU. Ensuite, nous considérons chaque terme du vocabulaire vocTerridoc comme élément de bas niveau car rattaché directement à des documents et nous enrichissons le thésaurus avec les termes vedettes plus génériques du thésaurus RAMEAU en ajoutant dès que nécessaire les relations de type générique, spécifique et associative entre ces nouveaux termes. Le but est d'obtenir une représentation sémantique de plusieurs niveaux d'abstractions. Nous formalisons le thésaurus en langage SKOS -*cf.* voir figure 7.10).

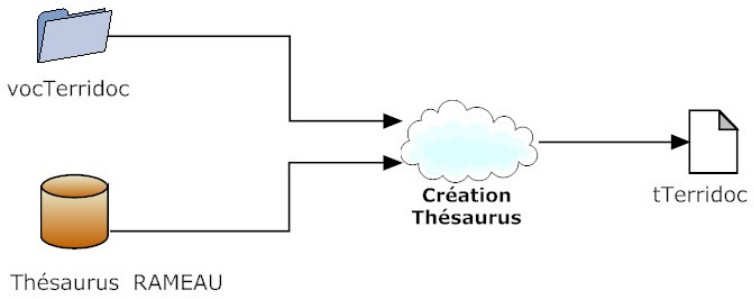


FIGURE 7.9 – Enrichissement du vocabulaire (création du thésaurus)

Nous appliquons ci-dessous les cinq patrons définis dans le processus de structuration du vocabulaire VocTerridoc en thésaurus tTerridoc (cf. définition de la méthodologie section 6.1.4.1 page 108) : (1) gestion des termes vedettes ; (2) gestion des termes « employé pour » ; (3) gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc ; (4) gestion des relations associatives entre termes du vocabulaire vocTerridoc ; (5) Structuration du thésaurus via les relations génériques.

7.2.3.1 Gestion des termes « vedettes »

Cette étape permet de créer un premier ensemble de vedettes qui sert ensuite de base pour construire le thésaurus tTerridoc. Nous définissons en début de structure un schéma (via *skos :ConceptScheme*) nommé tTerridoc qui regroupe l'ensemble des éléments (termes et relations) permettant de définir le thésaurus *tTerridoc* à partir du travail d'indexation des experts. Nous appliquons ensuite le patron défini tableau 6.1 (c.f. page 109). Pour chaque unité textuelle *UT* extraite du vocabulaire vocTerridoc qui correspond à une vedette dans le thésaurus RAMEAU, une vedette est définie en utilisant la balise SKOS *skos :Concept* avec comme identifiant (via l'attribut *rdf :about*) le nom du terme en remplaçant le caractère espace *« »* par le caractère *« _ »* . Un terme dans le formalisme SKOS peut avoir par langue un label préférentiel (*skos :prefLabel*) et plusieurs labels alternatifs (*skos :altLabel*). Dans cette première étape, nous définissons pour chacune des autorités uniquement le label préférentiel qui a pour valeur le nom du terme extrait de la notice descriptive. Dans l'exemple figure 7.10, le terme portant l'identifiant *Eaux_minérales* a pour label préférentiel « *Eaux minérales* ».

Nous utilisons la propriété *skos :example* pour définir un lien vers le document correspondant. Les termes représentent alors le niveau conceptuel et les documents le niveau

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
...
<skos:ConceptScheme rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc">
    <dc:title>Le Thesaurus tTerridoc</dc:title>
    <dc:description>Description du thésaurus tTerridoc pour le fonds de la MIDR</dc:description>
    <dc:creator>LIUPPA</dc:creator>
</skos:ConceptScheme>
    <skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Bareges_(Hautes-Pyrenees)">
        <skos:prefLabel>Barèges (Hautes-Pyérénées)</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:example rdf:ressource="http:/t2i.ujniv-pau.fr/docs/doc1.jpg"/>
    </Concept>
    <skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc.">
        <skos:prefLabel>Stations climatiques, thermales, etc.</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:example rdf:ressource="http:/t2i.ujniv-pau.fr/docs/doc1.jpg"/>
        <skos:example rdf:ressource="http:/t2i.ujniv-pau.fr/docs/doc2.pdf"/>
    </Concept>
    <skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Bagneres-de-bigorre_(Hautes-Pyrenees)">
        <skos:prefLabel>Bagnères-de-bigorre (Hautes-Pyérénées)</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:example rdf:ressource="http:/t2i.ujniv-pau.fr/docs/doc2.jpg"/>
    </Concept>
    <skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Eaux_minerale">
        <skos:prefLabel>Eaux minérales</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:example rdf:ressource="http:/t2i.ujniv-pau.fr/docs/doc1.jpg"/>
    </Concept>
...
</rdf:RDF>
```

FIGURE 7.10 – Extrait de la liste de termes formalisée en SKOS

physique. Par exemple, le document *doc2.pdf* décrit par la notice descriptive figure 6.3 sera relié aux termes *œuvres scentifiques*, *Stations climatiques, thermales, etc.*, *Bagnères-de-Bigorre (Hautes-Pyrénées)* et *19e siècle*.

A partir des 232 termes constituant le vocabulaire vocTerridoc, **35** sont présents en tant que vedettes dans RAMEAU et deviennent donc des vedettes dans le thésaurus tTerridoc. 197 termes sont à cette étape laissés de côté.

7.2.3.2 Gestion des termes « employé pour »

Cette étape, réalisée en simultané avec l'étape traitant les termes vedettes, permet d'appliquer le patron structurel défini tableau 6.2 (c.f. page 110) afin de prendre en compte les termes de vocTerridoc qui ne sont pas vedettes dans le thésaurus RAMEAU. Pour chacun des termes t1 du vocabulaire vocTerridoc qui est un « *terme rejeté* » à une vedette dans RAMEAU, si la vedette correspondante n'existe pas dans tTerridoc, elle est créée avec pour label préférentiel le nom de la vedette et un label alternatif lui est ajouté avec pour valeur le terme rejeté t1. Si la vedette est présente dans tTerridoc, le label alternatif avec pour valeur t1 est ajouté si il n'existe pas déjà. Dans l'extrait

du vocabulaire vocTerridoc (cf. figure 7.8), le terme « *villes thermales* » est un terme rejeté de la vedette « *Stations climatiques, thermales, etc.* » dans RAMEAU (cf. figure 7.5) et il est donc ajouté en tant que label alternatif à la vedette « *Stations climatiques, thermales, etc.* » (voir figure 7.11).

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
.....
    <skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc.">
        <skos:prefLabel>Stations climatiques, thermales, etc.</skos:prefLabel>
        <skos:altLabel>Villes thermales</skos:altLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
    .....
    </Concept>
.....
</rdf:RDF>
```

FIGURE 7.11 – Prise en compte des termes rejetés selon le formalisme SKOS

Parmi l'ensemble des termes constituant le vocabulaire vocTerridoc (232 termes), **70 termes** sont des termes rejetés dans RAMEAU. **6 termes** sont directement attachés aux vedettes existantes dans le thésaurus tTerridoc et le reste (64 termes) nécessite la création d'une vedette dans tTerridoc. Dans ce second cas, les termes sont ensuite attachés à la vedette créée en tant que label alternatif. A la suite de cette étape, 127 termes sont mis de côté faute de pouvoir identifier une correspondance avec la ressource RAMEAU. Parmi ces termes, nous identifions un nombre important de noms de personnes et de lieux-dits.

L'ensemble des termes étant traités, nous décrivons maintenant par type de relation l'étape de structuration du thésaurus.

7.2.3.3 Gestion des relations « génériques » et « spécifiques » entre termes du vocabulaire vocTerridoc

Nous appliquons ici la règle (c.f. tableau 6.3 page 111) permettant d'enrichir la structure avec les relations hiérarchiques lorsque deux termes du vocabulaire vocTerridoc existent dans RAMEAU en tant que vedette et qu'ils sont liés par une relation hiérarchique. A partir de l'extrait du thésaurus tTerridoc formalisé en SKOS figure 7.10, le thésaurus RAMEAU nous permet d'identifier une relation hiérarchique entre les vedettes *Barèges_(Hautes-Pyrénées)* et *Stations climatiques, thermales, etc.* indiquant que la première vedette est fils de la seconde. Une relation de type hiérarchique est donc créée entre ces deux vedettes en utilisant les éléments *skos :narrowerGeneric* et *skos :browderGeneric* (voir figure 7.12).

7.2.3.4 Gestion des relations associatives entre termes du vocabulaire vocTerridoc

Dans cette étape, nous appliquons la règle présentée tableau 6.4 (c.f. page 111) qui permet d'enrichir la structure avec les relations associatives lorsque deux termes du vo-

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
......
<skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Bareges_(Hautes-Pyrenees)">
        <skos:prefLabel>Barèges (Hautes-Pyérénées)</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:browderGeneric rdf:resource="http:/t2i.ujniv-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc."/>
    </Concept>

    <skos:Concept rdf:about="http:/t2i.univ-pau.fr/Concept/Stations_climatiques,_thermales,_etc.">
        <skos:prefLabel>Stations climatiques, thermales, etc.</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        <skos:narrowerGeneric rdf:resource="http:/t2i.ujniv-pau.fr/tTerridoc/Bareges_(Hautes-Pyrenees)"/>
    ......
    </Concept>
......
</rdf:RDF>
```

FIGURE 7.12 – Extrait SKOS du thésaurus tTerridoc intégrant la relation hiérarchique

cabulaire vocTerridoc existent dans RAMEAU en tant que vedettes et qu'ils sont liés par une relation de type « terme associé ». A partir de l'extrait du thésaurus tTerridoc présenté figure 7.10 et du thésaurus RAMEAU, une relation de type associatif est identifiée entre les vedettes *Stations_climatiques,_thermales,_etc.* et *Eaux_minérales*, ce qui nous permet de créer une relation associative entre ces deux vedettes dans le thésaurus tTerridoc. En SKOS, nous formalisons ce lien en utilisant l'élément *skos :related* (voir figure 7.13).

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
......
    <skos:Concept rdf:about="http:/t2i.univ-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc.">
        <skos:prefLabel>Stations climatiques, thermales, etc.</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        ......
        <skos:related rdf:resource="http:/t2i.univ-pau.fr/tTerridoc/Eaux_minerale"/>
    </Concept>
    <skos:Concept rdf:about="http:/t2i.univ-pau.fr/tTerridoc/Eaux_minerale">
        <skos:prefLabel>Eaux minérales</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        ......
        <skos:related rdf:resource="http:/t2i.univ-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc."/>
    </Concept>
......
</rdf:RDF>
```

FIGURE 7.13 – Extrait SKOS du thésaurus tTerridoc intégrant la relation associative

Ces deux étapes de structuration des vedettes sous forme de thésaurus tTerridoc nous permet d'identifier **117 relations de type hiérarchique** et **22 de type associative**. Nous obtenons à ce niveau un ensemble de petites structures sémantiques qui ne sont pas reliées entre elles. L'étape suivante dans le traitement consiste à identifier et ajouter

à la structure l'ensemble des termes génériques aux vedettes provenant du vocabulaire vocTerridoc afin de créer une structure thésaurus unique.

7.2.3.5 Structuration du thésaurus via les relations génériques

Nous appliquons le patron structurel présenté figure 6.8 page 113 afin de construire de façon récursive une structure de plusieurs niveaux hiérarchiques en partant des vedettes constituant le premier thésaurus tTerridoc et en exploitant les relations de généricité entre vedettes lorsqu'elles existent dans RAMEAU. Si nous reprenons l'exemple de la vedette *Stations_climatiques,_thermales,_etc.*, nous remarquons dans la fiche de la vedette (cf.figure 7.5 page 142) qu'elle possède deux vedettes génériques dans RAMEAU, à savoir « *Equipements sanitaires* » et « *Lieux de villégiature* ». Nous présentons figure 7.14 un exemple de cet enrichissement du thésaurus tTerridoc en utilisant les balises *skos :narrowerGeneric* et *skos :browderGeneric*.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
.....
<skos:Concept rdf:about="http:/t2i.ujniv-pau.fr/tTerridoc/Lieu_de_villegiature">
        <skos:prefLabel>Lieux de villégiature</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        .....
        <skos:narrowerGeneric rdf:resource="http:/t2i.univ-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc."/>
    </Concept>

    <skos:Concept rdf:about="http:/t2i.univ-pau.fr/tTerridoc/Stations_climatiques,_thermales,_etc.">
        <skos:prefLabel>Stations climatiques, thermales, etc.</skos:prefLabel>
        <skos:inScheme rdf:resource="http:/t2i.univ-pau.fr/tTerridoc"/>
        .....
        <skos:browderGeneric rdf:resource="http:/t2i.univ-pau.fr/tTerridoc/Lieux_de_villegiature"/>
    </Concept>
.....
</rdf:RDF>
```

FIGURE 7.14 – Structuration du thésaurus tTerridoc via les vedettes génériques

Cette dernière étape de structuration nous permet d'obtenir, à partir du corpus constitué de 900 documents indexé, un thésaurus tTerridoc constitué de **540 vedettes** (formant un vocabulaire de 3502 termes/labels si l'on prend l'ensemble des termes rejetés qui leur sont attachés). Les étapes de structuration nous permettent d'identifier **80 relations associatives** entre vedettes et **467 relations hiérarchiques**. L'étape de hiérarchisation nous permet de faire émerger une structure sur **32 niveaux** avec comme feuilles les vedettes provenant du vocabulaire vocTerridoc. Un extrait du thésaurus tTerridoc, édité via Protégé, est proposé figure 7.15. Notre méthode se veut générique et donc applicable à différents fonds documentaires et vocabulaires contrôlés associés. Cependant, nous sommes conscients que la qualité et la précision du thésaurus tTerridoc obtenu dépend en grande partie de l'exhaustivité de la ressource vocabulaire contrôlé utilisée.

Dans les phases d'extraction et de structuration des vedettes présentes dans les no-

tices descriptives, nous ne traitons pas les constructions syntaxiques spécifiques au travail d'indexation des experts bibliothécaires (têtes de vedettes, etc.) car les spécificités liées à ces constructions syntaxiques ne permettent pas de définir des patrons génériques et applicables à différents fonds indexés. Aussi SKOS est par définition un schéma de formalisation « simple » et ne permet pas d'expliciter ces constructions syntaxiques. La convention de nommage des URI, élément essentiel à la structure SKOS pour permettre un référencement précis de chacun des concepts identifiés dans RAMEAU, n'est pas encore stabilisée. L'idée d'un identifiant unique pour l'accès aux concepts n'est pas encore bien définie. Nous comptons cependant nous appuyer sur ARK (Archival Resource Key), un dispositif d'URL persistante (fr). Actuellement nous proposons des URL de la forme suivante : http :/t2i.univ-pau.fr/concept/Barèges_(Hautes-Pyrénées). Nous obtenons dans cette étape un thésaurus offrant une représentation structurée du travail d'indexation des experts. Le thésaurus tTerridoc obtenu sous forme de fichier SKOS est testé et analysé à l'aide de l'éditeur Protégé [87] auquel nous ajoutons le plugin SKOSEd [88] pour ouvrir et éditer des fichiers Skos (c.f. figure 7.15).

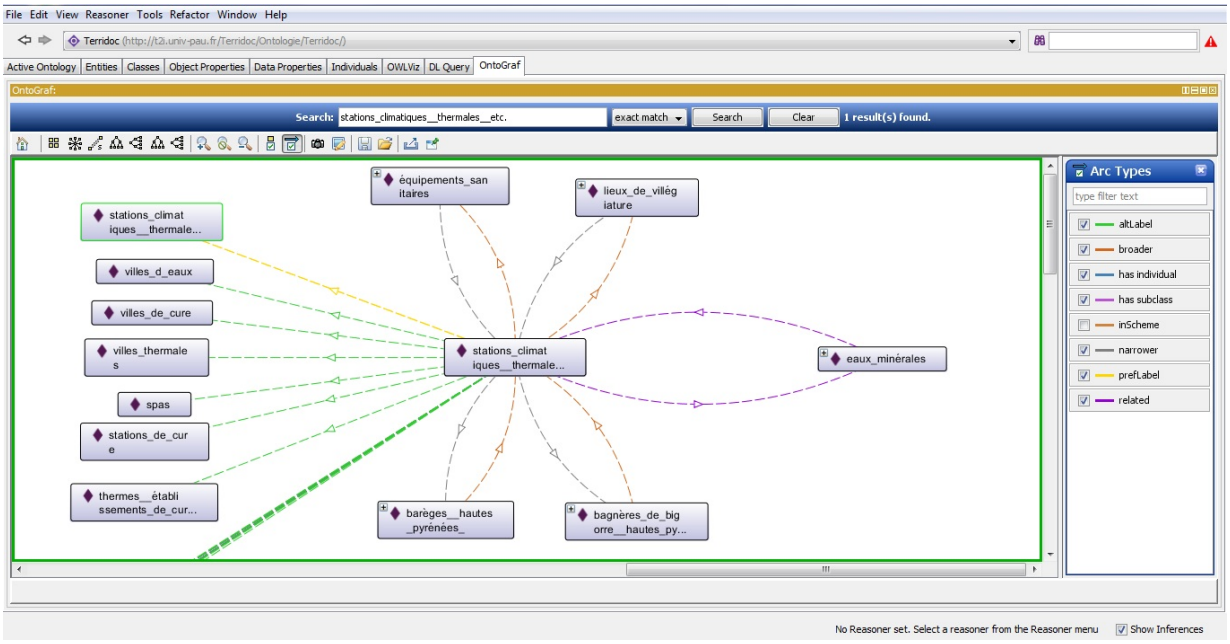


FIGURE 7.15 – Visualisation sous Protégé d'un extrait du thésaurus tTerridoc

Nous présentons maintenant la phase de transformation de ce thésaurus en une première ontologie d'un territoire. Nous utilisons comme lexiques toponymiques les bases BDTopo et BDCarto de l'IGN.

87. http ://protege.stanford.edu/ version 4.0
88. http ://protegewiki.stanford.edu/wiki/SKOS_Editor version 1.0.3

7.2.4 Module (3) de transformation du thésaurus tTerridoc en une première ontologie légère

Comme indiqué section 6.1.4.2 page 114, la transformation du thésaurus tTerridoc en ontologie d'un territoire se décompose en 4 actions : (i) regroupements des termes en concepts ; (ii) structuration via les relations hiérarchiques ; (iii) structuration via les relations associatives ; (iv) instanciation de l'ontologie pour la valorisation d'un territoire. L'ontologie générée est formalisée en OWL-Lite et nous utilisons actuellement l'éditeur Protégé pour en analyser la structure.

7.2.4.1 Regroupement des termes en concepts

La première étape consiste à regrouper les vedettes et labels qui ont le même sens dans tTerridoc en un seul concept dans l'ontologie. Cette étape intègre une phase de lemmatisation pour traiter les vedettes au pluriel et au singulier comme présenté section 1 page 114. tTerridoc est défini sur la base du thésaurus RAMEAU et ne contient donc pas de termes utilisés dans plusieurs groupes de termes décrivant chacun un sens spécifique. La polysémie est aussi traitée en spécifiant à chaque fois entre parenthèses le sens que l'on souhaite donner au terme (ex : *Grue (appareil)* et *Grue (animal)*). Il n'est donc pas nécessaire ici d'effectuer des traitements supplémentaires pour traiter ces cas. Nous présentons en exemple figure 7.16 le passage de la vedette « *Stations climatiques, thermales, etc.* » présente dans tTerridoc en concept dans l'ontologie.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#">

    <owl:Ontology rdf:about="http:/t2i.ujniv-pau.fr/Terridoc">
      <dc:title>Ontologie d'un territoire TERRIDOC_MIDR</dc:title>
       <dc:description>Ontologie décrivant un territoire décrit par le fonds documentaire de la MIDR</dc:description>
    </owl:Ontology>

    <owl:Class rdf:about="Station_climatique,_thermale,_etc.">
        <rdfs:label xml:lang="fr">Station climatique, thermale, etc.</rdfs:label>
        <rdfs:label xml:lang="fr">Station de cure</rdfs:label>
    </owl:Class>
</rdf:RDF>
```

FIGURE 7.16 – Formalisation du concept Station_climatique,_thermale,_etc en OWL-Lite

Le concept créé c1 a pour identifiant le nom de la vedette au singulier (ici *Station_climatique,_thermales,_etc.*) afin de garder un lien explicite entre la ou les vedettes du thésaurus et le concept c1. Nous utilisons la balise *rdfs :label* pour définir l'ensemble des noms portant le même sens. L'attribut lang permet de gérer plusieurs langues mais l'ontologie générée ne prend en compte que du français.

A partir du thésaurus tTerridoc obtenu 7.9 page 146, nous obtenons une liste de **507 concepts**. La gestion du pluriel permet ici de regrouper 540 vedettes en concepts.

7.2.4.2 Structuration via les relations hiérarchiques

Nous appliquons ici la règle (cf. section 2 pge 116) visant à structurer les concepts en exploitant les relations hiérarchiques présentes dans tTerridoc. Rappelons que pour chaque relation hiérarchique identifiée entre deux vedettes a1 et a2 du thésaurus, une relation de type « sous-classe » est créée entre les concepts relatifs à a1 et a2. Nous appliquons dans cette étape la règle de transitivité permettant de supprimer des redondances dans les relations hiérarchiques présentes dans le thésaurus. Dans cette étape, **371 relations « est un »** sont créées. A partir de l'extrait du thésaurus tTerridoc présenté figure 7.14 page 150, nous définissons en OWL une relation de type « sous-classe » entre les concepts *Station_climatique,_thermales,_etc.* et *Lieu_de_villegiature*.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#">
......
    <owl:Class rdf:about="Station_climatique,_thermale,_etc.">
        <rdfs:subClassOf rdf:resource="#Lieu_de_villegiature"/>
        <rdfs:label xml:lang="fr">Station climatique, thermale, etc.</rdfs:label>
        <rdfs:label xml:lang="fr">Station de cure</rdfs:label>
    </owl:Class>

    <owl:Class rdf:about="Lieu_de_villegiature">
        <rdfs:label xml:lang="fr">Lieu de villégiature</rdfs:label>
    </owl:Class>
......
</rdf:RDF>
```

FIGURE 7.17 – Formalisation du concept Barèges_(Hautes-Pyrénées) en OWL-Lite

Nous cherchons ensuite à identifier et ajouter des relations associatives (de type « associé à ») entre concepts de l'ontologie.

7.2.4.3 Structuration via les relations associatives

Nous appliquons la règle présentée section 3 page 117 qui consiste à définir dans l'ontologie les relations associatives provenant du thésaurus tTerridoc. Pour chaque relation associative entre deux termes de tTerridoc, nous définissons une relation associative dans l'ontologie entre les concepts dont ils sont labels. Nous obtenons à cette étape **63 relations associatives**.

Nous obtenons à cette étape une première ontologie légère offrant une représentation sémantique du fonds documentaire traité constituée de 507 concepts organisées selon 371 relations hiérarchiques et 63 relations associatives. Elle permet également de formaliser de de façon synthétique le travail d'indexation des experts. Cependant, la structure sémantique obtenue ne valorise pas les spécificités géographiques liées à un fonds documentaire territorialisé comme celui sur lequel nous travaillons et elle ne permet donc pas de représenter un territoire. Nous présentons dans le paragraphe suivant une applica-

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#">

    <owl:ObjectProperty rdf:ID="associe_A_1">
        <rdfs:domain rdf:resource="#Station_climatique,_thermale,_etc."/>
        <rdfs:range  rdf:resource="#Eau_minerale"/>
    </owl:ObjectProperty>

    <owl:Class rdf:about="Eau_minerale">
        <rdfs:label xml:lang="fr">Eaux minérales</rdfs:label>
    ......
    </owl:Class>
    <owl:Class rdf:about="Station_climatique,_thermale,_etc.">
        <rdfs:label xml:lang="fr">Station climatique, thermale, etc.</rdfs:label>
    ......
    </owl:Class>
......
</rdf:RDF>
```

FIGURE 7.18 – Formalisation du concept Barèges_(Hautes-Pyrénées) en OWL-Lite

tion de notre démarche automatisée permettant d'identifier des éléments décrivant un territoire et de les attacher ensuite à l'ontologie légère que nous venons de générer.

7.2.5 Module (4) d'instanciation de l'ontologie pour l'émergence d'un territoire

En nous appuyant sur les bases de données BDTopo et BDCarto fournies par l'IGN, une première étape nous permet d'identifier dans l'ontologie obtenue dans l'étape précédente l'ensemble des concepts qui sont des entités spatiales. Nous appliquons la règle exposant que pour chaque concept *C2* relié à un concept *C1* (identifié comme un nom_toponymique), par la relation hiérarchique indiquant que *C1 est « sous classe » de C2*, le concept C1 est supprimé, une instance *I1* est créée et la relation de type « sous classe de » dans laquelle le concept *C1* intervenait est redéfinie en relation de type « instance de » indiquant que *C1 est une « instance de » C2*. Une propriété *geometrie* est ajoutée au concept *C2* afin d'indiquer la représentation spatiale pour chacune de ses instances. Les autres concepts restent inchangés.

Par exemple, le label « Bagnères-de-Bigorre (Hautes-Pyrénées) » du concept *Bagnères-de-Bigorre_(Hautes-Pyrénées)* et« Barèges (Hautes-Pyrénées) » du concept *Barèges_(Hautes-Pyrénées)*, sont présents dans les bases de données et les deux concepts sont donc identifiés comme des entités spatiales. Ces deux concepts étant sous-classes de *Station_climatique,_thermale,_etc.* dans l'ontologie, ils sont supprimés et remplacés par des instances de même nom (figure 7.19), devenant ainsi instance du concept *Station_climatique,_thermale,_e*
La propriété *geometrie* ajoutée au concept *Station_climatique, _thermale,_etc.* permet

d'indiquer la représentation spatiale pour chacune de ses instances.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#">
  ......
<owl:DatatypeProperty rdf:ID="geometrie">
        <rdfs:domain rdf:resource="#Station_climatique_thermale_etc."/>
        <rdfs:subProperty0f rdf:resource="&higgins;simpleAttribute"/>
    </owl:DatatypeProperty>

<Station_climatique_thermale_etc. rdf:ID="Bagneres-de-Bigorre_(Hautes-Pyrénées)">
    <Station_climatique_thermale_etc.:geometrie rdf:ID="geom_Bagneres-de-Bigorre">43.0642,0.15</Station_climatique_thermale_etc.:geometrie>
</Station_climatique_thermale_etc.>
<Station_climatique_thermale_etc. rdf:ID="Bareges_(Hautes-Pyrénées)">
    <Station_climatique_thermale_etc.:geometrie rdf:ID="geom_ Bareges">42.8967,0.068</Station_climatique_thermale_etc.:geometrie>
</Station_climatique_thermale_etc.>
  ......
</rdf:RDF>
```

FIGURE 7.19 – Formalisation des entités spatiales en OWL

Nous obtenons après cette étape d'instanciation une première ontologie d'un territoire à partir du travail d'indexation d'experts. Cette première ontologie est constituée de **481 concepts** reliés à **26 instances** correspondant à des entités géographiques. Le résultat obtenu nous donne un premier aperçu des concepts décrits dans un espace géographique. Nous remarquons cependant que le nombre d'entités spatiales ne nous permet pas d'obtenir une description précise de l'espace décrit dans le fonds documentaire traité. Un aperçu de l'ontologie au format OWL-Lite est présenté figure 7.20.

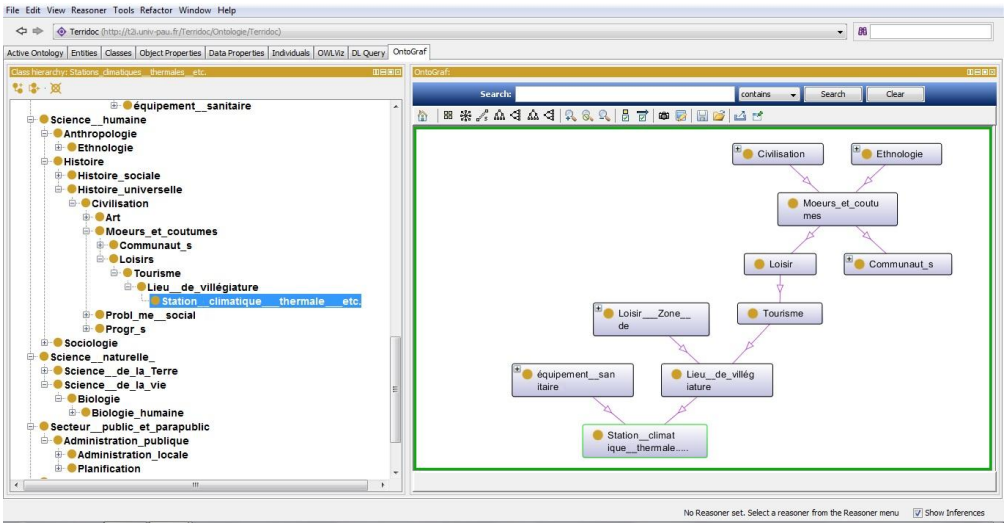


FIGURE 7.20 – Visualisation sous Protégé d'un extrait de l'ontologie formalisée en OWL

Nous proposons d'appliquer notre chaîne de traitement linguistique développée sous Linguastream afin d'identifier de nouvelles entités spatiales dans le reste des notices descriptives ainsi que dans le contenu des documents.

7.3 TALN pour la représentation d'un territoire

Reprenons la chaîne de traitement (*cf.* figure 6.12 page 125) de construction d'index particulièrement adaptée à l'aspect géographique des contenus [BGKS10]. Au sein de cette chaîne, nous nous appuyons sur l'annotation automatique d'informations géographiques consituées d'une Entité Nommée (EN) spatiale (un nom toponymique décrivant un espace) et dans le cas où elle est identifiable d'une EN temporelle calendaire décrivant une période. La chaîne de traitement est composée de cinq grandes phases [AFG03] : (a) la lemmatisation pour segmenter les mots ; (b) l'analyse lexicale et morphologique pour la reconnaissance des mots ; (c) l'analyse syntaxique, basée sur des grammaires, afin de trouver les relations entre les mots ; (d) enfin l'analyse sémantique pour réaliser une interprétation plus spécifique sur les syntagmes retenus ; (e) la validation des entités identifiées par des ressources géographiques (thèmes ou types, communes, lieux-dits, routes, pics, vallées, etc.). Résultent de ce traitement linguistique un ensemble d'entités validées et des relations existantes entre ces dernières.

Voici figure 7.21 une capture d'écran de notre chaîne de traitement implémentée sous Linguastream.

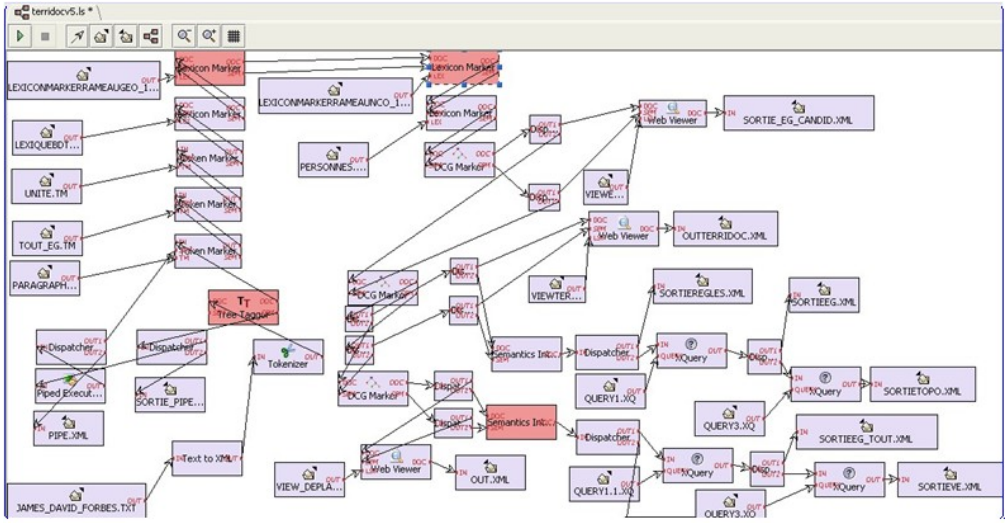


FIGURE 7.21 – Chaîne de traitement linguistique implémentée sous Linguastream

La chaîne de traitement mise en place intègre les éléments suivants :

- Module Tokenizer : ce service prend en charge la segmentation du texte. Il prend en entrée le texte et produit un autre flux dans lequel chaque mot est identifié et isolé dans une balise ;
- Module Tree-tagger : ce service se charge de l'analyse morpho-syntaxique du texte ;
- Module Token Marker : détection de motifs à partir d'expressions régulières. Ici pour nous, ce sont les entités géographiques candidates (entités à minima constituée d'une entité thématique ancrée sur une composante spatiale de type nom toponymique). Il prend en entrée le flux XML et un fichier de ressources contenant

les expressions régulières (au format .tm) qui définissent les patrons à détecter. Le flux de sortie est augmenté de balises pour les textes correspondant à ces patrons. Voici une règle, formalisée en XML mise en place dans notre chaîne de traitement pour marquer les noms toponymiques :

```
<?xml version="1.0" encoding="UTF-8"?>
<tokenMarker>
    <rule caseInsensitive="true">
        <pattern>[A-Z]{1}.+ </pattern>
        <featureSet>
                <egn>oui</egn>
        </featureSet>
    </rule>
</tokenMarker>
```

FIGURE 7.22 – Règle de marquage des noms toponymiques

– Module LexiconMarker : permet d'identifier dans le texte l'ensemble des mots et groupes de mots intégrés au lexique. Il donne la possibilité de typer une occurrence d'un mot par un autre mot renvoyant au même sens. Par exemple, en utilisant le lexique provenant du thésaurus RAMEAU, « *Gouffre* » identifié dans le texte sera marqué par le représentant « *Grottes* ». Nous utilisons notamment le fichier lexiconMarkerRameauNco_150609 reprenant l'ensemble des noms communs du thésaurus RAMEAU transformé au format lsl pour traitement sous Linguastream ;
– Module DCG Marker : ce service a pour but de détecter les relations qui existent entre les composantes d'entités géographiques candidates. Il s'appuie sur des grammaires DCG (implémenté à l'aide du langage Prolog dans les fichiers .pro). Ces grammaires permettent de s'appuyer sur les mécanismes d'inférence et d'unification de Prolog à l'aide de règles simples. Il nous permet ici d'identifier les expressions composées d'une entité spatiales candidates et d'une entité thématique contenant un nom commun (qu'il provienne du lexique RAMEAU ou non).
Voici un extrait figure 7.23 de règles Prolog permettant de capter les entités géographiques.

```
%ESA
...
es(es_a:X) --> es_a(X).

es_a(nom_toponymique:N..type:inconnu) --> prepOUprepartOUart, egn(N).
es_a(nom_toponymique:N..type:inconnu) --> egn(N).
...

%entite simple
...
egn(N) --> A@egn:oui, B@egn:oui, {string_concat(A,' ',W)}, {string_concat(W,B,N)}.
egn(N) --> N@egn:oui.
....
```

FIGURE 7.23 – Règle de grammaire DCG pour une analyse sémantique

La figure 7.24 décrit un deuxième extrait de patrons lexico syntaxiques défini en

Prolog permettant d'identifier une relation hiérarchique de type « instance_of ».

```
%marquage des ENs
egsimple(concept:X_es_a:Y) --> syntagme(X), rameaugeo(Y).
egsimple(concept:X_es_a:Y) --> syntagme(X),sep, rameaugeo(Y).
egsimple(concept:X_Y) --> syntagme(X), ls_token(_es:Y,es).
egsimple(concept:X_Y) --> syntagme(X),sep, ls_token(_es:Y,es).

%Marquage des termes provenant de notre première ontologie minimale (intégré dans notre chaîne
sous forme de lexique
syntagme(categorie:S_vedette:T_adjectif:I) -> adjectifs(I), ls_token(_categorie:S_vedette:T,
rameaunco).
syntagme(_categorie:S_vedette:T_adjectif:I) ->ls_token(_categorie:S_vedette:T, rameaunco),
adjectifs(I).
syntagme(categorie:S_vedette:T_adjectif:null)-->ls_token(_categorie:S_vedette:T,rameaunco).
```

FIGURE 7.24 – Patrons lexico-syntaxique pour capter les entités géographiques

– Modules XQuery : lors du traitement sémantique, chaque token est identifié par un numéro de paragraphe et un identifiant (dans le fichier doc). Ce module vient alors compléter le flux XML des identifiants des entités géographiques candidates. Il permet d'obtenir un flux XML valide par rapport au schéma défini (grâce aux fichiers.xslt donnés en entrées).

Concernant l'étape de validation des entités spatiales constituant les entités géographiques, nous utilisons pour notre part les bases de données BD-Topo et BD-Carto mises à disposition par l'IGN. Les patrons lexico-syntaxiques définis puis intégrés aux modules DCG Marker pour repérer des relations sémantiques (c.f. tableaux 6.6 et 6.7 page 116) exploitent les étiquettes morpho-syntaxiques ou sémantiques attribuées par Linguastream.

7.3.1 Modules (5 et 6) pour l'enrichissement de la représentation d'un territoire à partir du travail d'indexation d'experts

Nous appliquons notre chaîne de traitement aux notices descriptives pour obtenir un ensemble de représentations territoriales de l'ensemble du fonds. La figure 7.25 décrit un extrait de la sortie de notre chaîne de traitement.

L'extrait décrit figure 7.25 présente une liste d'entités géographiques marquées dans les notices descriptives parmi lesquelles nous pouvons citer « établissements thermaux des Pyrénées », « eaux de Barèges », « bains de Barège », etc.

Comme le montre la figure 7.26 présentant un bilan de l'exécution de notre chaîne de traitement sur les 900 notices descriptives constituant notre fonds documentaire, nous obtenons 266 entités géographiques dont 115, reliées à des labels de concepts de l'ontologie de départ, peuvent être ajoutées à l'ontologie pour enrichir la représentation du territoire des Pyrénées.

Cette étape nous permet d'enrichir la représentation du territoire en augmentant de façon intéressante le nombre d'entités géographiques identifiées (26 entités initialement à 141 entités en exploitant les notices). Seules, ces ressources permettent de définir une structure sémantique pour décrire notre territoire. Cependant, nous pensons pouvoir

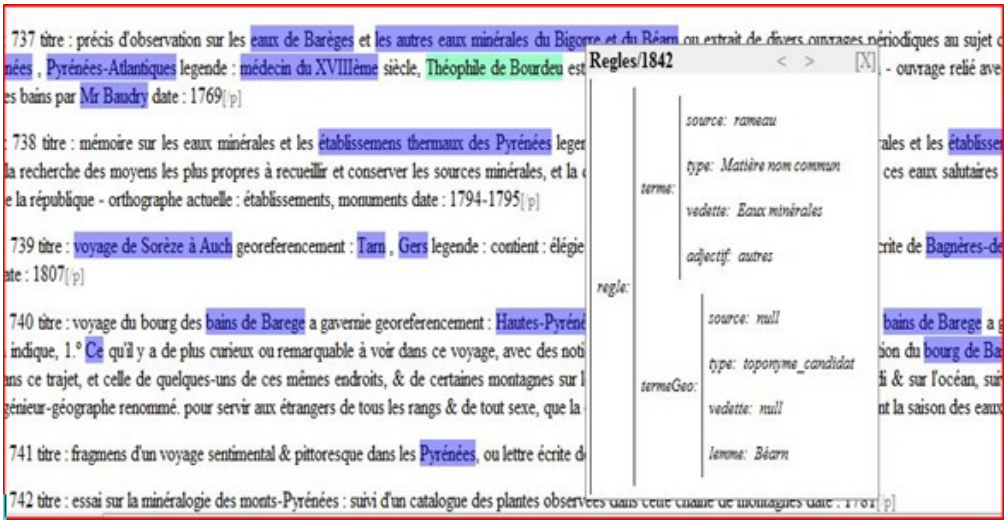


FIGURE 7.25 – Extrait du traitement Linguatream

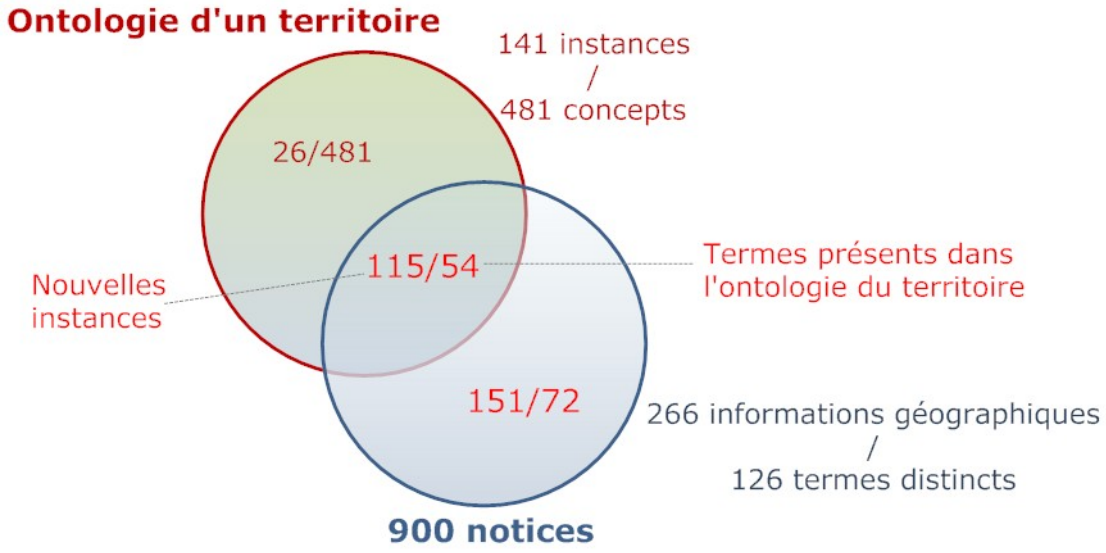


FIGURE 7.26 – Résultats de l'application de la chaîne de TAL sur les notices descriptives

enrichir cette première structure en utilisant comment point d'entrée le contenu de documents textes dits territorialisés. Rappelons que ce type de document se caractérise par une omniprésence des noms de lieux relatifs à un territoire particulier.

7.3.2 Modules (7 et 8) pour l'enrichissement de la représentation d'un territoire à partir du contenu des documents

Nous proposons d'appliquer notre chaine de traitement linguistique sur les documents eux-mêmes afin d'identifier des entités géographiques dans le contenu des documents constituant le corpus afin d'identifier un espace géographique le plus précis possible. Nous faisons l'hypothèse en traitant directement les documents de pouvoir obtenir une représentation plus précise du territoire décrit par le fonds documentaire indexé. Nous faisons le choix de restreindre l'identification des entités géographiques aux entités constituées d'une entité spatiale de type nom toponymique sur laquelle est ancrée une entité thématique correspondant à un label présent dans l'ontologie minimale.

L'intégration de l'ontologie obtenue à partir du travail d'indexation nous permet d'identifier une quantité importante d'entités géographiques (*cf.* figure 7.27).

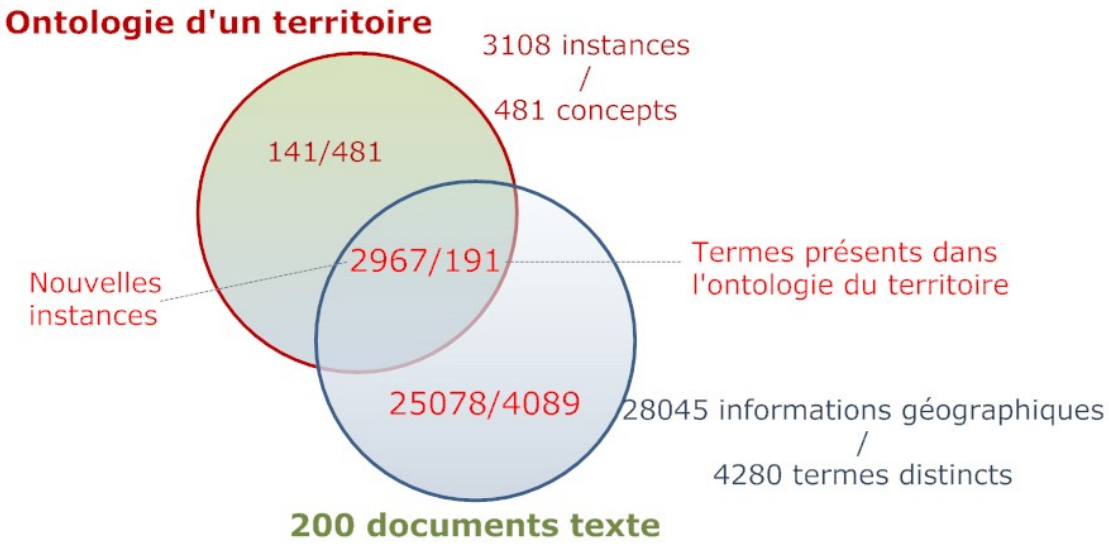


FIGURE 7.27 – Résultats de l'application de la chaîne de TAL sur les documents

Parmi ces entités, 2967 contiennent sont attachées à un label d'un concept présent dans l'ontologie, ce qui nous permet d'obtenir une ontologie contenant 3108 instances correspondantes à des informations géographiques se rapportant au territoire des Pyrénées. La quantité des entités géographiques identifiées comme instance de notre ontologie minimale augmente de façon significative lorsque nous appliquons notre chaîne de traitement linguistique sur le contenu des documents textes.

Nous remarquons d'après les statistiques présentées figure 7.27 que beaucoup d'informations géographiques, identifiées via notre chaîne de traitement, ne sont pas constituées d'entités thématiques correspondant à un concept de notre ontologie. Nous proposons de définir un processus automatisé afin de prendre en compte ces informations afin d'enrichir notre ontologie minimale par de nouveaux concepts et relations. Nous nous appuyons

pour cela sur la ressource RAMEAU, afin d'identifier des termes du thésaurus qui sont présents dans le contenu des documents.

7.4 Enrichissement de l'ontologie à partir du contenu des documents

Rappelons tout d'abord que l'enrichissement se fait par l'ajout de nouveaux labels/-concepts permettant d'élargir la couverture sémantique du domaine.

L'enrichissement de concepts est réalisé à partir de l'algorithme présenté figure 6.15 (*cf.* page 130) implémenté en JAVA. Cette étape nous permet de prévoir un enrichissement de l'ontologie de 940 concepts comme le montre la figure 7.28 (*cf.* page 161). Ces propositions d'enrichissement doivent en fin de traitement être validées par les experts bibliothécaires du domaine.

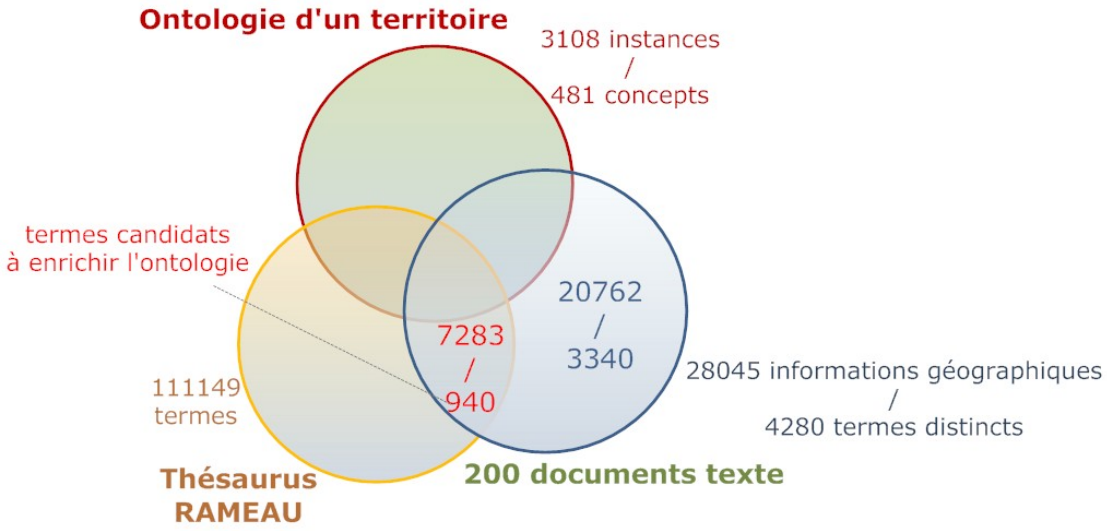


FIGURE 7.28 – Résultats de l'application de l'algorithme d'enrichissement sur les documents

Chapitre 8

Usages liés à notre approche

Sommaire

8.1 Géonto : la méthodologie Terridoc pour l'enrichissement d'une ontologie de domaine 164
8.1.1 Contribution au projet Geonto 164
8.1.2 Une chaine de traitement pour l'indexation de documents textes 165
8.1.3 Enrichissement de l'ontologie géographique 168
8.2 TERRIDOCViewer : Une application pour l'industrie permettant de naviguer dans des fonds documentaires . 170
8.2.1 Analyse des besoins 170
8.2.2 Architecture de l'application et Technologies 179
8.2.3 Schéma des bases de données 182
8.2.4 Fonctionnalités implémentées 185
8.2.5 Fonctionnement de l'application 191
8.2.6 Premiers bilans 192
8.3 Vers une application d'aide à l'indexation de documents pour les experts 192
8.3.1 Un premier contrôle 193
8.3.2 Propositions : vers une aide à la correction semi-automatisée 194
8.3.3 premiers bilans liés à l'analyse du travail d'indexation . . . 195

Nous avons proposé chapitre 6 page 101 une méthodologie automatisée Terridoc qui permet de construire une représentation sémantique d'un territoire sur la base d'un fonds documentaire indexé. Nous présentons dans ce chapitre des premiers usages qui nous permettent de mettre en application des modules de notre méthodologie dans des contextes divers.

Dans un premier temps, nous proposons dans le cadre du projet ANR Geonto, d'exploiter le module permettant d'extraire de textes grands publics des informations géographiques (cf. section 6.12 page 125) puis le module permettant d'enrichir une ontologie

d'un domaine cible (cf.6.3 129). Dans un second temps, nous présentons deux cas d'applications industrielles qui s'appuient sur des résultats intermédiaires obtenus lors de l'application de la méthodologie Terridoc sur un ensemble de documents. Nous présentons en premier lieu l'application Web TERRIDOCViewer qui permet de naviguer dans un fonds documentaire à travers le graphe de termes généré 8.2 page 170 sur la base du travail d'indexation. Nous travaillons actuellement à l'enrichissement de l'application pour intégrer les informations liées au territoire. Ensuite, nous décrivons un travail réalisé en collaboration avec les experts bibliothécaires, qui vise à identifier de façon automatique différents cas d'erreurs produites lors de l'indexation. Une pespective applicative qui correspond à un besoin des bibloithécaires est de proposer un module de correction semi automatisé des notices descriptives attachées aux documents fonctionnant en s'appyuant les résultats de l'analyse réalisée.

8.1 Géonto : la méthodologie Terridoc pour l'enrichissement d'une ontologie de domaine

Le projet Geonto intègre des équipes de recherche qui s'appliquent à proposer une méthodologie complète pour créer une ontologie géographique en s'appuyant notamment sur les spécifications des bases de données mises à disposition par l'IGN, l'objectif étant de définir une ontologie géographique qui fasse le lien entre l'ensemble des bases de données géographiques définies à l'IGN. Dans des expérimentations liées au projet Géonto, nous utilisons l'ontologie géographique produite pour améliorer les résultats obtenus à partir de notre chaîne d'indexation visant à identifier et qualifier des informations géographiques (cf. section 6.2.1 page 124) dans un ensemble de textes. L'objectif est ici d'identifier un volume plus important d'informations géographiques.

Si l'intégration de l'ontologie dans notre chaîne de traitement apporte des résultats encourageants avec une augmentation des entités géographiques identifiées, nous montrons que la couverture sémantique de l'ontologie reste limitée et que beaucoup d'entités géographiques ne peuvent encore être traitées. Dans ce cadre, nous proposons d'enrichir l'ontologie géographique produite dans Géonto par l'étape d'enrichissement proposée dans la méthodologie Terridoc présentée section 6.3 (cf. page 129 dans le but d'enrichir le vocabulaire de l'ontologie géographique et à terme d'améliorer les résultats de notre chaîne d'indexation. Avant de présenter en détail cette expérimentation, nous présentons brièvement le projet Geonto auquel nous apportons nos compétences au niveau de l'indexation automatisée d'informations géographiques.

8.1.1 Contribution au projet Geonto

Le projet GEONTO est un projet ANR mené dans le cadre de l'édition 2007 du programme « Masse de Données et Connaissances ». Le consortium correspondant regroupe quatre laboratoires spécialisés, dans le traitement de données cartographiques (COGIT - Paris), la construction d'ontologies (IRIT- Université de Toulouse) et leur alignement (LRI - Paris 11), ainsi que dans l'annotation et l'indexation automatisée d'informations

géographiques dans des fonds documentaires textuels (LIUPPA - Université de Pau). Le COGIT dispose de bases de données géographiques hétérogènes et a pour objectif l'interopérabilité de ces bases. Pour cela, le projet prévoit de fournir une ontologie par base de données, et d'aligner les ontologies obtenues avec une ontologie de référence construite semi automatiquement par le COGIT.

[KAG09] présente la démarche visant à construire automatiquement une première ontologie géographique à partir des spécifications au format Xml des bases de données. L'approche se veut globale et générique, et présente une double originalité. La structure présente dans les documents textuels est exploitée en plus du langage naturel, ce qui constitue une avancée dans la construction d'ontologies à partir de textes. Une application de la méthode a été réalisée sur la base de données BDTopo de l'IGN. Pour parler brièvement des résultats du traitement proposé dans cette méthode, l'analyse de la structure du document de spécifications de la base de données BDTopo a permis de construire une ontologie comportant 1258 concepts et 106 relations. Cette ontologie a été ensuite enrichie de 53 nouveaux concepts et de 11 nouvelles relations par l'analyse linguistique des définitions associées aux concepts déjà présents dans l'ontologie. 52 labels de cette ontologie ne font pas sens : ils correspondent à des valeurs numériques (21), à des caractères de ponctuation (4), à des termes neutres comme sans objet, inconnu, etc. (11), ou à des adjectifs (16). Par ailleurs 32 labels sont des énumérations et il existe 16 cas d'inclusion lexicale entre concepts frères. Une règle de correction est tout d'abord appliquée sur 12 labels énumératifs : 4 cas n'ont pas été validés. Ce traitement a produit 12 nouveaux concepts labélisés par des adjectifs, ce qui amène le nombre de labels adjectivaux à 28. Sur ces 28 labels, 21 répondent aux critères requis, à savoir que tous leurs frères ont des labels adjectivaux et que leur père a un label nominal. Ces 21 labels ont été automatiquement corrigés. Enfin 13 cas d'inclusion lexicale avec expansion droite ont été également automatiquement corrigés. Au total, 42 corrections valides ont été apportées à l'ontologie.

L'ontologie générée apporte un premier élément de réponse à la volonté dans le projet de proposer au grand public un accès structuré aux ressources d'un domaine cible. Cependant, afin d'assurer la couverture sémantique du domaine cible la plus importante possible, nous proposons d'enrichir l'ontologie à partir d'un échantillon de documents textes grand public, représentatif du domaine cible. Ce choix s'impose notamment par le fait que nous souhaitons que l'ontologie s'adapte au plus près du domaine d'application. La méthodologie proposée (figure 8.1) [KKS+09] se décompose de la façon suivante : (i) identification et validation des entités géographiques dans le texte ; (ii) identification des termes, constituants des entités géographiques, non présents dans l'ontologie présentée dans [KAG09] ; (iii) enrichissement de l'ontologie.

8.1.2 Une chaine de traitement pour l'indexation de documents textes

Nous appliquons la chaîne de traitement linguistique présentée section 6.3 (cf. page 129) sur un corpus de 14 livres de type récit de voyage qui mettent en avant les Pyrénées entre le XIX et le XX siècle.

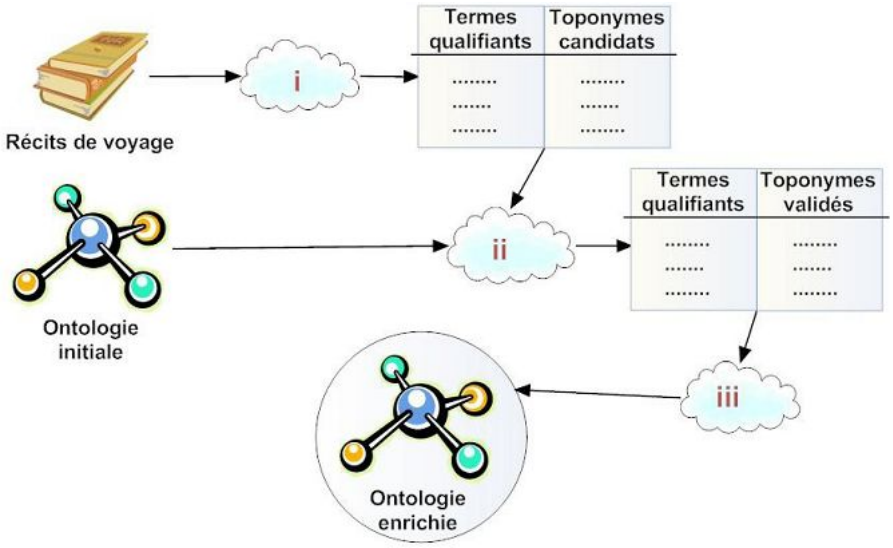


FIGURE 8.1 – Démarche d'enrichissement d'ontologie dans Géonto

8.1.2.1 Analyse qualitative de termes constituants des entités géographiques

La phase d'indexation nous permet d'obtenir deux ensembles de termes : ceux associés à des toponymes validés par des ressources géographiques et ceux associés à des toponymes candidats non validés. Par exemple :

- « vallée d'Ossau » : le terme « vallée » est associé au toponyme « Ossau », validé par la ressource IGN BD-Nyme ;
- « fontaine de Visos » : le terme « fontaine » est associé au toponyme-candidat « Visos » non validé par nos ressources.

Si l'on considère l'extrait de l'ontologie géographique générée (figure 8.2) et la liste des termes communs constituants des entités géographiques dans des récits de voyage, nous observons que 50% des termes sont communs à des concepts correspondants au niveau des feuilles de l'ontologie.

Ainsi, un grand nombre des termes distincts associés à des toponymes dans des récits de voyages pourraient venir enrichir les concepts de niveau feuille. Prenons l'exemple des termes *abîme*, *antre*, *caverne* qui reviennent régulièrement dans le corpus analysé ; il apparait clairement qu'ils pourraient enrichir l'ontologie par autant de nouveaux concepts (cf. figure 8.2 la branche relative au concept Grotte). Nous envisageons d'utiliser l'échantillon de termes distincts ainsi extrait et de l'associer au potentiel du thésaurus RAMEAU en vue d'enrichir l'ontologie géographique.

Si l'on prend l'exemple du toponyme « Crabioules » (cf. tableau 8.1.2.1) extrait de notre échantillon et associé à 8 termes dans ces mêmes textes, nous pouvons remarquer (figure 8.3) que le qualifiant « abîme » a pour terme vedette « Grottes » dans RAMEAU et que les termes rejetés pour « Grottes » sont « abîme », « antre », « aven », « caverne », « gouffre », etc.



FIGURE 8.2 – Extrait de l'ontologie géographique

Crabioules **Occurrences**	**Terme associé**	**Ontologie Géographique**	**Thesaurus RAMEAU**
1	abîme		✓
			✓
2	col	✓	✓
1	Corniche		✓
1	crête	✓	✓
1	Mont		✓
1	promenade		✓
1	route	✓	✓
1	sommet	✓	✓

« Crabioules » est qualifié dans l'échantillon de textes analysés par divers termes qui peuvent renvoyer à des représentations spatiales différentes du toponyme.

L'ontologie géographique offre donc un premier élément de réponse à la nécessité de typage des entités géographiques détectées dans des textes. Elle permet notamment d'identifier une représentation spatiale pour les entités géographiques « Col de Crabioules » et « Mont de Crabioules » (cf. *Col* et *Mont* dans le tableau 8.1.2.1). Le potentiel de notre échantillon de récits de voyages associé au thésaurus RAMEAU est relativement important. Cependant, un certain nombre de termes pose encore problème et nous souhaitons exploiter la structure du thésaurus RAMEAU pour enrichir cette ontologie et lever des ambigüités. Dans la partie suivante, nous proposons d'utiliser les termes associés à des toponymes dans des textes grand public afin d'enrichir l'ontologie géographique. La méthode décrite ci-après s'appuie sur le thésaurus RAMEAU et le recoupement de sous-arbres de l'ontologie avec des sous-arbres du thésaurus RAMEAU.

8.1.3 Enrichissement de l'ontologie géographique

Reprenons l'exemple de l'entité géographiques « les abîmes de Crabioules » (figure 8.1.2.1). Le terme *abîme* est alors identifié par notre chaîne de traitement comme un qualifiant dans une entité géographique. L'ontologie géographique ne permet pas de typer cette représentation spatiale car le concept « Abîme » n'existe pas. Dans ce cas, nous proposons d'enrichir l'ontologie géographique. Ainsi, pour le terme candidat « abîme », nous identifions trois équivalences entre les concepts de l'ontologie géographique sous le terme *Grotte* (figure 8.2) et la liste de termes provenant de RAMEAU liés à la vedette *Grottes* (figure 8.3). En effet, *grotte*, *aven* et *gouffre* sont présents dans les deux ensembles et cela nous permet de créer un nouveau concept *Abîme* en tant que fils du concept *Grotte*.


Grottes [+ subd. géogr.]

Vedette matière nom commun . S'emploie en tête de vedette

<Employé pour :
Abîmes
Antres
Avens
Cavernes *Ancienne vedette*
Cavernes préhistoriques
Cavités souterraines
Gouffres
Grottes ornées
Grottes préhistoriques
Préhistoire -- Grottes
Spélonques

<<Terme(s) générique(s) :
Habitat préhistorique
Relief (géographie)
Zones souterraines

FIGURE 8.3 – Exemple de notice descriptive dans RAMEAU décrivant le terme « Grottes »

La démarche proposée permet ainsi d'enrichir l'ontologie en ajoutant de nouveaux concepts afin de mieux couvrir le domaine géographie. L'ontologie ainsi mise en place donne alors la possibilité de lever des ambiguïtés dans la phase de qualification de l'entité géographique identifiée dans un texte. Il reste cependant des termes pour lesquels RAMEAU ne nous permet pas d'apporter de réponse. De plus, certains des termes qui enrichissent l'ontologie géographique peuvent engendrer des contresens. Par exemple, les termes glacier et gorges ont un double sens (géographique et autre), et sont chacun présents plusieurs fois dans RAMEAU, pour des contextes d'usage différents. Nous travaillons actuellement sur ces points en étudiant l'utilisation de ressources complémentaires (EuroWordNet, Larousse) qui permettraient de lever de telles ambiguïtés.

Nous proposons ci-après une évaluation de l'apport de l'ontologie géographique enrichie dans la chaîne de traitement linguistique présentée figure **chainePIV** (cf. page

125).

8.1.3.1 Analyse quantitative des termes associés aux EN spatiales

L'analyse des termes associés aux entités nommées et l'utilisation des données des BD de l'IGN nous permet de qualifier/typer automatiquement 5% des informations spatiales annotées (15% du nombre total d'occurrences). Ce taux passe à 10% de typage automatique des informations spatiales (33% du nombre total d'occurrences) si l'on utilise l'ontologie géographique générée dans le cadre du projet GEONTO. Nous obtenons 50% (75% du nombre total d'occurrences) de typage automatique des informations spatiales après l'enrichissement de cette ontologie à partir d'une étude combinée d'échantillons de textes grand public (récits de voyage) et du thésaurus RAMEAU. Nous avons détaillé dans [KKS+09] ces diverses statistiques. Ces traitements sont repris et enrichis dans [KKB+re]. Elles nous ont permis d'estimer l'enrichissement de l'ontologie via RAMEAU comme présenté figure 8.4. Cette figure résulte de l'analyse de 14 livres : 2388 occurrences distinctes de termes qualifiants y sont associées à des toponymes candidats.

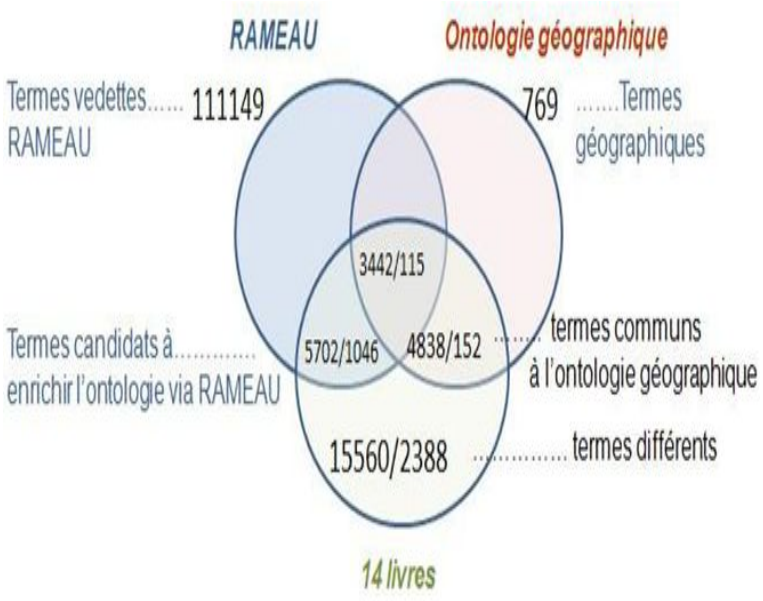


FIGURE 8.4 – Toponymes candidats et termes associés

Ainsi, 1046 termes RAMEAU distincts sont associés à ces qualifiants et sont candidats à l'enrichissement de l'ontologie. Nos travaux actuels affinent cette dernière statistique afin de détecter les termes RAMEAU porteurs d'un sens géographique. Nous proposons d'utiliser une méthode d'analyse des relations entre les syntagmes nominaux et les syntagmes verbaux dans les textes [BGKS10]. Nous faisons l'hypothèse en appliquant cette méthodologie de pouvoir réduire les ambigüités et le bruit dans l'ensemble des termes candidats à l'enrichissement.

8.2 TERRIDOCViewer : Une application pour l'industrie permettant de naviguer dans des fonds documentaires

Le travail présenté chapitre 7 page 135, permet dans un premier temps d'obtenir une représentation sémantique synthétisant le travail d'indexation des experts réalisé sur un fonds documentaire. Cette représentation, structurée sous forme de thésaurus, offre par la même occasion une vue générale du contenu de l'ensemble des documents constituant le fonds traité. Bien que limitée dans la description d'un territoire, cette représentation apporte des informations sur les thèmes (activités, constructions, éléments naturels, etc.) caractérisant le fonds traité.

Dans un second temps, cette représentation est transformée en ontologie et enrichie par des informations relatant d'un territoire. La version actuelle de TERRIDOCViewer s'appuie sur la représentation générale du fonds documentaire structurée sous forme de thésaurus pour proposer des outils de recherche et de navigation à travers les documents. Des travaux sont en cours pour intégrer dans une deuxième version l'ontologie de territoire obtenue afin de proposer des outils de recherche mettant en avant les spécificités spatiales et temporelles caractérisant le territoire.

8.2.1 Analyse des besoins

Les diverses réunions avec les experts bibliothécaires de la médiathèque et les responsables de la ressource RAMEAU ont permis de rédiger un cahier des charges complet définissant notamment les besoins des experts bibliothécaires. Nous identifions les différents acteurs, puis les cas d'utilisation de l'application.

Dans notre situation, nous nous limitons à trois types d'acteurs : l'utilisateur visiteur de la médiathèque, l'utilisateur catalogueur et spécialiste de l'indexation qui est un employé d'une médiathèque ou d'une bibliothèque, et enfin l'utilisateur responsable des mises à jour de Rameau à la BnF. L'application doit permettre aux experts, qu'ils soient catalogueurs ou de l'équipe RAMEAU, de naviguer à l'aide d'un graphe dans l'ensemble des termes du thésaurus RAMEAU. Elle doit également permettre aux catalogueurs de visualiser de façon synthétique le travail d'indexation réalisé sur un ensemble de documents. Enfin, TERRIDOCViewer doit proposer des outils permettant de consulter les documents de centres documentaires indexés à l'aide de la ressource RAMEAU.

L'objectif est de proposer des outils informatiques ergonomiques et simples d'utilisation qui facilitent l'accès au thésaurus RAMEAU ainsi qu'aux documents tout en valorisant le travail d'expertise des experts. Ici nous entendons par ergonomie la manière de mettre en adéquation les caractéristiques des machines avec les caractéristiques des hommes qui utilisent l'application. La réflexion menée se concentre sur le type de module à proposer à l'utilisateur pour l'assister dans ses recherches et sur la façon de représenter visuellement les résultats des requêtes pour en faciliter la compréhension.

En analysant les besoins des utilisateurs ciblés, quatre cas d'utilisation ont été identifiés :

– Navigation à travers un thésaurus tTerridoc généré sur la base d'un fonds docu-

mentaire annoté ;
- Visualisation des documents ;
- Consultation d'une fiche RAMEAU ;
- Paramétrage de l'affichage du module de recherche.

Afin de bien les illustrer, voici sur la figure 8.5, un diagramme général des cas d'utilisation.

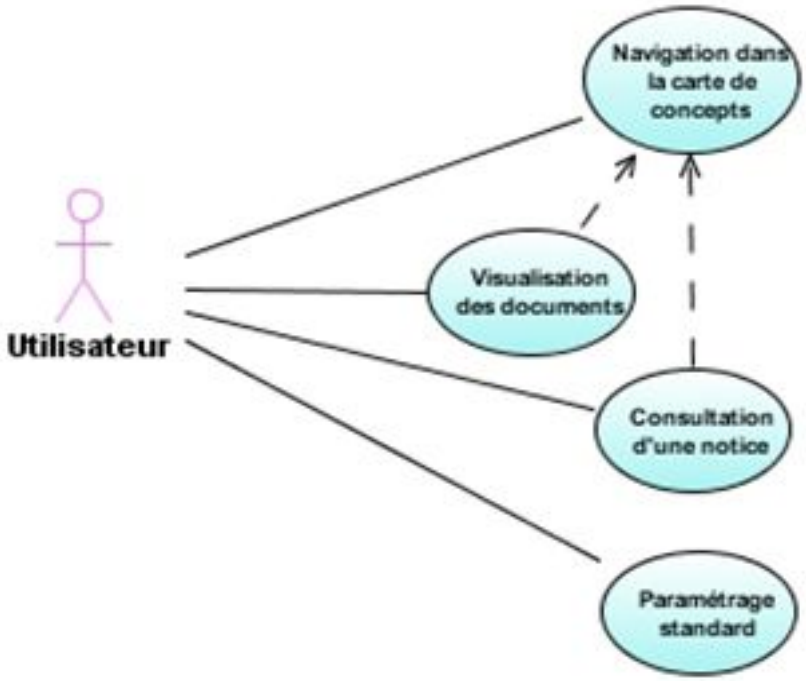


FIGURE 8.5 – Diagramme général des cas d'utilisations

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	Le spécialiste catalogueur peut naviguer dans le graphe de terme correspondant à tTerridoc, visualiser des documents, consulter la fiche RAMEAU associée à un termes ou bien encore paramétrer l'affichage du graphe.

8.2.1.1 Navigation à travers un thésaurus tTerridoc généré sur la base d'un fonds documentaire annoté

Etant donné le panel d'utilisateurs ciblés, il est important de prévoir différents modes de recherches d'informations pour permettre de répondre à la fois aux demandes pré-

cises lorsque l'utilisateur sait précisément ce qu'il recherche, et à des demandes moins précises lorsque l'utilisateur n'a pas une idée claire de ce qu'il recherche. Dans tous les cas, les thésaurus RAMEAU et tTERRIDOC (cf. section 7.9, page 146) sont des ressources intéressantes sur lesquelles nous pouvons nous appuyer pour proposer des outils de recherches avancés (exploitant la sémantique et la structure des thésaurus).

Une demande forte des catalogueurs et des responsables de RAMEAU est de pouvoir visualiser leur travail de façon synthétique. L'outil que nous proposons d'utiliser pour modéliser le thésaurus RAMEAU, ainsi que le travail d'indexation de fonds documentaires, est un graphe de termes. Les termes utilisés dans le graphe sont extraits du thésaurus RAMEAU, mais peuvent être issus d'un autre thésaurus. La navigation dans ce graphe de termes doit permettre la consultation des documents associés.

Il n'est pas évident de pouvoir se représenter instinctivement un ensemble de termes liés entre eux, et la quantité souvent importante de termes à disposition est un obstacle majeur à cette représentation et à la navigation dans cet ensemble. Pour tenter de répondre à ces problématiques, nous proposons une représentation visuelle structurée et hiérarchisée d'un ensemble de termes afin d'en faciliter la compréhension. Aussi lorsque l'on souhaite avoir une vision « locale » détaillée d'un sous ensemble des termes, un graphe semble être un outil adapté car il permet la mise en évidence des relations entre les termes et des termes eux-mêmes. Il permet notamment à un utilisateur qui ne connait pas forcément ce qu'il recherche de découvrir et de représenter visuellement cette hiérarchie de termes ainsi que les documents qui leurs sont reliés.

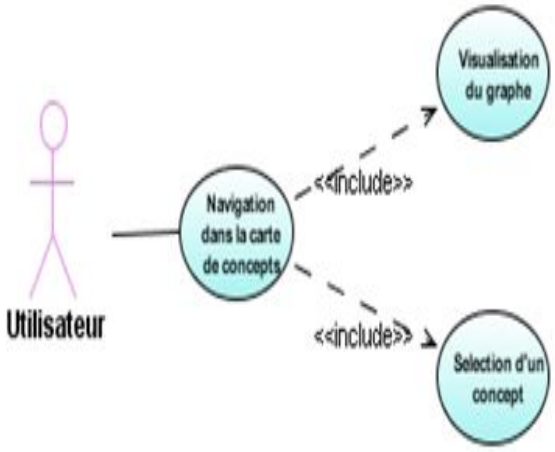


FIGURE 8.6 – Navigation dans le graphe de termes

Visualisation du graphe de termes

Plusieurs éléments principaux ont été établis, il est important d'avoir un graphe structuré qui respecte tout le temps une même logique. Le graphe doit ainsi com-

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	L'acteur peut naviguer dans le graphe de termes, cette navigation s'effectue en deux étapes principales : la sélection d'un terme central tout d'abord suivie de la visualisation du graphe correspondant

porter un terme central autour duquel les autres termes gravitent (figure 8.7) et doit être accessible constamment sur la fenêtre principale.

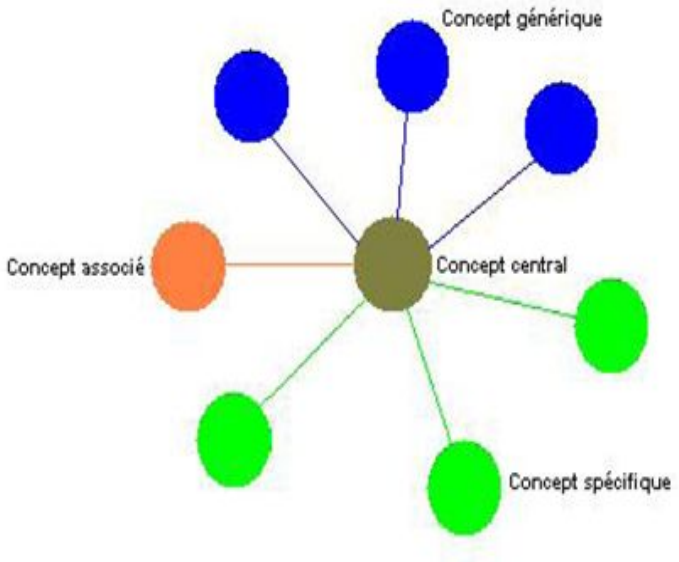


FIGURE 8.7 – Structure du graphe

Il est primordial de prendre en compte le nombre de termes ainsi que le nombre de niveaux à afficher pour ajuster la taille des nœuds afin d'avoir une meilleure visibilité. En effet, plus le nombre de termes et le niveau choisi sont importants plus il y a de nœuds et moins le graphe est lisible. Aussi, afin de pallier ce problème, il est important de prendre en compte ces facteurs afin de dimensionner les nœuds en conséquence.

Afin de clarifier l'affichage, il a été fait le choix de différencier les types de relations entre les termes, pour cela l'utilisation de couleurs différentes semble être une solution simple et efficace (par exemple, la couleur bleu pour l'icône d'un terme générique) (figure 8.7). Par défaut, les types des relations ne sont pas affichés, pour ne pas surcharger le graphe. De plus, il est utile de positionner les termes par type de relation entretenu avec le terme central (par exemple les termes génériques sont affichés au dessus du terme central), cela permet d'ordonner le graphe, le graphe est

donc plus lisible. Lorsqu'il y a plusieurs occurrences reliées à un terme, une icône représentant une pile de documents est affichée, pour la représentation d'un seul document, l'utilisation d'une icône représentant le type de document (un haut parleur si c'est un document son), permet d'indiquer la nature du document présent. Ainsi l'information est affichée sans qu'aucune autre action ne soit effectuée.

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	Il est possible de visualiser un terme central, ainsi que les termes parents, fils et associés voire plus si le niveau sélectionné est supérieur à 1

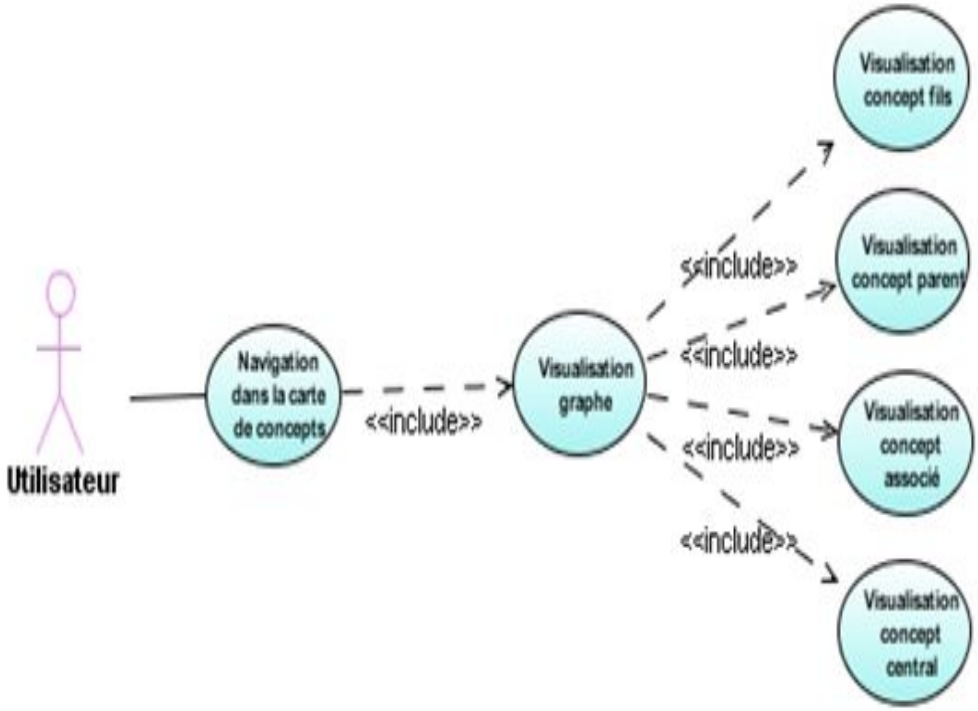


FIGURE 8.8 – Visualisation du graphe

Sélection d'un terme

Trois types de navigation sont possibles : une recherche par mots clés proposant l'auto complétion pour permettre à l'utilisateur d'être assisté dans sa recherche, un index qui permet d'avoir une vue globale de l'ensemble des termes et la navigation via le graphe qui permet d'avoir une vue locale de l'information. L'utilisateur doit ainsi avoir à disposition un index (liste) hiérarchique de l'ensemble des termes lui

permettant de naviguer dans le fonds documentaire. La sélection d'un terme (via un des trois types de navigation) permet la génération du graphe correspondant avec pour terme central le terme sélectionné. La mise à disposition d'un historique est intéressante afin de revenir à des recherches précédentes et de pouvoir retracer le chemin de navigation.

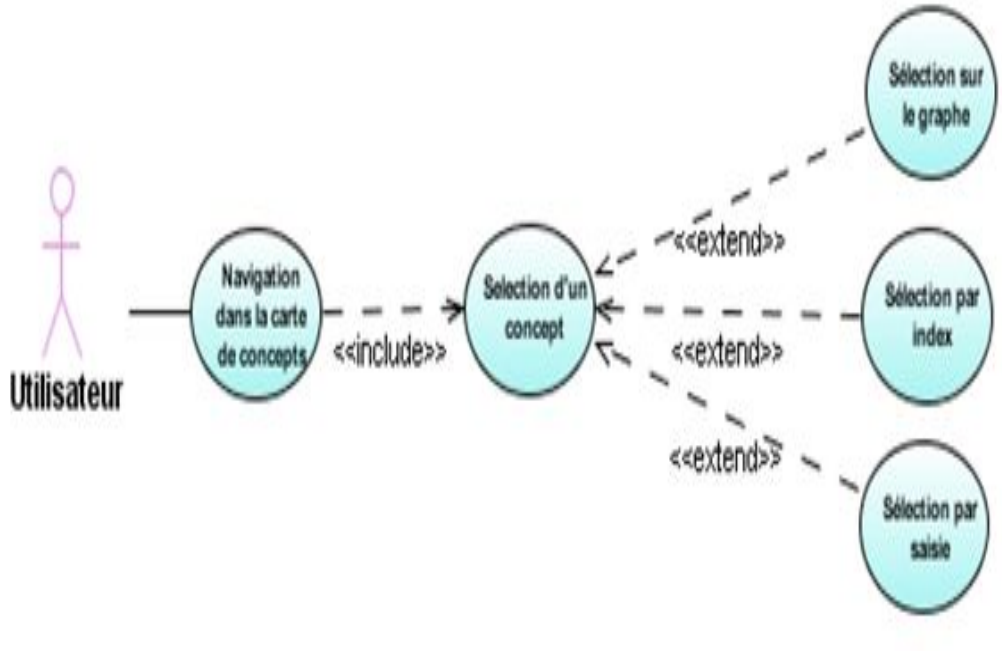


FIGURE 8.9 – Sélection d'un terme

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	L'utilisateur peut naviguer dans le graphe de trois manières différentes (sélection dans le graphe, par index ou bien par saisie). Après sélection d'un nouveau terme central, on peut visualiser un nouveau graphe avec comme terme central le terme sélectionné

Diagramme de séquence pour la navigation dans le graphe

La sélection d'un terme s'effectue à l'aide d'un index (l'utilisateur sélectionne un terme dans la liste hiérarchique des termes), de la recherche par mots clés (l'utilisateur doit effectuer une saisie dans une zone texte) ou bien en cliquant sur un terme du graphe. Cela entraîne l'envoi d'une requête au serveur avec l'identifiant du terme sélectionné. Le serveur effectue des appels vers la base de données dans le but de récupérer les données nécessaires à la génération du graphe. Une fois les données acquises, le serveur envoie les données au client. Le client génère alors le

graphe correspondant. L'utilisateur voit le graphe demandé s'afficher sur la page de son navigateur. Nous présentons ce diagramme figure 8.10.

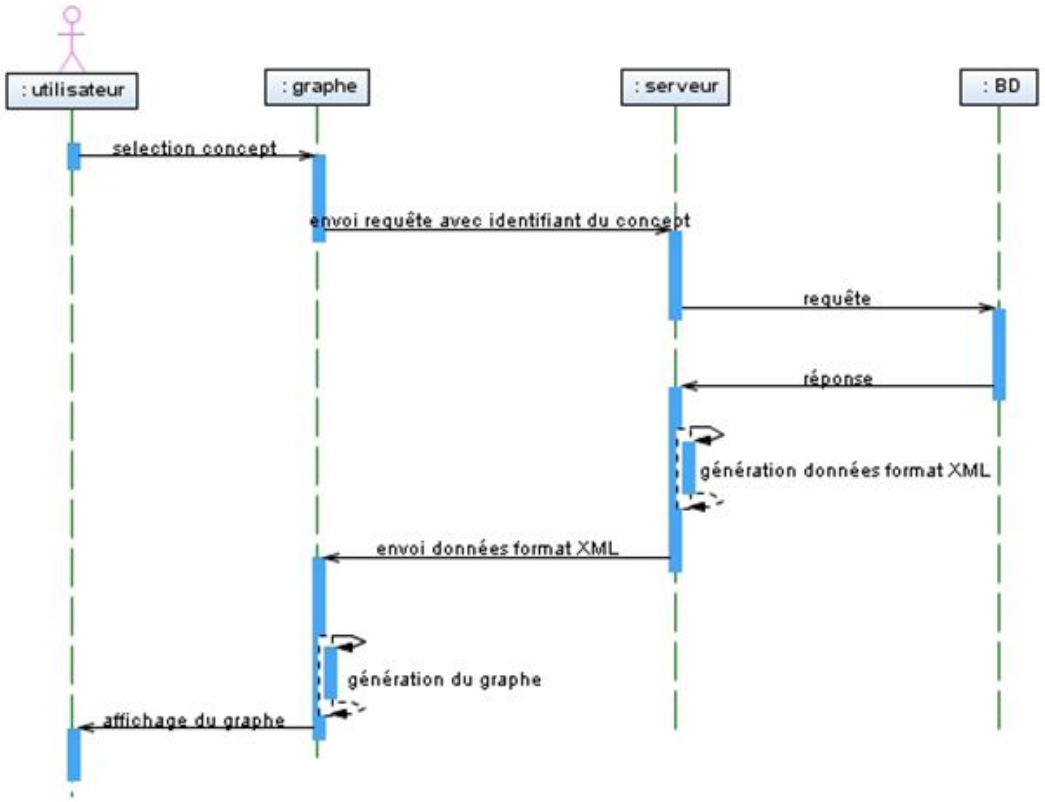


FIGURE 8.10 – Diagramme de séquence pour la navigation dans le graphe

8.2.1.2 Visualisation des documents

Un élément important de l'application est l'accès aux documents. Mais pour préserver une bonne interactivité, la visualisation dans des documents se doit d'être rapide en temps de chargement quels que soient la taille et le type du document. En cliquant sur une occurrence dans le graphe, l'utilisateur peut dans la même fenêtre, visualiser une liste de documents en miniatures, et peut ensuite, s'il le souhaite, visualiser en détails un des documents dans un nouvel onglet de l'application : description du document et visualisation du document en taille réelle.

8.2.1.3 Consultation d'une fiche RAMEAU

Comme présenté chapitre 7.2.1.1 page 140, les fiches RAMEAU sont utilisées dans les médiathèques et bibliothèques afin de faciliter la réalisation des notices décrivant les documents (phase d'indexation). Nous nous sommes rendu compte en analysant le travail

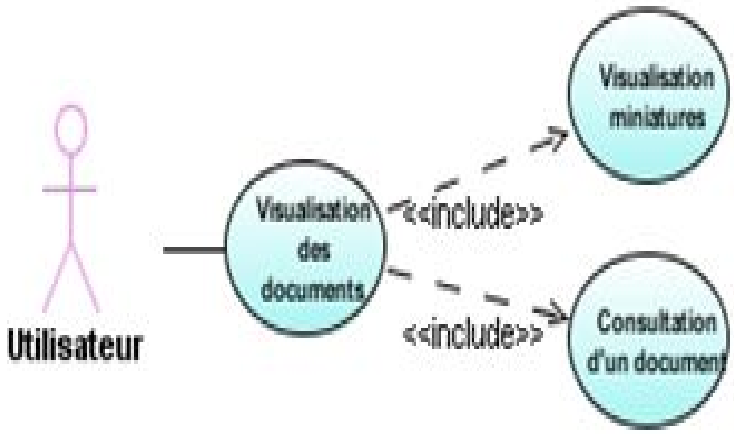


FIGURE 8.11 – Visualisation des documents

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	Le spécialiste catalogueur peut visualiser les documents, pour cela il doit sélectionner l'icône documents associé à un terme du graphe. Dans un premier temps, il visualise une liste de documents en miniatures puis après sélection, il a la possibilité de consulter en détail le document (description et affichage en grandeur nature)

réalisé par les experts catalogueurs qu'ils s'appuient régulièrement sur ces fiches pour indexer de nouveaux documents. Actuellement, la seule façon d'y accéder est de consulter le catalogue RAMEAU en ligne [89], obligeant à ouvrir une nouvelle application Web. Nous proposons donc d'intégrer ces informations dans l'application TERRIDOC pour éviter un accès externe, engendrant des "allers et retours" réguliers entre l'application hébergeant le thésaurus et l'application utilisée pour indexer les documents.

L'utilisateur doit avoir la possibilité d'afficher une fiche descriptive pour chaque terme, les informations ont la structure des fiches RAMEAU, comme présenté sur la figure 7.5 142. L'accès à des fiches évite à l'utilisateur de devoir aller chercher les fiches sur le catalogue RAMEAU.

8.2.1.4 Paramétrage de l'affichage

Quel que soit la recherche de l'utilisateur, des éléments de paramétrage lui permettant, de façon simplifiée, de sélectionner les informations à afficher ont été prévus.

Paramétrage standard : L'utilisateur peut sélectionner les termes à afficher en fonc-

89. http ://catalogue.bnf.fr/jsp/recherche_autorites_rameau.jsp ?nouvelleRecherche=O&host=catalogue

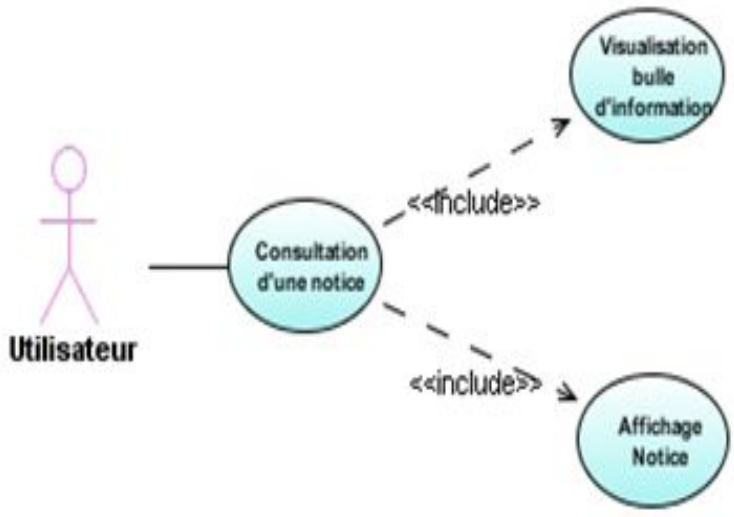


FIGURE 8.12 – Consultation d'une fiche RAMEAU

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	A tout moment, il peut consulter la notice descriptive associée à un des termes affichés sur le graphe. Pour cela, il lui suffit d'effectuer un clic gauche sur le terme voulu. Là, une bulle d'information apparaît, pour en savoir plus l'utilisateur doit cliquer sur un lien de la bulle qui ouvre la notice descriptive

tion de leur niveau de relation avec le terme central (N+2 par exemple un parent au terme parent du terme central, N-2 un fils au terme fils du terme central), ce qui permet une vue plus élargie par rapport à un niveau 1 et par la même occasion une navigation plus large aussi. L'application doit pouvoir permettre de choisir les types d'associations à afficher : « génériques », « spécifiques », « associés » ou encore « employé pour ». Par exemple, pour le terme Faillite on a va avoir pour terme spécifique liquidation et pour terme générique Finances. Si seuls les termes « spécifiques » sont sélectionnés pour être affichés, on aura un graphe avec le terme central Faillite et le terme Liquidation seulement. Ce paramétrage permet d'alléger la visualisation du graphe et de cibler un peu plus ses recherches.

Paramétrage avancé : Il est possible d'afficher ou non les noms des relations à côté des liens (spécifique, générique, etc.), la couleur et le style pour chaque type de relation peuvent être choisis. De plus, l'application doit permettre de sélectionner le type d'icône pour les termes navigables, pour les termes reliés à des documents, ainsi que pour les documents afin d'avoir un affichage plus clair et personnalisé. Enfin, le thème graphique de l'interface peut être changé.

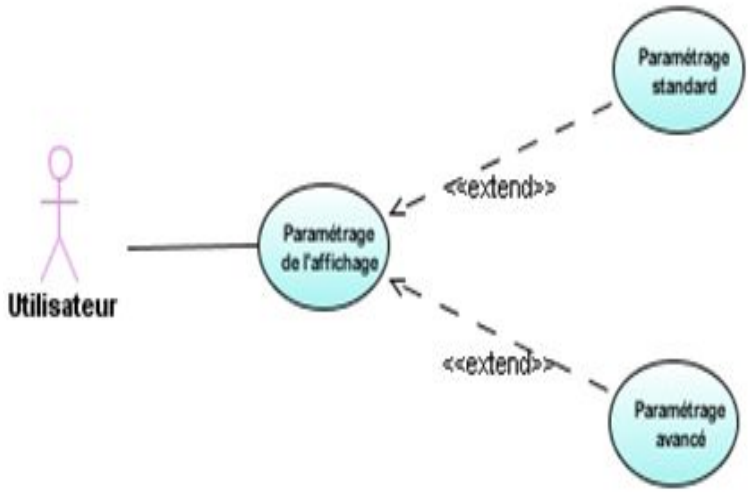


FIGURE 8.13 – Paramétrage de l'affichage

Acteurs	Visiteurs Spécialiste catalogueur et spécialiste Rameau
Description	L'utilisateur a la possibilité de paramétrer certains aspects de l'affichage du graphe.

8.2.2 Architecture de l'application et Technologies

L'application a une architecture 3-Tiers avec un tiers client (navigateur), un tiers applicatif (serveur Apache Tomcat) et un tiers données (la base de données et le fonds documentaire) (figure 8.14). L'architecture de l'application respecte le framework Struts 2 [90]. Aussi la partie applicative met en œuvre le modèle MVC2 [91] avec une partie Contrôleur utilisant une servlet afin de gérer les requêtes clientes, une partie Modèle pour la récupération et la gestion des données à envoyer, et une partie Vue utilisant des JSP [92] et Ajax [93].

8.2.2.1 Tiers client

Le tiers client a pour rôle l'affichage de l'interface utilisateur. Le client utilise AJAX pour l'affichage du graphe de termes avec le framework JSViz [94] et l'affichage de l'in-

90. Framework MVC issu de WebWork et Struts, http ://struts.apache.org/2.x/index.html. documentation : Venkatray Kamath, Struts 2.0 in action, JavaWorld.com, http ://www.javaworld.com/javaworld/jw-10-2007/jw-10-struts2inaction.html ?page=1
91. Modèle Vue Contrôleur 2, http ://fr.wikipedia.org/wiki/Modèle-Vue-Contrôleur
92. http ://www.abrillant.com/doc/fiches/FD_01.html
93. https ://developer.mozilla.org/fr/AJAX
94. http ://code.google.com/p/jsviz/

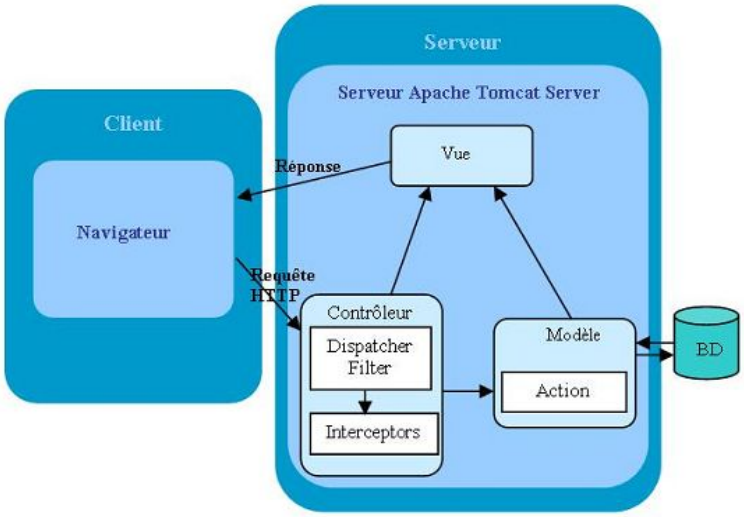


FIGURE 8.14 – Architecture de l'affichage

terface globale avec le framework Dojo Toolkit [95]. La communication avec le serveur s'effectue grâce à l'objet XMLHttpRequest [96].

8.2.2.2 Tiers applicatif

Le tiers applicatif a pour rôle de gérer les requêtes clientes et de renvoyer les données nécessaires au client en réponse à leurs requêtes. Il se décompose en trois modules selon le frameworks Struts 2 qui est basé sur le pattern MVC2 : le contrôleur, la vue et le modèle. La figure 8.15 présente le fonctionnement de ces modules [97].

8.2.2.3 Tiers données

Le tiers données est représenté par la base de données comportant l'ensemble des termes du thésaurus RAMEAU ainsi que par le fonds documentaire dans lequel l'application doit permettre de naviguer.

8.2.2.4 Aspects techniques

Comme pour la partie construction de l'ontologie de territoire, le langage JAVA est utilisé côté serveur car cette technologie est maitrisée par l'équipe de développement

95. http ://dojotoolkit.org/
96. http ://www.w3.org/TR/XMLHttpRequest/
97. En résumé, le fonctionnement est le suivant : une requête cliente est envoyée au conteneur de servlets. Le FilterDispatcher (un filtre servlet standard) va intercepter cette requête. Il délègue alors le contrôle à l'ActionProxy. Les Interceptors sont appelés et invoquent les classes actions. Celles-ci accèdent à la couche business et aux données nécessaires. La requête est renvoyée à la vue (JSP) avant de repasser par les Interceptors. La réponse est alors renvoyée au Filter Dispatcher avant d'être retournée au client.

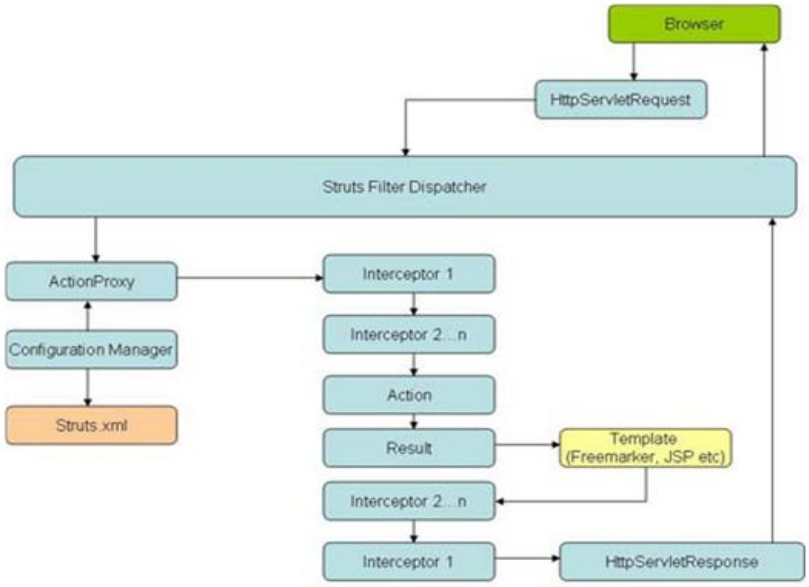


FIGURE 8.15 – Architecture de l'affichage selon le framework MVC Struts 2

de DIS et facilite l'intégration dans les portails Web et applications existantes dans l'entreprise. De plus, elle apporte l'ensemble des fonctionnalités requises dans TERRIDOCViewer. Dans le but de mettre en place une application structurée impliquant une séparation du modèle et de l'interface utilisateur, il a été choisi de mettre en place le pattern MVC. Une étude comparative des différents frameworks MVC JAVA [98] existants réalisée au sein de l'entreprise DIS, nous conforte dans le choix d'utiliser Struts 2. Nous nous accordons avec une étude réalisée par Matt Raible [99] pour dire que Struts 2 facilite une maintenance à long terme de l'application et permet dans le même temps de modifier ou d'enrichir le volet interface pour différents clients sans devoir intervenir sur le cœur de l'application. Aussi, ce framework est open source et possède une communauté de pratique bien plus importante que les autres technologies disponibles. Il intègre une liste importante de composants et de plugins qui permettent de prévoir des améliorations à long terme.

Deux Systèmes de Gestion de Bases de données (SGBD) différents sont utilisés MySQL 6.0 et PostgreSQL 8.2, pour permettre de s'interfacer rapidement avec les différentes applications de l'entreprise. En effet, diverses solutions proposées par DIS utilisent les deux SGBD et il est indispensable de respecter ce prérequis. Il est donc important que le requêtage soit valide aussi bien pour l'un que pour l'autre. L'application doit intégrer les outils DIS pour la visualisation ou l'écoute des documents : Viewer VLS-Image [100]

98. Nous pouvons notamment citer Swing (http ://www.java2s.com/Tutorial/Java/0240_Swing/SwingMVC.htm) , Spring (http ://www.springsource.org/) , JSF (http ://mbaron.developpez.com/javaee/jsf/), etc.
99. https ://equinox.dev.java.net/framework-comparison/WebFrameworks.pdf
100. http ://www.docimsol.eu/index.php ?id=23

pour les images, Viewer FLV [101] pour les vidéos , Viewer PDF pour les textes.

L'application Web doit être accessible depuis tout poste ayant un navigateur Web tel que Mozilla Firefox 3.0 ou bien Internet Explorer 8. Elle doit être utilisable depuis un poste distant. L'installation de l'application doit s'effectuer à partir d'un installateur semi-automatique ou automatique.

Les outils utilisés doivent être des logiciels libres. De plus, les technologies du Web évoluent très vite et le choix d'outils payants implique des frais de mise à jour importants. Ajax est utilisé comme technologie côté client, avec le framework JSViz pour la création du graphe et Dojo Toolkit [102] pour l'interface graphique globale. Ce choix résulte d'une étude comparative des différents frameworks Ajax permettant de développer des interfaces Web riches « EtudeComparativeFrameworksAjax.doc », l'IHM de l'application doit être ergonomique et intuitive.

8.2.3 Schéma des bases de données

8.2.3.1 Base de données thesaurus

La base de données « thesaurus » permet de stocker des thésaurus qui sont utilisés pour indexer les documents des fonds documentaire qui seront pris en compte dans l'application TERRIDOCViewer. La base de données a été définie sur le modèle de la base de données utilisée pour le thésaurus RAMEAU. Elle est explicitée plus en détails dans un document [103] décrivant la base de données qui lui est entièrement consacrée, un schéma de la base est disponible figure 8.17.

La base de données se décompose en dix tables, voici un bref descriptif des tables utiles à la création d'un graphe :

- Term qui définit un terme du thésaurus RAMEAU ;
- Associated_term qui permet de modéliser les associations entres les termes (génériques, spécifiques, associés, et autres) ;
- Rejected_term qui correspond aux termes exclus.

Pour générer un graphe, la table Term est utilisée pour récupérer les informations sur le terme dont l'identifiant est passé en paramètre. Il faut ensuite accéder à la table Associated_term pour obtenir les termes liés à ce terme, ainsi qu'à la table Rejected_term pour les termes exclus correspondants.

8.2.3.2 Base de données bdterridoc

Afin de pouvoir naviguer dans un fonds documentaire à partir d'un graphe de termes issus d'un thésaurus, il était important de faire la correspondance entre un fonds documentaire et un thésaurus associé. C'est à dire de pouvoir faire la liaison entre les entités

101. Technologie intégrée entre autre à une application de recherche d'informations dans flux vidéos d'actualités : http ://www.docimsol.eu/index.php ?id=50
102. www.dojotoolkit.org
103. Rapport technique interne de la base de données Thesaurus

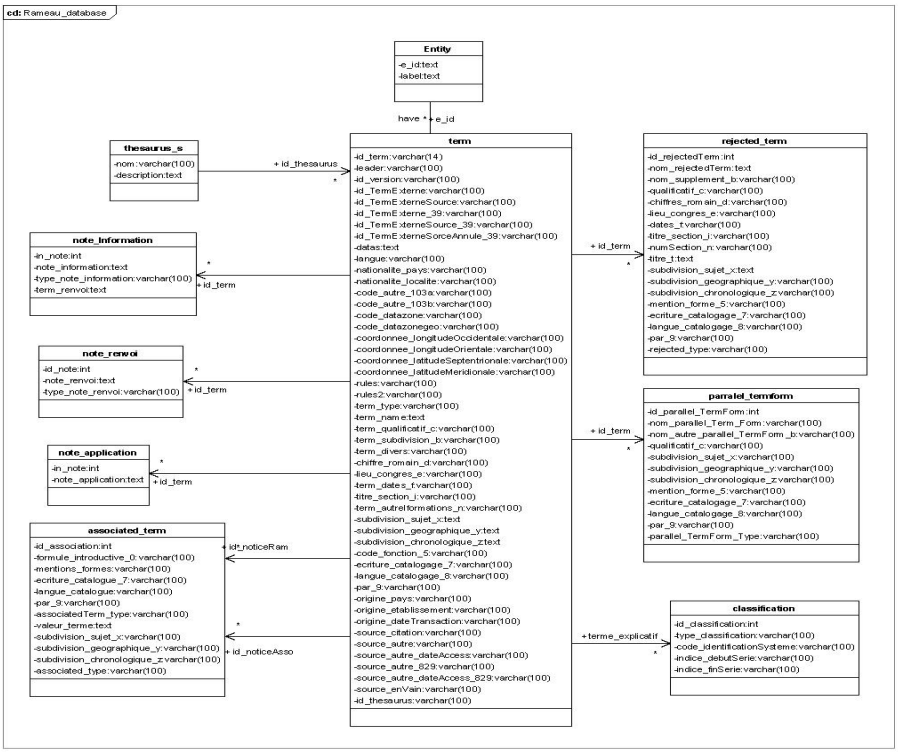


FIGURE 8.16 – Structure de la base de données *thesaurus*

nommées contenues dans les notices descriptives des documents, et les termes correspondants dans le thésaurus. Ainsi, une base de données « bdterridoc » (figure 7.2.1.1) a été élaborée, permettant de stocker les informations concernant un fonds documentaire.

Entre autres, sont stockées des informations concernant les documents, les notices descriptives et les entités nommées contenues dans celles-ci. Les informations sur les documents permettent ainsi la récupération des documents là où ils sont stockés. *bdterridoc* se décompose en 10 tables :

- fondsdocumentaire qui fait la liaison entre un fonds documentaire et le thésaurus associé ;
- notice qui modélise les notices descriptives des documents ;
- Les tables entity, spatialentity et temporalentity représentant les entités nommées contenues dans les notices descriptives des documents. Il existe les entités nommées spatiales (spatialEntity), les entités nommées temporelles (temporalentity) et les autres (entity), elles sont associées à la table notice grâce aux tables d'association index_e, index_se et index_te. Les entités nommées peuvent faire référence ou non à un terme du thésaurus associé (attribut « id_term ») ;
- Les rejectedEntity correspondent aux entités nommées exclues ;
- Et enfin la table document contient les informations sur les documents du fonds documentaire.

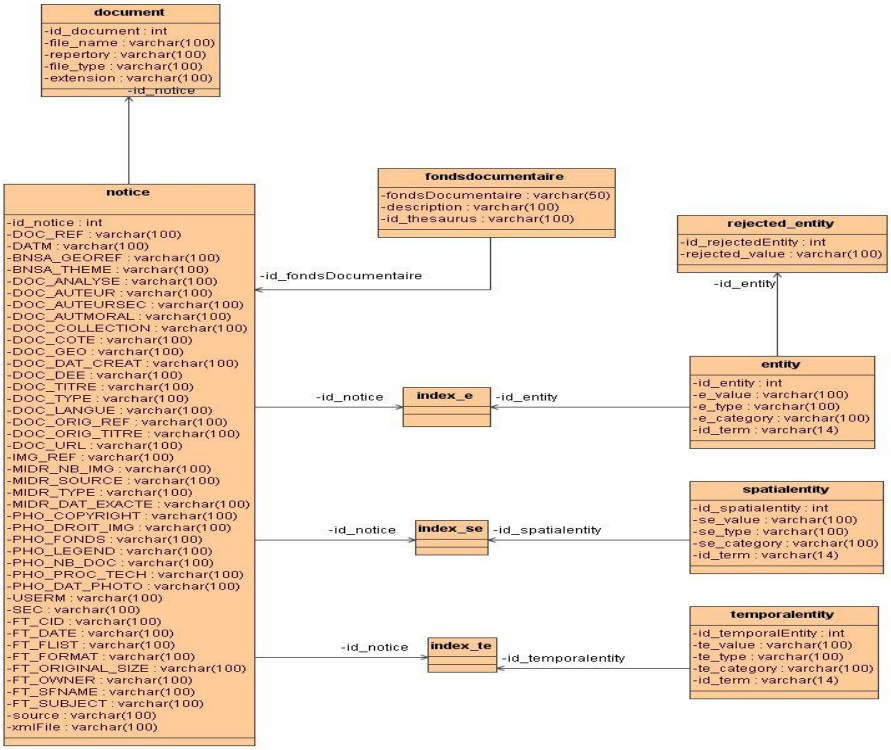


FIGURE 8.17 – Schéma de la base de données *bdterridoc*

8.2.3.3 Gestion des paramètres de visualisation

L'application devant permettre un paramétrage de l'interface, il était important de pouvoir conserver les paramètres d'un utilisateur après fermeture de l'application. Ainsi, à chaque lancement de l'application, celle-ci conserve les choix d'affichage définis précédemment, ce qui évite à l'utilisateur de paramétrer l'interface à chaque ouverture de l'application. Une base « paramétrage » a donc été mise en place et stocke les préférences de l'utilisateur (figure 8.18). Actuellement, TERRIDOCViewer ne permet pas de prendre en compte des configurations de navigations dédiées : prise en compte des choix par utilisateurs, stockage de ces choix pour une navigation ultérieure, etc. Ce mode de configuration dédié nécessite un développement logiciel important qui ne rentre pas dans le cadre des travaux de thèse. Cependant, ce mode est prévu dans une version à venir de TERRIDOCViewer.

La base de données est constituée de :

- Une table utilisateur permet d'identifier l'utilisateur connecté ;
- Une table parametrage_simple permet de conserver les informations concernant l'affichage des relations ou du niveau d'affichage du graphe ;
- Une table configuration stocke les données concernant les sources de données à utiliser (thesaurus, fonds documentaire) ;
- Les autres tables correspondent au paramétrage avancé.

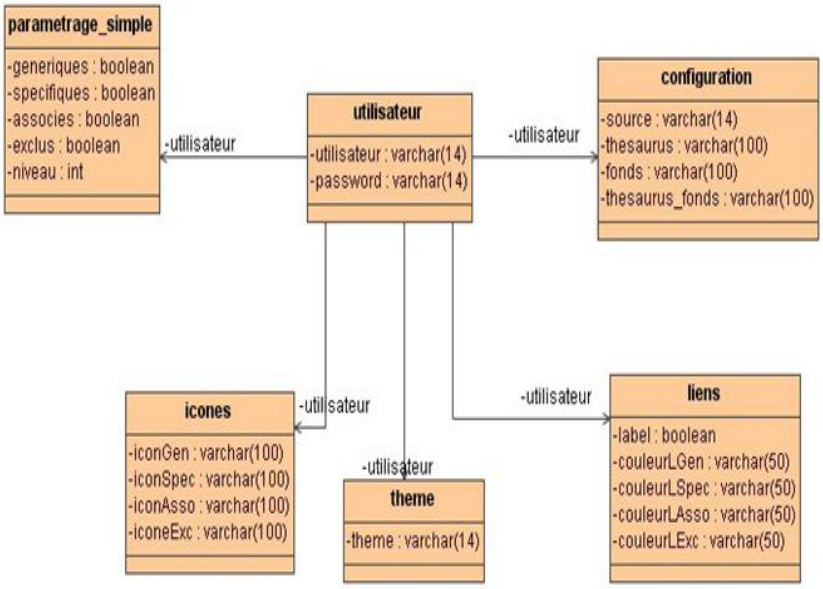


FIGURE 8.18 – Schéma de la base de données *paramétrage*

Il est spécifié dans la définition du tiers client de l'application (8.2.2.1) que la technologie utilisée pour la création de l'interface est le framework AJAX Dojo Toolkit, nous allons voir à présent un récapitulatif de l'étude réalisée pour parvenir à ce choix.

Maintenant que les besoins sont explicités et que les choix techniques sont présentés, nous décrivons les différentes fonctionnalités implémentées. Nous présentons ensuite de façon succincte le fonctionnement de l'application.

8.2.4 Fonctionnalités implémentées

Présentons maintenant les diverses fonctionnalités mises en place dans l'application TERRIDOCViewer.

8.2.4.1 Visualisation du graphe

Un travail a été effectué au niveau de l'affichage du graphe pour qu'il soit le plus lisible possible, des couleurs ont été utilisées afin de différencier le type de relations, de plus la taille des nœuds tient compte du nombre de niveau ainsi que du nombre de termes à afficher. La figure 8.19 est une capture d'écran de l'application fonctionnant sur un fonds documentaire. Le graphe généré a pour terme central *Faillite* (recherche effectuée via la rubrique Recherche sur la gauche). Les termes génériques sont représentés en bleu, les spécifiques en vert, et les associés en rose (cf. la légende sur la Figure 8.20). Nous pouvons voir par exemple que *Faillite* a pour terme générique *Finances*. Tous les termes présents sur l'interface sont reliés directement à *Faillite*, en effet, l'application est paramétrée avec un niveau d'affichage égal à 1 (termes distants de 1 relation du terme

sélectionné). On peut voir également via une icône (représentant des documents) que des ressources documentaires sont rattachées à ce terme. La visionneuse en dessous du graphe permet de les visualiser rapidement (la visionneuse apparaît après une action de l'utilisateur).



FIGURE 8.19 – Visualisation du graphe



FIGURE 8.20 – Légende du graphe

Bien qu'intéressante et adaptée à la visualisation d'un graphe selon la composante

thématique comme présenté figure 8.19, TERRIDOCViewer ne permet pas en l'état actuel de traiter au niveau des modules de recherche et de l'interface les composantes spatiales et temporelles. Nous travaillons actuellement au sein de DIS à la mise en place d'un prototype faisant intéragir des modules de recherches caractérisants les trois composantes formant une entité géographiques (thème, spatial et temporel).

8.2.4.2 Paramétrage

Le paramétrage répond aux éléments exigés dans le cahier des charges, un paramétrage standard est possible via la rubrique paramètres, et permet le choix des relations à afficher ainsi que le niveau. La figure 8.21 représente un graphe de termes ayant le terme *Faillite* comme terme central et un niveau d'affichage N+3, N-3, comparé à la figure 8.19, les termes sont plus rapprochés entre eux et la taille des icones est réduite. L'affichage d'un tel graphe permet d'avoir une vue plus large de la hiérarchie de termes, ainsi que des documents rattachés aux différents termes à différents niveaux. Il apparait par exemple que le terme *Faillite* a pour terme générique *Finances* qui lui-même a pour terme générique *Affaires*, il est donc plus facile de faire la correspondance entre *Faillite* et *Affaires* grâce à cet affichage.

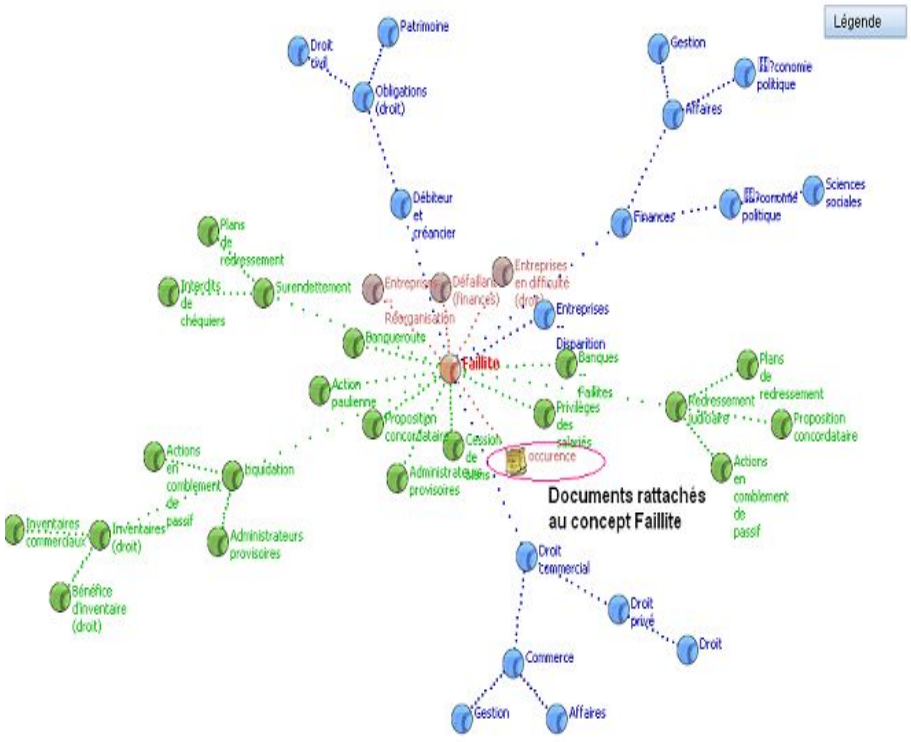


FIGURE 8.21 – Graphe au niveau N+3, N-3

Un paramétrage avancé permet le choix des icônes, de la couleur des liens et du thème

de l'interface (cf. Figure 8.22).

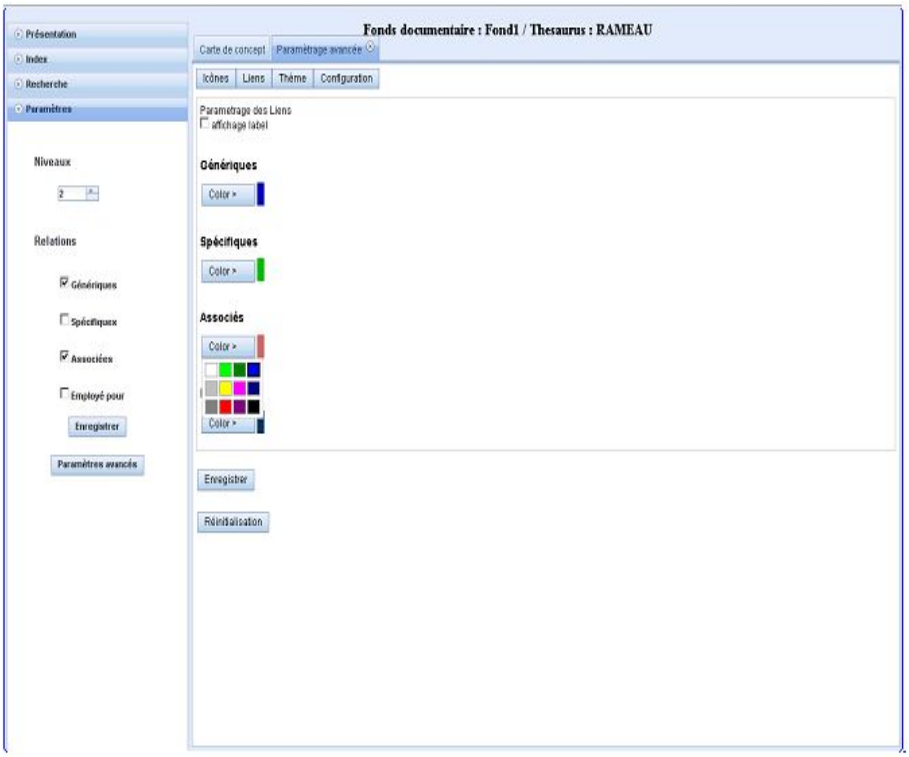


FIGURE 8.22 – Eléments de paramétrage

Ce paramétrage avancé permet aussi de configurer la source utilisée pour l'application. En effet, l'application peut fonctionner sur un fonds documentaire avec un thesaurus associé ou sur un thésaurus seulement, dans ce cas là, il n'y aura pas de documents associés aux termes. La navigation ne s'effectuera seulement que dans un ensemble de termes. Un changement de la source entraîne une réinitialisation de l'application. Le paramétrage est stocké pour tous les utilisateurs et permet de conserver d'une utilisation à une autre les préférences.

8.2.4.3 Recherche par mots clés avec auto complétion

La recherche par mots clés propose l'auto complétion (cf. Figure 8.23(a), si la saisie ne correspond pas à une des propositions de l'auto complétion alors une recherche plus large est effectuée et une liste de termes similaires est affichée (cf. Figure 8.23(b)). La recherche diffère suivant la source sélectionnée.

Avec un thesaurus comme source, les termes utilisés pour l'auto complétion sont issus du thésaurus, en effet il est proposé l'ensemble des termes de celui-ci. Si le mot saisi n'appartient pas à la liste d'auto complétion, une recherche est effectuée dans l'ensemble des termes rejetés du thésaurus.

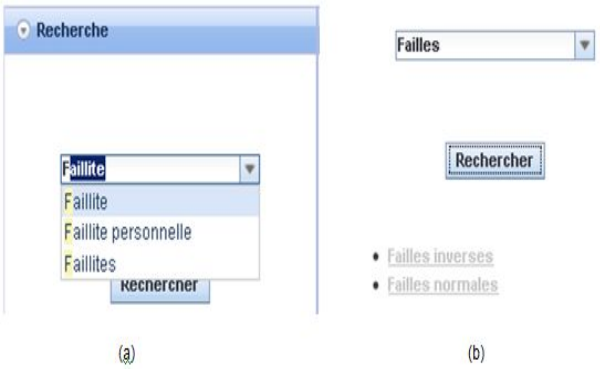


FIGURE 8.23 – Barre de recherche. (a) Auto complétion. (b) Terme non proposé avec l'auto complétion.

Si nous prenons un fonds documentaire comme source de l'application, les termes utilisés pour l'auto complétion sont les entités et les entités rejetées contenues dans le fonds documentaire, ainsi que les termes rejetés correspondant dans le thesaurus. Si le terme saisi n'est pas connu, une recherche plus large est réalisée dans l'ensemble des termes du thésaurus associé. La sélection d'un nouveau terme entraîne la génération du graphe correspondant.

8.2.4.4 Recherche à partir d'un arbre hiérarchique

Avec un thesaurus comme source de l'application, la rubrique par index comporte les termes les plus génériques contenus dans le thésaurus. Pour RAMEAU, la liste contient près de 5 000 termes de haut niveau, que nous choisissons de classer alphabétiquement dans des rubriques (A-B, C-D, etc.) pour éviter de troubler l'utilisateur avec une liste trop importante de termes de haut niveau.

Avec un fonds documentaire comme source, l'arbre hiérarchique présente les entités en tant que feuille et leurs termes génériques (dans le thésaurus) en tant que nœuds supérieurs (figures 8.24 et 8.25). Ce choix de ne reprendre que les termes génériques aux termes utilisés pour décrire les documents traités limite la liste des termes de haut niveau, liste que nous choisissons alors d'afficher telle quelle.

La sélection d'un terme dans l'index génère un nouveau graphe.

8.2.4.5 Fiche RAMEAU

L'affichage de la fiche RAMEAU d'un terme vedette est possible en effectuant un clic gauche sur un des termes du graphe, la fiche descriptive du terme est alors affichée dans un nouvel onglet (figure 8.25). Ce choix permet d'avoir un lien vers la fiche visualisable dans un onglet, tout en gardant la possibilité de naviguer dans le fonds documentaire. Il est ainsi possible de visualiser plusieurs fiches RAMEAU et de les garder disponibles pour de nouvelles recherches.



FIGURE 8.24 – Liste hiérarchique en mode fonds documentaire

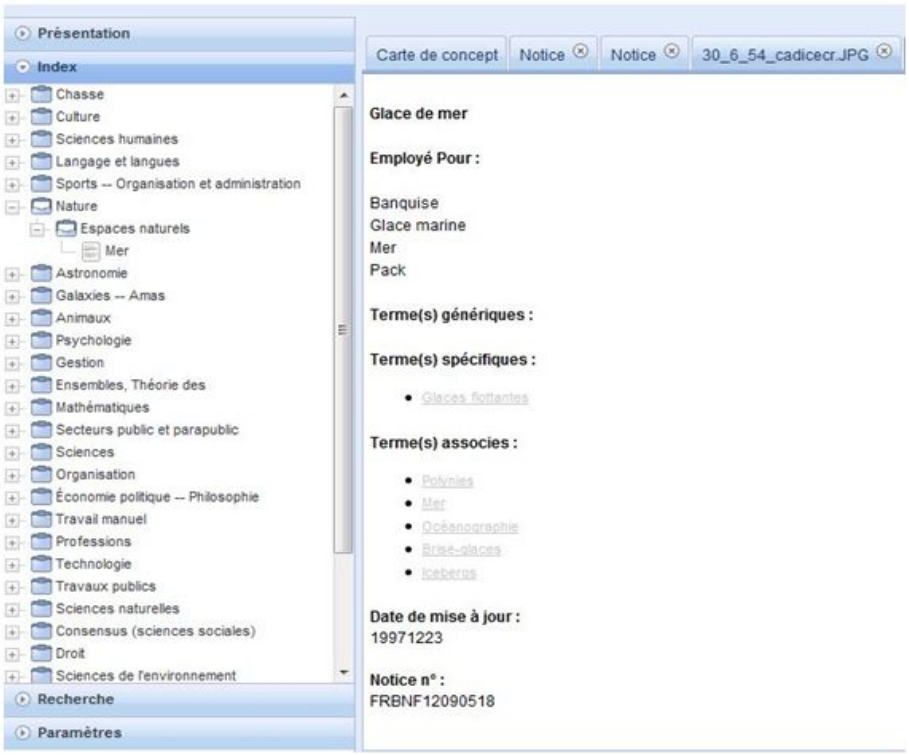


FIGURE 8.25 – Fiche RAMEAU dans TERRIDOCViewer pour le terme « Glace de mer »

8.2.4.6 Visionneuse de documents

L'affichage du document est possible via une petite visionneuse accessible sous le graphe (cf. Figure 8.19 page 186), la sélection d'un des documents avec la souris permet sa visualisation dans un nouvel onglet (cf. Figure 8.26).

Pour l'exemple ici, le terme utilisé pour la recherche est Montagnes, ce terme est relié

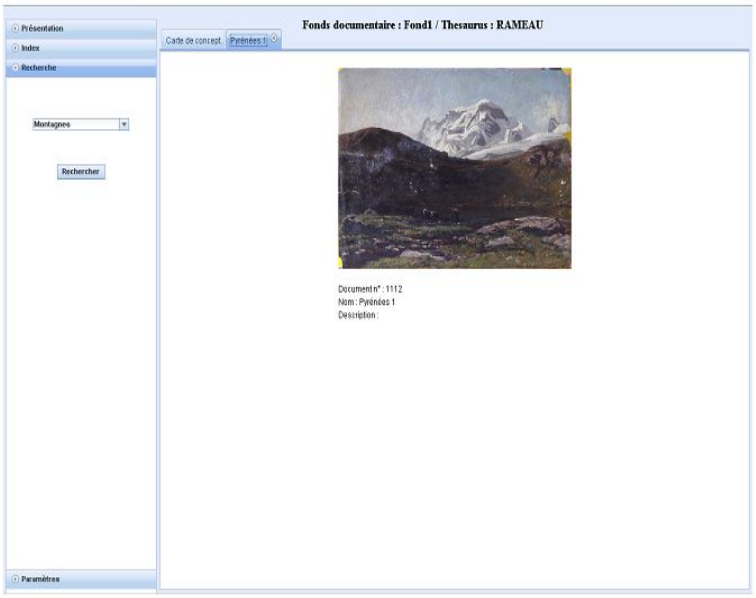


FIGURE 8.26 – Affichage d'un document

à des documents d'où la possibilité de les consulter comme on peut le voir sur la figure 8.26 décrivant une peinture représentant une montagne).

8.2.4.7 Historique

Un historique, sous forme de boîte permettant le défilement des termes qu'elle contient, visualise les termes consultés et donne la possibilité de les consulter de nouveau. Cet historique se trouve sous la carte de concepts afin que l'utilisateur puisse y accéder facilement (cf. Figure 8.19).

8.2.5 Fonctionnement de l'application

L'application doit être lancée à partir d'un navigateur Web. La page principale comporte deux parties, un accordéon sur la gauche contenant les différentes rubriques (« Présentation », « Index », « Recherche », « Paramètres ») et un onglet graphe qui permet d'afficher le graphe de termes. Initialement, un message d'accueil est présent, indiquant la marche à suivre à l'utilisateur. L'utilisateur peut effectuer les actions suivantes :

- Consulter les informations sur l'application dans la rubrique « Présentation », telles que l'aide par exemple ;
- Paramétrer l'application via la rubrique « Paramètres » ;
- Sélectionner un terme dans la rubrique « Recherche » ou « Index » afin de générer un graphe de termes sur l'onglet principal.
-

Après sélection d'un terme dans *l'Index* ou la *Recherche*, un graphe est généré (cf. Figure 8.19). L'utilisateur peut par la suite naviguer directement dans le graphe en double cliquant sur un terme ou bien visualiser la fiche RAMEAU d'un terme en cliquant dessus. Si un terme est relié à des documents, une icône permet de le signaler. Un clic sur cette icône permet l'affichage d'une visionneuse sous le graphe, l'utilisateur peut à présent consulter les documents individuellement dans un nouvel onglet (cf. Figure 8.26 page 191).

8.2.6 Premiers bilans

La structure proposée pour la base de données thesaurus permet, pour un grand groupe industriel par exemple, de traiter plusieurs thésaurus (chacun étant utilisé dans un service spécifique), et pour chaque thésaurus, il est ainsi possible de représenter plusieurs fonds documentaires. Nous définissions une troisième base permettant de gérer les paramètres généraux de l'application. A partir de ces bases de données, l'application TERRIDOCViewer propose différents modules de recherches dans des fonds documentaires annotés.

Les premiers retours d'expériences des experts de la médiathèque sont encourageants, voyant ici la possibilité de visualiser de façon globale et dans des temps intéressants leur travail d'indexation. La structure définie sous forme d'ontologie représente en effet une base de connaissances regroupant le travail de l'ensemble des bibliothécaires sur un fond documentaire donné, ce qui était jusqu'alors difficile à représenter étant donné le nombre important de documents et de notices descriptives associées. De plus, les bibliothécaires étant régulièrement amenés à travailler sur des sites distants les uns des autres (7 sites à ce jour dépendent de Pau), une telle base de connaissances facilite les échanges lors des phases de description et d'indexation du fonds documentaire.

Cependant, en l'état actuel, TERRIDOCViewer ne permet pas de mettre en avant les informations géographiques caractérisant un territoire. La composante thématique est traitée comme le montrent les figures 8.19 et 8.21 mais la représentation sous forme de graphe de termes ne suffit pas lorsque le besoin est de valoriser un territoire.

8.3 Vers une application d'aide à l'indexation de documents pour les experts

Suite à une présentation de TERRIDOCViewer auprès de personnels de la MIDR, les bibliothécaires sont en mesure d'appréhender de façon globale puis détaillée l'ensemble du travail d'indexation réalisé et d'identifier ainsi plus aisément les erreurs d'indexation possibles. Le processus automatisé visant à construire une ontologie du territoire intègre des phases de contrôle des notices descriptives qui nous permettent de citer quelques limites dues à la saisie opérée par les experts de la médiathèque. Certaines associations ne pourront hélas être explicitées automatiquement, certains documents ont subi une phase de description mais la phase d'indexation a été omise. Ainsi, ces documents ne sont pas décrits par des vedettes provenant de RAMEAU et ne pourront être reliés à notre thésau-

rus dans cette étape. De plus, nous avons identifié des erreurs de saisies (orthographe, pluriels, termes exclus) ainsi que des termes ne semblant pas provenir du thésaurus. Un travail plus approfondi doit nous permettre d'exploiter au mieux ces connaissances cachées. Nous précisons dans cette partie les différentes que nous identifions lors de la création de l'ontologie et nous initions un réflexion sur un module semi-automatisé d'aide à la correction du travail d'indexation réalisé.

8.3.1 Un premier contrôle

Lors de la création automatique du thésaurus présentée précédemment, une phase de vérification est lancée de façon automatique afin de s'assurer que les termes utilisés dans les notices descriptives pour indexer le document sont bien présents en tant qu'autorité-matière dans RAMEAU. Nous avons ainsi pu identifier quatre cas d'erreurs liées à l'utilisation de la ressource RAMEAU :

1. **Notices vides d'indexation** Il apparaît que des notices ne contiennent pas de descriptifs du contenu utilisant la ressource RAMEAU. Il est donc impossible de lier les documents finaux décrits par ces notices dans notre cartographie ;
2. **Nommage erroné des vedettes** : Des erreurs de nommage des vedettes sont apparues dû à une mauvaise utilisation de la casse et du singulier/pluriel : le langage d'indexation impose des règles de désignation de concepts en utilisant le pluriel pour désigner ce qui est dénombrable (par exemple *Montagnes*, *Châteaux*, etc.) et le singulier pour ce qui ne l'est pas (*Paysage* ou encore *Mer*). Il est possible cependant qu'un même terme puisse figurer dans la liste au singulier et au pluriel, le singulier désignant le concept abstrait (par exemple *Cinéma* pour désigner le 7e art) et le pluriel désignant les notions concrètes (par exemple *Cinémas* pour désigner les salles de cinéma). Un nombre important d'erreurs de ce type est identifié dans l'utilisation de la casse et du pluriel (Figure 8.27).
3. Gestion des termes exclus : En tant que langage contrôlé, RAMEAU limite l'usage du vocabulaire en proscrivant l'emploi des termes considérés comme équivalents (synonymie et quasi-synonymie). Parmi ces termes équivalents, une seule forme est retenue pour représenter le concept que l'on définit comme vedette. Elle est le seul terme autorisé pour l'indexation. Les termes exclus peuvent être des synonymes ou des quasi-synonymes (mots de sens voisin, ou ayant entre eux des rapports d'inclusion) et sont reliés à la vedette par la relation « employé pour ». Nous avons identifié des termes définis comme exclus mais utilisés pour décrire les documents (par exemple Maisons qui devrait être remplacé par la vedette Habitations) ;
4. Gestion de l'homonymie : Une vedette doit représenter un seul concept. Cependant un terme peut avoir différents sens et la distinction des homonymes se fait en utilisant des qualificatifs (par exemple *Grues (appareils)* et *Grues (animaux)*). Actuellement, nous savons distinguer l'ensemble des propositions possibles provenant de RAMEAU lorsque le terme est utilisé dans la notice sans ces qualificatifs alors qu'ils sont nécessaires (le terme *Grues* ne peut être utilisé seul dans RAMEAU par exemple).

La figure 8.27 présente les résultats de la phase de vérification effectuée sur 900 documents.

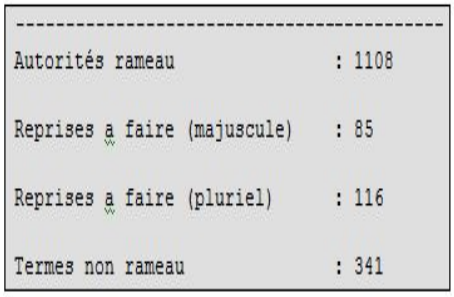


FIGURE 8.27 – Résultat du traitement de vérification pour validation de l'indexation

8.3.2 Propositions : vers une aide à la correction semi-automatisée

Une fois identifiés, les cas d'erreurs sont stockés pour une phase de correction ultérieure via l'interface Web.

A partir de l'application TERRIDOCViewer permettant de naviguer dans le fonds documentaire présenté chapitre 8.2 page 170, il est possible d'accéder via un onglet à une liste de termes identifiés (d'après les différents cas d'erreurs énoncés précédemment) comme ne faisant pas partie de RAMEAU. Pour chaque terme, le nombre d'occurrence du cas d'erreur est affiché. En ce qui concerne les cas liés à des erreurs de nommage et à l'utilisation de termes exclus, des solutions sont proposées. Lorsque l'usager, par exemple, en parcourant la liste des termes, souhaite corriger le terme *Grues* qui y figure, il clique alors sur le terme défini comme un lien afin d'accéder à une nouvelle page listant la ou les solution(s) proposée(s). Dans ce cas, une liste de 3 propositions lui sera fournie : *Grues (appareils)*, *Grues (animaux)*, et *Ajout à la liste de termes candidats*. L'usager peut alors faire son choix pour correction et validation. Cette approche n'est que partiellement satisfaisante car pour un terme polysémique tel que montagne, le bibliothécaire se voit alors proposer une liste de plus de 50 propositions.

La correction d'un terme peut être faite de façon locale (pour une notice en particulier) dans le cas où l'utilisation du terme se veut spécifique au document traité ; ou de façon globale lorsque l'erreur est le pluriel (par exemple à l'autorité *Montagnes*) et que l'on souhaite reporter cette correction sur l'ensemble des notices descriptives.

Il est possible d'indexer des documents en utilisant des termes ne faisant pas partie de RAMEAU. Prenons le cas d'une photographie de la *Place Royale à Pau*, l'expert qui effectue l'indexation pourra choisir d'utiliser le terme *Place Royale (Pau)*, terme spécifique au fonds documentaire que le comité chargé de maintenir RAMEAU n'a pas incorporé au thésaurus. Ce terme apparaîtra dans la liste de cas d'erreurs car il n'existe pas de correspondance dans RAMEAU. Lors de sa correction, l'usager peut alors choisir de l'ajouter à la liste de termes candidats. Cette liste apparaît comme une nouvelle source de connaissances sur laquelle pourront s'appuyer à l'avenir les bibliothécaires pour de

nouvelles tâches d'indexation. Cette liste de termes dits candidats peut aussi être formatée et proposée au comité de décision de la BnF en tant que demande d'enrichissement du thésaurus RAMEAU.

Concernant les cas d'erreurs liés à l'homonymie, nous limitons l'assistance à la correction en proposant, comme pour d'autres cas d'erreurs, une liste de termes contenant le terme provoquant un cas d'erreur (par exemple *Grues* sans qualifiant supplémentaire). Il est envisageable de prévoir des solutions tentant d'identifier, en calculant la fréquence des qualifiants possibles (appareils et animaux pour l'exemple de *Grues*) la solution qui semble la plus probable dans la notice descriptive voire même, pour les documents texte, dans le contenu même du document.

8.3.3 premiers bilans liés à l'analyse du travail d'indexation

Tous les cas recensés permettent de créer et d'enrichir une base d'expérience que nous pouvons exploiter. D'une part, cette base de connaissance peut être utilisée pour assister les experts pour identifier les erreurs lors de l'analyse a posteriori des notices descriptives produites. D'autre part, cette base de connaissance peut servir de point de départ pour mettre en place des scénarios pédagogiques afin de faciliter le travail des bibliothécaires néophytes dans l'apprentissage de l'utilisation de RAMEAU pour indexer des documents. Nous travaillons actuellement au développement d'une extension de TERRIDOCViewer visant à proposer à des experts un workflow complet intégrant des interfaces Web pour les aider à corriger un ensemble de notices descriptives.

Cinquième partie

Conclusion générale

Chapitre 9

Synthèse et perpectives

Sommaire

9.1 Synthèse générale 199
9.1.1 Modélisation d'un territoire 200
9.1.2 Choix de l'ontologie pour la représentation sémantique d'un territoire 202
9.1.3 Découverte incrémentale d'un territoire à partir de documents annotés 203
9.2 Usages mis en place et perspectives applicatives 207
9.2.1 Application TerridocViewer pour la RI 207
9.2.2 Enrichissement d'ontologie géographique et améliorations de l'indexation spatiale de fonds documentaires 209
9.3 Perspectives scientifiques 209
9.3.1 Prise en compte des relations spatiales pour l'enrichissement de l'ontologie d'un territoire 209
9.3.2 Création d'une ontologie d'un territoire adaptée à plusieurs fonds documentaires 211
9.3.3 Perspective à plus long terme : Application de la méthodologie Terridoc dans une autre langue 211

9.1 Synthèse générale

Comme la majorité des bibliothèques et des médiathèques en France, la MIDR de Pau a comme souci majeur de structurer au mieux la connaissance présente dans ses fonds documentaires afin d'en faciliter l'accès et le partage par une large communauté d'utilisateurs. Une caractéristique importante de ce type de fonds documentaire est qu'ils contiennent d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire, et il est primordial de valoriser ces spécificités territoriales pour répondre à ces objectifs d'information et d'éducation.

Les travaux présentés dans ce manuscrit se placent dans le vaste domaine de la gestion de connaissances appliquée à la géographie, domaine connu sous le nom de géomatique. La gestion de connaissances est l'ensemble des initiatives et des techniques permettant d'identifier, d'analyser, d'organiser, et de partager des connaissances entre les membres des organisations. La géomatique, domaine d'application de notre approche, a pour objet la gestion des données à référence spatiale et qui fait appel aux sciences et aux technologies reliées à leur acquisition, à leur stockage, à leur traitement et à leur diffusion. Dans ce cadre, nous nous intéressons plus précisément aux trois axes de recherche que sont l'Extraction et la Structuration de la Connaissance, l'Organisation des Connaissances, le Traitement Automatique du Langage Naturel s'appuyant sur des ressources de type Gazetteer pour le traitement des données géographiques sur les Systèmes d'Information Géographique. La problématique choisie, de l'extraction et de la structuration d'informations géographiques contenues dans des fonds documentaires indexés, nécessite en effet une approche pluridisciplinaire afin de mener à bien des propositions de solutions efficaces. Notre démarche vise à proposer une méthodologie opérationalisée permettant de définir de façon automatique une ontologie légère d'un territoire implicitement décrit dans des fonds documentaires indexés. Nous faisons l'hypothèse que l'extraction d'informations locales dans les documents et les notices descriptives, avec un filtre géographique, fait émerger une représentation territorialisée de la base documentaire traitée.

Pour ce faire, nous proposons tout d'abord une définition de ce que nous entendons par territoire en nous appuyant sur les travaux de recherche des géographes, linguistes et informaticiens. Avant de synthétiser notre méthodologie incrémentale visant à définir une représentation d'un territoire à partir de documents annotés, nous précisons notre choix d'utiliser l'ontologie pour normaliser la représentation du territoire obtenue. Nous décrivons ensuite la méthodologie que nous appliquons sur un fonds documentaire de la MIDR de Pau et nous faisons un état des usages que nous avons expérimenté. Nous terminons en présentant les perspectives applicatives et scientifiques liées à nos travaux.

9.1.1 Modélisation d'un territoire

Dans le cadre de cette thèse, nous proposons tout d'abord une modélisation de ce que nous entendons par la notion de territoire (cf. section 5.2 page 88). Un état des travaux réalisés dans différents domaines montre que la définition de la notion de territoire est encore sujette à discussion. Nous nous appuyons sur les travaux des géographes pour qui la notion de territoire est au cœur de leurs préocupations. Il existe au niveau mondial une définition conventionnelle, décrivant le territoire comme un espace sur lequel s'exerce une autorité limitée par des frontières politiques et administratives. De nombreux travaux [Pio92, Sch94, LL03, GV04, DMB05, Gui07] enrichissent cette définition en mettant en avant les pratiques sociales de l'Homme dans un espace ainsi que la relation entre les composantes thématique (coutumes, activités, etc.), spatiale (ensemble de lieux) et temporelle (periode plus ou moins longue). Cependant, peu de travaux cherchent à représenter le territoire en mettant en avant ces pratiques à travers la composante thématique [LMP01]. A l'inverse, de nombreux travaux, que ce soit en géographie [DS05, UTC04], en sciences cognitives et en psychologie [Tol48, Lyn60, Den97]

et plus récemment en informatique [Gai01, PSA07], s'appliquent à traiter la notion d'information géographique en vue de proposer des représentations administratives ou politiques d'espaces géographiques, mettant ainsi de côté l'aspect culturel. [UTC04] définit l'information géographique (ou entité géographique) comme la relation entre une entité thématique, une entité spatiale et une entité temporelle. Cette définition est maintenant admise dans de nombreux travaux et notamment en informatique [UTC04, PSA07] et dans notre équipe [Gai01, Mal03, Les07]. Les trois composantes (spatiale, thématique et temporelle) faisant partie intégrante de la notion de territoire, nous nous appuyons sur cette définition pour proposer une modélisation de la notion de territoire. Dans ce travail, nous ne nous limitons pas à une représentation politique ou administrative car nous cherchons également à identifier et modéliser des élements liés à la culture (activités, évènements, coutumes, etc.) pour un espace géographique donné.

9.1.1.1 Définition de ce que l'on entend par territoire

Nous nous accordons avec [LMR93, Les07] qui précisent la définition d'information géographique comme une entité thématique ancrée sur une localisation spatiale explicite et, régulièrement, temporelle. Cette définition met à la fois en avant l'importance de la composante spatiale et ajoute que l'entité temporelle n'est pas toujours explicite et du coup plus difficile à identifier dans le contenu des documents.

Sur la base des travaux de [DMB05] et [Gui07] qui s'attachent à mettre en valeur la notion de lieux au sein d'un territoire, nous pouvons affirmer tout d'abord que : *le territoire décrit dans un fonds documentaire correspond à un réseau défini par des lieux et par les relations identifiées entre ces lieux.* Cette première définition fait intervenir la composante spatiale, à travers des lieux, comme un élément important.

Sur la base des travaux de [Ent96, BE98], [DMB05] décrit le territoire comme *le contact vécu de l'homme avec le milieu* et intègre les composantes thématique (ce qui a trait à l'Homme) et temporelle (le contact vécu instaure une notion de temps plus ou moins long) dans la définition d'un territoire. Nous nous appuyons sur ces travaux pour proposer un modèle de la notion de territoire, définie comme la composition d'informations géographiques (cf. figure 9.1).

Dans ce modèle, nous définissons la composante spatiale comme un nom toponymique, et la composante temporelle comme une information calendaire. La composante thématique est définie dans un contexte comme un groupe nominal « qualifiant » une entité spatiale. Le contexte ici est défini sur la base du travail d'indexation d'experts en s'appuyant sur un vocabulaire contrôlé riche (de type taxonomie, thésaurus ou ontologie). Cela nous permet de penser que l'extraction et la structuration de ce travail d'expertise, en appliquant un filtre géographique, fait émerger une représentation d'un territoire.

Le modèle proposé permet d'intégrer un ensemble d'informations géographiques formant une représentation administrative d'un territoire (« les montagnes des Pyrénées-Atlantiques », « la commune de Pau », etc.) et apporte également des éléments de réponses dans la représentation des activités sociales de l'Homme dans un espace. Prenons l'exemple des informations géographiques « l'alpinisme au Pic D'Ossau » et « rugby à

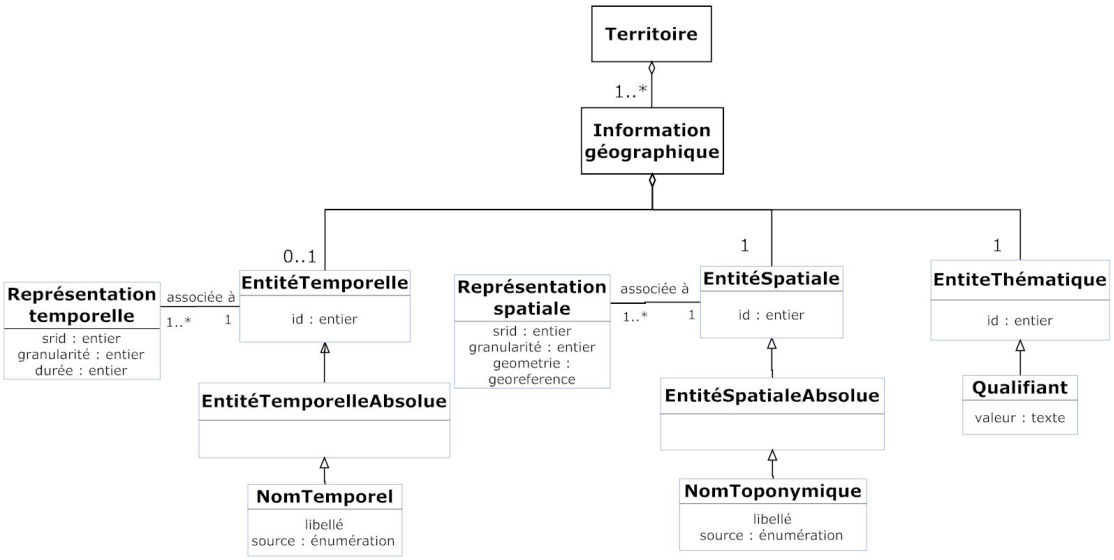


FIGURE 9.1 – Modélisation du territoire

Pau » présentes dans le fonds documentaire traité, « Pic d'Ossau » et « Pau » sont deux lieux sur lesquels sont ancrés respectivement les termes « alpinisme » et « rugby », tous deux présents dans le vocabulaire contrôlé sur lequel nous nous appuyons. Dans notre démarche, ces deux termes nous permettent d'expliciter les activités sportives dans les Pyrénées-Atlantiques.

9.1.2 Choix de l'ontologie pour la représentation sémantique d'un territoire

Rappelons tout d'abord l'objectif lié à ces travaux de recherche. Nous souhaitons, en nous appuyant sur le travail d'annotation d'experts bibliothécaires, modéliser et structurer la connaissance mettant en avant un territoire implicitement décrit dans un fonds documentaire. Les annotations des experts, réalisées sur la base d'un vocabulaire contrôlé, sont des informations porteuses de sens qu'il nous semble important de prendre en compte.

Parmi les outils existants permettant de structurer la connaissance que sont les Topic Maps, les folksonomies, ou encore les vocabulaires contrôlés tels que les thésaurus ou les ontologies, nous avons choisi d'utiliser les ontologies. Le fait que les annotations expertes soient extraites d'un vocabulaire contrôlé nous incite à choisir une structure du même type afin d'exploiter au mieux le travail d'expert notamment sur le choix de la sémantique, de la terminologie et la syntaxe des termes sélectionnés pour décrire des documents. Aussi, en tant que vocabulaire contrôlé, l'ontologie possède des règles sémantiques, terminologiques et syntaxiques qui permettent de modéliser et structurer la connaissance de façon précise dans un domaine cible. Contrairement au thésaurus,

l'ontologie permet de préciser la définition des concepts, avec notamment la possibilité de leurs attacher des propriétés et des relations qu'il est possible d'expliciter. Aussi, l'ontologie permet d'effectuer des inférences, limitant ainsi la quantité d'informations à structurer. Si nous reprenons l'exemple des châteaux, indiquer dans l'ontologie qu'un « château fort » est un « château » implique indirectement qu'un « château fort » est un « monument ». Ce modèle sémantique est alors un moyen efficace de modéliser de façon précise la connaissance d'un domaine et de l'utiliser dans diverses applications. Nous nous accordons avec B. Bachimont pour dire que l'usage prévu de l'ontologie contraint et encadre sa construction pour un domaine cible [Bac00,BAG03]. Parmi les différents types d'ontologies, nous positionnons nos travaux dans la construction d'une ontologie légère de domaine décrivant un territoire. L'ontologie légère est composée d'un lexique, de concepts et de relations hiérarchiques et associatives entre concepts et permet malgré un niveau formel moindre de proposer une représentation de ce que nous entendons par territoire. Ce type d'ontologie permet également de mettre en place des mécanismes de recherche élaborés. L'utilisation d'ontologies lourdes a été envisagée, mais leur élaboration très coûteuse et les difficultés à réutiliser l'ontologie lourde pour d'autres tâches que celle pour laquelle elle est définie nous a conforté dans le choix de construire des ontologies légères.

9.1.3 Découverte incrémentale d'un territoire à partir de documents annotés

Un état des travaux présenté section 4.4.1 (cf. page 80) montre qu'il n'existe pas à notre connaissance une méthodologie complète et automatisée permettant de construire un modèle formel offrant une représentation sémantique d'un territoire à partir d'un ensemble de documents indexés. La méthodologie TERRIDOC que nous proposons de mettre en place se décompose en quatre étapes principales que nous schématisons figure 9.2 et que nous résumons ci après.

L'approche se veut générique et applicable sur différents types de fonds documentaires indexés. L'un des intérêts de l'approche est de pouvoir proposer de façon incrémentale une représentation sémantique d'un territoire après chacune des étapes. Les utilisateurs sont ainsi à même de choisir parmi les différentes représentations celle qui correspond le mieux à leurs attentes. Dans le cadre de nos expérimentations, nous nous appuyons sur un fonds documentaire hybride de la MIDR de Pau constitué de 900 documents notices Xml associées, sur le thésaurus RAMEAU utilisé par les experts pour réaliser les notices descriptives ainsi que sur les bases de données BDTopo et BDCarto mises à disposition par l'IGN, et servant de lexiques toponymiques pour valider les entités nommées identifiées dans les documents. Nous présentons maintenant les quatre étapes constituant notre approche schématisée figure 9.2 en faisant un état rapide des expérimentations réalisées.

– **Construction d'une première ontologie d'un territoire à partir de la connaissance experte** : Nous nous appuyons sur le travail d'indexation d'experts réalisé à partir d'un vocabulaire contrôlé car, comme [Sve86,APC01], nous pensons qu'il apporte un plus sémantique indéniable à la représentation du domaine étudié.

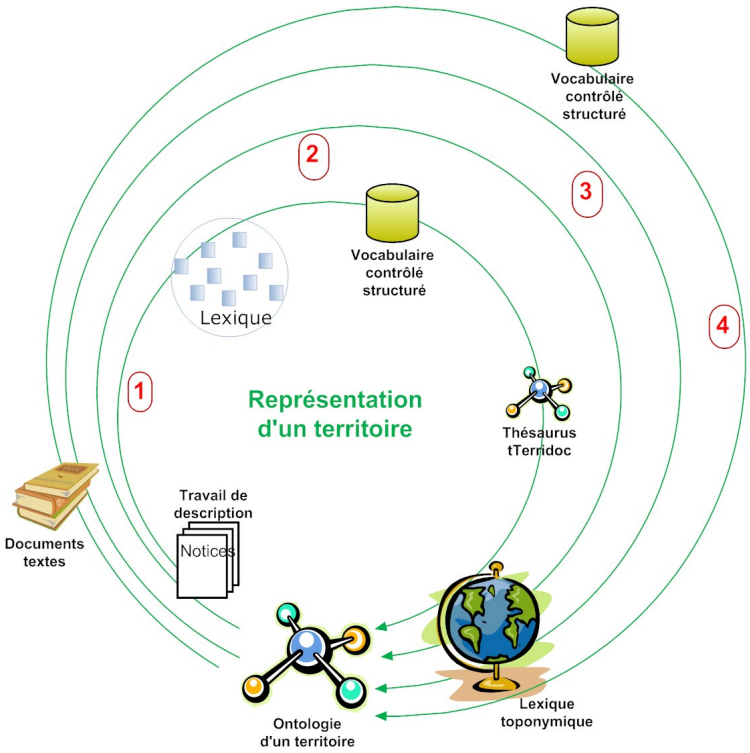


FIGURE 9.2 – Méthodologie incrémentale de construction d'ontologie d'un territoire à partir de documents annotés

Parmi les différentes méthodologies existantes dans la communauté pour construire une ontologie de façon automatique, nous mettons en application la méthodologie proposée dans les travaux de [Her05] et [CHGM06] qui s'appuient eux mêmes sur la méthodologie TERMINAE composée des cinq étapes suivantes : (i) analyse des besoins ; (ii) constitution du corpus ; (iii) analyse linguistique ; (iv) normalisation en réseau sémantique ; (v) formalisation du réseau sémantique. A la différence de ce qui est présenté dans les travaux de recherche [SLL+04, Her05, CHGM06, SGD04], nous nous appuyons sur le travail d'indexation des experts pour définir de façon automatique le thésaurus tTerridoc de départ sur lequel nous nous appuyons ensuite pour créer une ontologie de domaine. Le thésaurus tTerridoc obtenu dans une étape intermédiaire, que nous préconisons de formaliser en SKOS, offre une représentation sémantique structurée synthétisant le travail d'indexation des experts bibliothécaires qui peut être utilisée telle quelle pour analyser ou encore valider le travail réalisé. Ensuite, nous réalisons la tranformation du thésaurus en ontologie en 3 étapes : (i) regroupement des concepts, (ii) traitement des relations hiérarchiques, (iii) traitement des relations associatives. Nous préconisons de formaliser l'ontologie obtenue après transformation du thésaurus tTerridoc en OWL (en version OWL-Lite) car il offre plus de vocabulaire pour décrire les propriétés et les

classes. Nous utilisons enfin une ressource de type gazetteers afin d'identifier dans l'ontologie l'ensemble des concepts qui sont des entités spatiales. Dans l'ontologie obtenue, les concepts correspondent aux entités thématiques et les instances à des entités géographiques. La ressource gazetteer apporte également des informations sur les informations géographiques que nous choisissons de stocker en tant que propriété des instances correspondantes. Cette étape d'instanciation permet par la même occasion de désambigüiser une partie des relations hiérarchiques provenant du thésaurus en transformant la relation concernée en « instance de ».
A partir des expérimentations réalisées sur l'ensemble des 900 notices fournies par la MIDR, nous obtenons une première ontologie constituée de 481 concepts reliés à 26 instances (entités spatiales). L'ontologie obtenue offre une première représentation d'un territoire. Si l'on souhaite proposer une représentation complète d'un territoire décrit dans l'ensemble du fonds documentaire (documents + notices), cette représentation offre une vue très générale de l'espace géographique décrit dans le fonds documentaire. Cela s'explique par le fait que les spécificités spatiales se résument à une liste d'entités géographiques choisies par les bibliothécaires pour décrire de façon très générale le contenu des documents, ce qui reste relativement limité en informations.

– **Chaîne de traitement linguistique pour l'identification d'entités géographiques à partir du travail d'indexation d'experts** : Une contribution importante de notre travail concerne les étapes permettant d'enrichir de façon incrémentale la première ontologie de territoire obtenue en s'appuyant sur les documents textes et notices descriptives attachées. Pour cela, nous définissons une chaîne de traitement linguistique automatisée (cf. figure 9.3) qui permet de marquer dans des documents textes l'ensemble des EGs (nom toponymique + qualifiants). La chaîne de traitement intègre l'ontologie construite sur la base du travail des experts afin d'identifier des informations géographiques faisant le lien entre les concepts de l'ontologie et des lieux, propres à un territoire. Lorsqu'un lien est identifié entre un concept et un lieu dans un texte, le lieu est ajouté à l'ontologie sous forme d'instance du concept concerné par la relation, l'instance étant l'information géographique provenant du texte.
L'application de la chaîne de traitement linguistique sur les 900 notices descriptives attachées aux documents constituant notre corpus de référence nous permet d'identifier un ensemble de 115 nouvelles EGs qui viennent enrichir la représentation du territoire. L'ontologie obtenue offre, par l'ajout sous forme d'instances de ces EGs normalisées, une représentation plus précise du territoire décrit dans les notices attachées à un ensemble de documents. De plus, à cette étape, l'ensemble des documents son, image, texte et vidéo sont pris en compte par l'intermédiaire de leur notice descriptive. Cependant, malgré la prise en compte des différents constituants de la notice descriptive, la représentation que nous obtenons propose une vue encore générale du territoire décrit par le fonds traité. Le résultat peut ne pas s'avérer assez précis si l'objectif est de pouvoir faire ressortir du fonds documentaire un maximum d'entités géographiques.

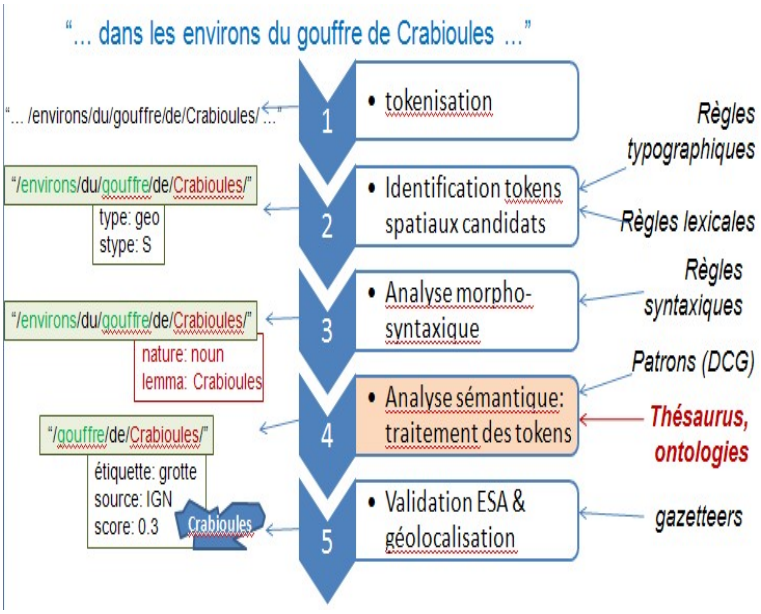


FIGURE 9.3 – Chaîne de traitement d'informations spatiales dans les documents textuels

– **Application de la chaîne de traitement linguistique pour l'identification d'entités géographiques à partir du contenu des documents** : Nous proposons dans cette troisième étape d'enrichir l'ontologie en utilisant comment point d'entrée le contenu de documents textes dits territorialisés. Rappelons que ce type de document se caractérise par une omniprésence des noms de lieux relatifs à un territoire particulier. A noter que nous ne prenons pas en compte dans cette étape les documents de type image, son et vidéo car le traitement de leur contenu implique des actions particulières qui vont bien au-delà du cadre de nos travaux de thèse. L'application de notre chaîne linguistique sur le contenu des documents nous permet d'identifier une quantité importante d'EGs. En effet, nos expérimentations réalisées sur le contenu des documents de la MIDR nous permettent d'identifier 28045 entités géographiques. Parmi ces EGs, 2967 sont attachées un label correspondant à un concept présent dans l'ontologie, et sont donc ajoutées à l'ontologie en tant qu'instances. Le reste, 25078, ne peut être traité dans cette étape. Nous obtenons une ontologie constituée de 481 concepts instanciés par 3108 informations géographiques lorsque nous appliquons notre chaîne de traitement linguistique sur le contenu des documents textes.
– **Enrichissement de l'ontologie minimale à partir du contenu des documents** : Nous remarquons en analysant les résultats obtenus que beaucoup d'EGs, identifiées via notre chaîne de traitement, ne sont pas constituées de qualifiant correspondant à un concept de notre ontologie. Cette quatrième et dernière étape intègre un processus automatisé prenant en compte ces informations afin d'enrichir notre ontologie. L'enrichissement se fait en deux étapes : élargissement de

la couverture sémantique et enrichissement ensuite de la structure de l'ontologie. L'élargissement de la couverture sémantique correspond à ajout de nouveaux labels/concepts permettant d'élargir la couverture sémantique du domaine. Dans cette étape notre chaîne de traitement TAL permet d'une part de repérer les entités géographiques candidates constituées d'un qualifiant qui n'est pas présent dans l'ontologie, et d'autre part, de sélectionner parmi ces entités le syntagme nominal candidat à l'enrichissement. Le terme est ajouté en tant que label si il se rapporte à un concept existant ou directement en tant que concept si aucun lien de « synonymie » n'est identifié. Pour cela, nous mobilisons la ressource vocabulaire contrôlé mise à disposition par les experts, afin d'associer au syntagme un ensemble d'informations permettant de mieux caractériser son champ lexical. Nos expérimentations nous permettent de prévoir à partir des 25078 entités spatiales ne pouvant être reliées à l'ontologie dans l'étape précédente, un enrichissement de 940 concepts (et potentiellement 7283 instances).
La phase d'enrichissement, bien qu'automatique, doit être validée en fin de processus par les experts bibliothécaires. L'approche présentée [KKS+09] permet d'exploiter de façon automatisée la liste des EGs pour enrichir l'ontologie d'un territoire.

9.2 Usages mis en place et perspectives applicatives

Les usages liés à l'exploitation de l'ontologie résultante de nos travaux sont nombreux. Les trois usages que nous avons expérimentés montrent l'étendue des possibilités : (i) système TerridocViewer de recherche d'information s'appuyant sur l'ontologie générée, (ii) module d'aide à la validation du travail d'indexation et (iii) module d'enrichissement d'ontologies géographiques.

9.2.1 Application TerridocViewer pour la RI

Nous avons développé sous forme d'application Web 3Tiers un système de recherche d'information que nous nommons TERRIDOCViewer [104] qui s'appuie sur l'ontologie créée pour permettre à tout type d'utilisateurs de découvrir un territoire en naviguant à travers les documents. L'intérêt majeur de cette application pour les utilisateurs visiteurs est de pouvoir appréhender dans un premier temps le fonds documentaire de façon globale (vue sémantique globale du fonds sous forme de hiérarchie de concepts) puis sous forme plus précise (vue sémantique locale sous forme de graphe structuré de termes). Un second usage de ces travaux est la possibilité pour les experts du domaine de naviguer à travers les documents via une représentation sémantique synthétisant leur travail d'indexation. De leur côté, les experts de l'équipe RAMEAU ont à leur disposition un outil offrant une représentation graphique du thésaurus RAMEAU.

Les premiers retours intéressants de l'équipe de travail RAMEAU nous ont amenés à présenter l'application TerridocViewer à ses collaborateurs à la BnF travaillant autour de l'indexation. Nous avons prévu de nous rencontrer à nouveau après la thèse pour discuter

104. http ://t2i.univ-pau.fr/

des perspectives à venir. Au niveau de la médiathèque, les premiers retours d'expériences des experts sont également encourageants, voyant ici la possibilité de visualiser de façon globale et dans des temps intéressants leur travail d'indexation. La structure définie sous forme de thésaurus représente en effet une base de connaissances regroupant le travail de l'ensemble des bibliothécaires sur un fond documentaire donné, ce qui était jusqu'alors difficile à représenter étant donné le nombre important de documents et de notices descriptives associées. De plus, les bibliothécaires étant régulièrement amenés à travailler sur des sites distants les uns des autres, une telle base de connaissances facilite les échanges lors des phases de description et d'indexation du fonds documentaire.

Nous travaillons actuellement à l'enrichissement de l'application pour intégrer l'ontologie contenant des informations liées au territoire. Nous souhaitons notamment intégrer les composantes spatiales et temporelles dans les différents modules de recherche, ce qui implique entre autre un enrichissement des interfaces. Dans ce sens, des premiers travaux ont permis d'enrichir les modules de navigation en intégrant une recherche multicritères (spatial et thématique) permettant de répondre aux requêtes du type « les châteaux au sud de Pau ». Au niveau des interfaces, nous travaillons également sur un module qui propose d'afficher en parallèle les résultats de la recherche multicritère sur une carte géographique pour indiquer l'espace représenté par les résultats de la requête, ainsi que sur une carte de concepts pour en indiquer une représentation sémantique.

Concernant la composante temporelle, un travail d'analyse des résultats, obtenus en appliquant notre chaîne de TAL sur des textes, est nécessaire.

9.2.1.1 Aide à l'indexation

Suite à une présentation de TerridocViewer auprès du personnel de la MIDR, les bibliothécaires sont en mesure d'appréhender de façon globale puis détaillée l'ensemble du travail d'indexation réalisé et d'identifier ainsi plus aisément les erreurs d'indexation possibles. Les différentes rencontres avec les experts nous ont permis d'intégrer dans le processus automatisé visant à construire une ontologie du territoire des phases de contrôle des notices descriptives afin d'enrichir une liste de cas d'erreurs possibles classifiées en quatre catégories (notices vides d'indexation, nommage erronés des autorités, utilisation de termes exclus, utilisation de termes homonymes). D'une part, cette base de connaissance peut être utilisée pour assister les experts pour identifier les erreurs lors de l'analyse « à posteriori » des notices descriptives produites. D'autre part, cette base de connaissances peut servir de point de départ pour mettre en place des scénarios pédagogiques afin de faciliter le travail des bibliothécaires néophytes dans l'apprentissage de l'utilisation de RAMEAU pour indexer des documents. Nous travaillons actuellement au développement d'une extension de TerridocViewer visant à proposer à des experts un workflow complet intégrant des interfaces Web pour les aider à corriger un ensemble de notices descriptives.

9.2.2 Enrichissement d'ontologie géographique et améliorations de l'indexation spatiale de fonds documentaires

Dans le cadre du projet Géonto [105], nous avons expérimenté notre approche visant à enrichir une ontologie géographique à partir d'un échantillon de documents textes grand public, représentatif du domaine cible. L'ontologie cible présentée dans [KAG09] est définie sur la base de spécifications de bases de données proposée par l'IGN. Ce choix s'est imposé notamment par le fait que nous souhaitions que l'ontologie s'adapte au plus près du domaine d'application. Nous avons appliqué la méthodologie automatisée définie dans nos travaux de thèse qui se décompose de la façon suivante : (i) identification et validation des EGs dans le texte ; (ii) identification des termes liés aux EGs non présents dans l'ontologie ; (iii) enrichissement de l'ontologie. Les expérimentations menées sur un ensemble de 14 livres de type récits de voyage nous ont permis d'identifier 140 termes candidats à l'enrichissement de l'ontologie.

Concernant l'amélioration de la phase d'indexation spatiale de fonds documentaire, nos expérimentations détaillées dans [KKS+09] montrent que l'utilisation de l'ontologie enrichie améliore grandement l'indexation des informations spatiales. En effet, après l'application de notre chaîne de traitement linguistique sur les 14 livres, l'analyse des termes associés aux entités nommées et l'utilisation de l'ontologie géographique générée dans le cadre du projet Géonto nous permet de qualifier/typer automatiquement 10% des informations spatiales annotées et nous obtenons 50% de typage automatique des informations spatiales après l'enrichissement de cette ontologie.

9.3 Perspectives scientifiques

9.3.1 Prise en compte des relations spatiales pour l'enrichissement de l'ontologie d'un territoire

La méthodologie visant à construire une ontologie d'un territoire intègre une étape visant à enrichir l'ontologie à partir des résultats obtenus par notre chaîne de traitement linguistique. Dans cette étape, nous sommes en mesure d'identifier de nouveaux labels ainsi que de nouveaux concepts candidats à enrichir le vocabulaire de l'ontologie. Afin d'identifier un premier ensemble de relations spatiales entre concepts, nous préconisons de nous appuyer directement sur les représentations spatiales des informations géographiques identifiées dans le corpus. A partir des représentations spatiales pouvant être identifiées entre les instances de l'ontologie, nous proposons d'identifier des relations spatiales entre les concepts attachés aux instances. L'objectif est ici d'expliciter des relations de type « termes associés » lorsqu'elles existent ou d'en créer de nouvelles dans le cas contraire. En nous appuyant notamment sur les travaux de [DMGVOM07], nous travaillons actuellement à la définition de patrons pour chaque type de relations spatiales. Prenons la relation d'adjacence permettant d'indiquer, à partir de variables prédéfinies, si une instance de concept est proche, loin ou à la périphérie d'une autre.

105. projet ANR mené dans le cadre de l'édition 2007 du programme « Masse de Données et Connaissances »

Si chaque instance d'un concept, est reliée à au moins une des instances d'un deuxième concept par une relation de type adjacence de même valeur (proche, loin de, etc.), alors une relation d'adjacence entre deux concepts est identifiée et créée.

Conditions	- C1 et C2 deux concepts
	- I1, I2, I3 les 3 instances de C1 et I4, I5 les 2 instances de C2
Actions	- Si I1, I2, « proche » de I4 et I3 « proche » de I5
	- Créer une relation d'adjacence entre C1 et C2 : C1 « est proche de » C2

Pour exemplifier cette étape, prenons les groupes nominaux suivants identifiés dans les textes pour les concepts « cours_d'eau » et « station_climatique,_thermale,_etc. » :

- « Nous avons longé **le gave de Pau**. »
- « la visite **des cures thermales de Cauterets**. »
- « La traversée du **gave d'Oloron** fut périlleuse. »
- « Nous pouvons citer les **stations thermales des Eaux-Bonnes et des Eaux-Chaudes**. »

Parmi les EGs identifiées en gras dans les exemples ci-dessus, le qualifiant *gave* est un label ajouté au concept « cours_d'eau ». *Cures thermales* et *stations climatiques* sont deux labels ajoutés au concept « station_climatique,_thermale,_etc. » (*cf.* section 9.1.3). Validés dans la chaîne TALN comme des noms_toponymiques, *Pau* et *Oloron* sont formalisés en tant qu'instances du concept « cours_d'eau ». *Cauterets, Eaux-Bonnes et Eaux-Chaudes* comme instances du concept « station_climatique,_thermale,_etc. » (cf. section 9.1.3). La dernière phase de l'enrichissement consiste alors à étudier pour chaque concept les relations spatiales existantes entre ses instances et les instances de chaque autre concept de l'ontologie. Dans l'exemple ci-dessus, nous identifions une relation de proximité (type adjacence avec une distance entre ces deux entités spatiales inférieure à une variable prédéfinie) entre « le gave de Pau » et les « cures thermales de Cauterets » et à nouveau une relation de proximité entre le « gave d'Oloron » et les « stations thermales des Eaux-Bonnes et des Eaux-Chaudes ». Une relation de proximité existe donc entre chaque instance du concept « Cours_d'eau » et une instance du concept « station_climatique,_thermals,_etc. », ce qui nous permet de proposer la création d'une relation de type adjacence avec pour qualifiant « est localisé proche de » entre ces deux concepts. Cette relation doit permettre, par calcul d'inférence, d'améliorer la recherche d'information dans le fonds documentaire. L'exemple de la relation présentée ci-dessus permet d'apporter des éléments de réponse à la requête « voie d'eaux proches des stations climatiques ». Ces propositions doivent ensuite être validées par les experts bibliothécaires. Aussi, nous souhaitons en perspective de ces travaux affiner l'analyse en ajoutant un poids à chaque relation identifiée afin d'aider l'expert lors de la phase de validation. Si nous reprenons l'exemple de la relation d'adjacence « est localisé proche de », cela consiste à ajouter un niveau de proximité (en nous appuyant sur les coordonnées des deux entités spatiales analysées). Nos travaux actuels portent enfin sur l'intégration des autres types de relations spatiales dans ce processus d'enrichissement d'ontologie de territoire. Nous cherchons ici à définir des patrons génériques de relations entre concepts.

9.3.2 Création d'une ontologie d'un territoire adaptée à plusieurs fonds documentaires

Nous souhaitons étendre nos expérimentations en appliquant notre approche sur plusieurs fonds documentaires provenant de divers centres documentaires de l'hexagone. Nous souhaitons tout d'abord pouvoir intégrer dans TerridocViewer l'ensemble des représentations sémantiques de territoires correspondantes aux fonds traités pour analyser la structure et le contenu de chacune d'elle. L'intérêt majeur est ensuite de nous appuyer sur des techniques d'alignement d'ontologies [SE05, SR09, MAM10], afin d'identifier :

- les éléments communs qui serviront de bases pour construire une ontologie d'un territoire commun/global ;
- les éléments distincts, caractérisant un territoire d'un autre.

Si nous prenons l'exemple des activités sportives « football » et « rugby », nous faisons l'hypothèse ici que l'activité « football » est référencée dans l'ensemble des fonds documentaires traités alors que l'activité « Rugby » n'est présente que dans certains fonds documentaires indiquant qu'elle n'est pratiquée que sur certains territoires. Le premier objectif est de pouvoir définir de façon automatique une ontologie de territoire national, permettant notamment de mettre en avant que le *football* est une activité s'exerçant sur l'ensemble du territoire français. Le deuxième objectif est de pouvoir identifier les éléments distincts caractérisant un territoire limité, comme par exemple l'activité « Rugby » qui ne se pratique que dans certaines régions du territoire national.

9.3.3 Perspective à plus long terme : Application de la méthodologie Terridoc dans une autre langue

L'application de notre méthodologie complète ou partielle dans les différents cas d'usages présentés dans ce mémoire et notamment dans le projet Geonto nous aide à penser que notre approche peut être proposée à l'international, en prenant bien entendu en compte des modifications des modules liés au traitement linguistique notamment.

Dans les diverses conférences et rencontres scientifiques auxquelles j'ai eu la chance d'assister, nous avons notamment pu rencontrer Elena Montiel-Ponsoda et Mari Carmen Suárez-Figueroa, toutes deux chercheuses de l'UPM (Universidad Politécnica de Madrid) qui travaillent principalement sur la définition d'une méthodologie permettant de créer semi-automatiquement une ontologie et sur l'alignement d'ontologies. L'UPM collabore actuellement avec l'IGN espagnol afin de définir une ontologie géographique à partir de la structure des bases de données. Actuellement, leur démarche est manuelle [GPGPB08]. Nous souhaitons reprendre contact avec l'UPM afin de tester et valider notre approche à partir de données provenant de fonds documentaires étrangers. Aussi, nous avons initié ensemble des travaux sur l'identification de liens entre leur première ontologie et celle définie automatiquement dans le cadre du projet Geonto, uniquement pour le sous-arbre hydrographie. Pour enrichir notre analyse, nous avons ajouté à l'analyse l'ontologie géographique britannique OrdonanceSurvey en ne prenant que la partie 'hydrography'. Les premiers résultats, intéressants, renforcent l'intérêt de notre approche visant à construire une ontologie à partir de ressources structurées et non structurées du domaine cible. Une

collaboration intéresse également les personnes de l'UPM travaillant actuellement avec l'IGN espagnol. Ces contacts nous permettent d'entrevoir l'idée d'un projet autour de la définition d'une ontologie géographique au niveau « européen ».

Bibliographie

[AAHM00] Eneko Agirre, Olatz Ansa, Eduard H. Hovy, and David Martínez. Enriching very large ontologies using the www. In *ECAI Workshop on Ontology Learning*, 2000.

[AB08] A. Auger and C. Barriere. Pattern based approaches to semantic relation extraction : a state-of-the-art. *Terminology*, 19 :1–19, 2008.

[ACFLGP03] J. C. Arpirez, O. Corcho, M. Fernandez-Lopez, and A. Gomez-Perez. Webode in a nutshell. *AI Mag*, 24 :37–47, September 2003.

[AFG03] Mohammad Abolhassani, Norbert Fuhr, and Norbert Gövert. Information extraction and automatic markup for xml documents. In Henk M. Blanken, Torsten Grabs, Hans-Jörg Schek, Ralf Schenkel, and Gerhard Weikum, editors, *Intelligent Search on XML Data*, volume 2818 of *Lecture Notes in Computer Science*, pages 159–178. Springer, 2003.

[AGBS00] Nathalie Aussenac-Gilles, Brigitte Biebow, and Sylvie Szulman. Revisiting ontology design : A methodology based on corpus analysis. In Rose Dieng and Olivier Corby, editors, *EKAW*, volume 1937 of *Lecture Notes in Computer Science*, pages 172–188. Springer, 2000.

[AGDS08] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The terminae method and platform for ontology engineering from texts. In Paul Buitelaar and Philipp Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. IOS Press, http ://www.iospress.nl/, janvier 2008.

[AGM04] Nathalie Aussenac-Gilles and Josiane Mothe. Ontologies as Background Knowledge to Explore Document Collections . In *RIAO 2004 , Avignon, 26/04/04-28/04/04*, pages 129–142. ., avril 2004.

[AGPLTP99] Julio César Arpírez, Asunción Gómez-Pérez, Adolfo Lozano Tello, and Helena Pinto. How to find suitable ontologies using an ontology-based www broker. In José Mira and Juan Sánchez-Andrés, editors, *Engineering Applications of Bio-Inspired Artificial Neural Networks*, volume 1607 of *Lecture Notes in Computer Science*, pages 725–739. Springer Berlin / Heidelberg, 1999. 10.1007/BFb0100540.

[AGPTP98] Julio César Arpírez, Asuncion Gómez-Pérez, Adolfo Lozano Tello, and Helena Sofia Andrade Pinto. $(ONTO)^2$ Agent An ontology-based WWW broker to select ontologies. In *Proceedings of the Workshop on Applications of Ontologies and Problem solving Methods at the 13th European Conference on Artificial Intelligence - ECAI'98*, Brighton, England, 1998.

[AHB01] N. Asher, D. Hardt, and J. Busquets. Discourse Parallelism, Scope, and Ellipsis. *Journal of Semantics*, 18 :1–16, 2001.

[AHV05] Michel Aurnague, Maya Hickmann, and Laure Vieu. Les entités spatiales dans la langue : étude descriptive, formelle et expérimentale de la catégorisation. In Catherine Thinus-Blanc & Jean Bullier, editor, *Agir dans l'espace*, Cognitique, pages 217–232. Editions de la Maison des Sciences de l'Homme, 2005.

[AM10] Nathalie Abadie and Sébastien Mustière. Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*, 20(2) :145–174, 2010.

[APC01] James D. Anderson and José Pérez-Carballo. The nature of indexing : how humans and machines analyze messages and texts for retrieval : part i : research, and the nature of human indexing. *Inf. Process. Manage.*, 37 :231–254, March 2001.

[Aur08] Aurnague, M. Qu'est-ce qu'un verbe de déplacement ? : critères spatiaux pour une classification des verbes de déplacement intransitifs du français. page 176, 2008.

[Bac00] Bruno Bachimont. *Engagement Sémantique et Engagement Ontologique : Conception et Réalisation D'ontologies En Ingénierie Des Connaissances*, chapter 19, pages 305–324. 2000.

[Bac04] Bruno Bachimont. Arts et sciences du numerique : Ingenierie des connaissances et critique de la raison computationnelle. phd, Universite Technologique de Compiegne, janvier 2004.

[BAG03] Didier Bourigault and Nathalie Aussenac-Gilles. Construction d'ontologies à partir de textes . In *TALN 2003 : 10e conférence sur le Traitement Automatique des Langages Naturelles , Batz-sur-Mer (F), 11/06/03-14/06/03*, pages 27–47. Université de Nantes, juin 2003.

[BAGC04] Didier Bourigault, Nathalie Aussenac-Gilles, and Jean Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA), Numéro spécial sur les techniques informatiques de structuration de terminologies, M. Slodzian (Ed.)*, 18(1/2004) :87–110, 2004.

[Ban07] Audrey Baneyx. *Construire une ontologie de la Pneumologie Aspects théoriques, modèles et expérimentations.* PhD thesis, Université Pierre et Marie Curie - Paris VI, 02 2007.

[Bar00] Lawrence W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(04) :577–660, 2000.

[Bar05] Nicolas Barbey. Guide pour l'indexation et le stockage des documents à la midr de pau. Technical report, MIDR de Pau, 2005.

[BB01] A. Burgun and O. Bodenreider. Mapping the UMLS Semantic Network into general ontologies. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 81–85, 2001.

[BBCZ95] J. Bouaud, B. Bachimont, J. Charlet, and P. Zweigenbaum. Methodological principles for structuring an ontology. In *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, pages 95–148, Août, 1995.

[BBG+07] R. Brisson, O. Boussaïd, P. Gançarski, A. Puissant, and N. Durant. Navigation et appariement d'objets géographiques dans une ontologie. In *EGC'07*, pages 391–396, 2007.

[BCEM03] Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, and Yann Mathet. Geographic reference analysis for geographic document querying. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, pages 55–62, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[BCM05] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology Learning from Text : Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, 2005.

[BE98] Vincent Berdoulay and Jean Nicholas Entrikin. Lieu et sujet. perspectives theoriques. *L'Espace geographique*, 2 :111–121, 1998.

[BE05] F. Bilhaut and P. Enjalbert. *Sémantique et traitement automatique du langage naturel*, chapter 10, Recherche d'information géographique, pages 371–406. Hermes, Lavoisier, 2005.

[Ber93] M. Bergeron. *Vocabulaire de la Géomatique*. Les publications du Québec, 1993.

[Bes06] François Besancenot. Le territoire : un espace à identifier, 20 janvier 2006.

[BFGPGP98] M. Blázquez, M. Fernández, J.M. García-Pinar, and A. Gómez-PÉrez. Building ontologies at the knowledge level using the ontology design environment. In *11th Knowledge Acquisition Workshop (KAW'98)*, Banff, Canada, 1998.

[BFT94] R. Brunet, R. Ferras, and H. Thery. *Les mots de la géographie*. Collection Dynamiques du territoire. Reclus-La documentation française, 1994.

[BGKS10] Marie-Noëlle Bessagnet, Mauro Gaio, Eric Kergosien, and Christian Sallaberry. Extraction automatique d'un lexique à connotation géographique à des fins ontologiques dans un corpus de récits de voyage.

In *Conférences sur le Traitement Automatique des Langues Naturelles, Montréal, 19-23/07/2010 TALN 2010 - 17e Conférence sur le Traitement Automatique des Langues Naturelles*, page 10 pages, Montréal Canada, 07 2010.

[BGM96] Stefano Borgo, Nicola Guarino, and Claudio Masolo. Stratified ontologies : the case of physical objects. In *ECAI-96 Workshop on Ontological Engineering*, pages 5–15, 1996.

[BGMG96] Didier Bourigault, Isabelle Gonzalez-Mullier, and Cécile Gros. Lexter, a natural language processing tool for terminology extraction. In *Seventh EURALEX International Congress on Lexicography In (EURALEX96), Part II*, pages 771–779, Chichester, UK, 1996.

[BHGS01] Sean Bechhofer, Ian Horrocks, Carole Goble, and Robert Stevens. Oiled : A reason-able ontology editor for the semantic web. In Franz Baader, Gerhard Brewka, and Thomas Eiter, editors, *KI 2001 : Advances in Artificial Intelligence*, volume 2174 of *Lecture Notes in Computer Science*, pages 396–408. Springer Berlin / Heidelberg, 2001. 10.1007/3-540-45422-5_28.

[BHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5) :34–43, 2001.

[Bil06] F. Bilhaut. *Analyse automatique de structures thématiques discursives - Application à la recherche d'information.* PhD thesis, Université de Caen, 2006.

[BIT02] Bruno Bachimont, Antoine Isaac, and Raphaël Troncy. Semantic commitment for designing ontologies : A proposal. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 114–121, London, UK, 2002. Springer-Verlag.

[BKDB07] Marie-Noëlle Bessagnet, Eric Kergosien, and Alain Du Boisdulier. Vers une indexation sémantique d'images dans un fonds iconographique territorialisé. In *XXVème congrès Inforsid*, pages 327–343, Palais des Congrès, Perros-guirec, 22700, 2007.

[BKG09] Marie-Noëlle Bessagnet, Eric Kergosien, and Mauro Gaio. Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus. In *Patrimoine 3.0 CIDE'12 : 12e Colloque International sur le Document Electronique*, pages 275–286, Montréal Canada, 10 2009. Europia.

[BL02] Didier Bourigault and Guiraude Lame. Analyse distributionnelle et structuration de terminologie. application a la construction d'une ontologie documentaire du droit. *Traitement automatique des langues*, 43(1), 2002.

[BLC96] A. Bernaras, I. Laresgoiti, and J. Corera. Building and reusing ontologies for electrical network applications. In Wolfgang Wahlster, editor, *12th*

European Conference on Artificial Intelligence (ECAI'96), pages 298–302, Chichester, UK, 1996. John Wiley and Sons.

[BLCL+94] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The world-wide web. *Commun. ACM*, 37(8) :76–82, 1994.

[BMRS06] Delphine Battistelli, Jean-Luc Minel, and Sylviane R. Schwer. Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *Traitement Automatique des Langues*, 47(3) :11–37, 2006.

[BN03] Franz Baader and Werner Nutt. The description logic handbook. chapter Basic description logics, pages 43–95. Cambridge University Press, New York, NY, USA, 2003.

[Bor97] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse.* PhD thesis, Enschede, September 1997.

[Bor98] A. Borillo. *L'espace et son expression en français.* L'essentiel. Ophrys, 1998.

[BP06] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy, 2006.

[Bra77] Ronald J. Brachman. What's in a concept : structural foundations for semantic networks. *International Journal of Man-Machine Studies*, 9(2) :127 – 152, 1977.

[Bri93] D. Brill. *LOOM Reference Manual (Version 2.0).* University of Southern California, Marina del Rey (CA), USA, 1993.

[Bri04] Lauent Brisson. Mesures d'intérêt subjectif et représentation des connaissances. Master's thesis, Université de Nice, Sophia Antipolis,France, 2004.

[Bro04] J. Brodeur. *Interopérabilité des données géospatiales : Élaboration du concept de proximité géosémantique.* PhD thesis, Université Laval, Québec, 2004.

[BS99] Brigitte Biebow and Sylvie Szulman. Terminae : A linguistic-based tool for the building of a domain ontology. In Dieter Fensel and Rudi Studer, editors, *EKAW*, volume 1621 of *Lecture Notes in Computer Science*, pages 49–66. Springer, 1999.

[Can13] Augustin Pyramus de Candolle. *Theorie elementaire de la botanique, ou, Exposition des principes de la classification naturelle et de l'art de decrire et d'etudier les vegetaux / par A. P. de Candolle.* Deterville, Paris, 1813.

[CBBZ96] J. Charlet, B. Bachimont, J. Bouaud, and P. Zweigenbaum. Ontologie et réutilisabilité : expérience et discussion. *Aussenac-Gilles N., Laublet P. et Reynaud C., coordinateurs, Acquisition et ingénierie des connaissances : tendances actuelles*, pages 69–87, 1996.

[CBT05] Jean Charlet, Bruno Bachimont, and Raphaël Troncy. Ontologies pour le web sémantique. *Noûs*, Hors série, 2005.

[CDMV89] Ghislaine Chartron, Sylvie Dalbin, Marie-Gaëlle Monteil, and Monique Vérillon. Indexation manuelle et indexation automatique : dépasser les oppositions. *Documentaliste*, 26(4-5) :181–187, juillet-octobre 1989.

[CFLGP03] O. Corcho, M. Fernandez-Lopez, and A. Gomez-Perez. Methodologies, tools and languages for building ontologies : where is their meeting point ? *Data Knowl. Eng.*, 46 :41–64, July 2003.

[CFLGPLC05] Oscar Corcho, M. Fernández-López, Asunción Gómez-Pérez, and A. López-Cima. Building legal ontologies with methontology and webode. *Law and the Semantic Web*, 3369 :142–157, 2005.

[Cha97] M. Charolles. L'encadrement du discours : Univers, champs, domaines et espaces. Cahier de recherche linguistique, Université de Nancy2, 1997.

[Cha02] Jean Charlet. L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Mémoire d'habilitation à diriger des recherches, 12 2002.

[Cha10] J. Charre. Échelle en géographie : de la carte à l'espace. In *Colloque Géopoint 2010*, Avignon, France, 3 et 4 juin 2010.

[CHGM06] Claude Chrisment, Nathalie Hernandez, Françoise Genova, and Josiane Mothe. D un thesaurus vers une ontologie de domaine pour l exploration d un corpus. *AMETIST*, 0 :59–92, septembre 2006.

[CJB99] B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What are ontologies, and why do we need them ? *IEEE Intelligent Systems*, 14 :20–26, 1999.

[CM92] M. Chein and M. Mugnier. Conceptual Graphs : Fundamental Notions. *Revue d'intelligence artificielle*, 6(4) :365–406, 1992.

[COL09] G. COLLETIS. Local development, proximities et productive encounters : The case of development dynamics in the region of toulouse. *Canadian Journal of Regional Science*, 32, 2009.

[CPSV03] N. Cullot, C. Parent, S. Spaccapietra, and C. Vangenot. Des sig aux ontologies géographiques. *Revue Internationale de Géomatique*, 13(3/2003) :285–306, 2003.

[CV05] Philipp Cimiano and Johanna Völker. Text2onto - a framework for ontology learning and data-driven change discovery. In Andres Montoyo, Rafael Munoz, and Elisabeth Metais, editors, *Proceedings of the 10th*

International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, June 2005. Springer.

[DAP06] L. Dufy, V. Abt, and P. Poyet. Giea : Gestion des informations de l'exploitation agricole un projet au service de l'interopérabilité sémantique de la profession agricole. *Ingénieries E A T*, 48 :27–36, 2006.

[Den97] Michel Denis. The Description of Routes : A Cognitive Approach to the Production of Spatial Discourse. *Cahiers Psychologie Cognitive*, 16(4) :409–458, 1997.

[Dia06] Gayo Diallo. *Une Architecture à base d'Ontologies pour la Gestion Unifiées des Données Structurées et non Structurées*. PhD thesis, Université Joseph-Fourier - Grenoble I, 12 2006.

[DMB05] G. Di Meo and P. Buleon. *L'espace social : Lecture géographique des sociétés*. Armand Colin, Paris, 2005.

[DMCG99] J. Domingue, E. Motta, and O. Corcho Garcia. *Knowledge Modelling in WebOnto and OCML : A User Guide*, 1999.

[DMGVOM07] Alina Dia Miron, Jérôme Gensel, Marlène Villanova-Oliver, and Hervé Martin. Towards the geo-spatial querying of the semantic web with ontoast. In *7th international conference on Web and wireless geographical information systems*, W2GIS'07, pages 121–136, Berlin, Heidelberg, 2007. Springer-Verlag.

[Dom98] J. Domingue. Tadzebao and webonto : Discussing, browsing, editing ontologies on the web. In *11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1998.

[DS05] J. Denègre and F. Salgé. *Les systèmes d'informations géographique, 2e éd.* P.U.F. « Que sais-je ? », Paris, 2005.

[Dub04] Karl Dubost. Ontologie, thésaurus, taxonomie et web sémantique, 2004.

[EFF08] Bernard Espinasse, Sébastien Fournier, and Fred Freitas. Agent and ontology based information gathering on restricted web domains with agathe. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC'08, pages 2381–2386, New York, NY, USA, 2008. ACM.

[Ehr08] M. Ehrmann. *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. PhD thesis, Thèse de doctorat, Université Paris 7 Denis Diderot, 2008.

[EJ06] M. Ehrmann and G. Jacquet. Vers une double annotation des entités nommées. In *Traitement Automatique du Langage*, volume 47, pages 63–88, 2006.

[Eli02] B. Elissalde. Une géographie des territoires. *L'information Géographique*, 65(3) :193–205, 2002.

[EMC06] Patrick Etcheverry, Christophe Marquesuzaà, and Sandrine Corbineau. Designing suited interactions for a document management system handling localized documents. In *Proceedings of the 24th annual ACM international conference on Design of communication*, SIGDOC '06, pages 188–195, New York, NY, USA, 2006. ACM.

[Ent96] J. N. Entrikin. Place and region 2. *Progress in Human Geography*, 20(2) :215–221, 1996.

[FFP+95] Adam Farquhar, Richard Fikes, Wanda Pratt, A Pratt, and James Rice. Collaborative ontology construction for information integration. Technical report, 1995.

[FFR97] Adam Farquhar, Richard Fikes, and James Rice. The ontolingua server : a tool for collaborative ontology construction. *Int. J. Hum.-Comput. Stud.*, 46(6) :707–727, 1997.

[FGM+05] L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. Tides 2005 standard for the annotation of temporal expressions. Technical report, MITRE, 2005.

[Fis98] D.H. Fischer. From thesauri towards ontologies ? In *5th Int. ISKO Conference*, 25-29th August, 1998.

[FJA05] F. FU, C.B. Jones, and A.I Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In *IASTED International Conference on Databases and Applications*, 2005.

[FK85] Richard Fikes and Tom Kehler. The role of frame-based representation in reasoning. *Commun. ACM*, 28 :904–920, September 1985.

[FL99] M. Fernández-López. Overview of methodologies for building ontologies. In *Workshop on Ontologies and Problem-Solving Methods (KRR5), IJCAI-99*, pages 4–1 à 4–13, Stockholm, Sweden, August 2, 1999.

[FLGP02] Mariano Fernández-López and Asunción Gómez-Pérez. Overview and analysis of methodologies for building ontologies. *Knowl. Eng. Rev.*, 17 :129–156, June 2002.

[FLGPJ97] Mariano Fernandez-Lopez, Asuncion Gomez-Perez, and Natalia Juristo. Methontology : from ontological art towards ontological engineering. In *Proceedings of the AAAI97 Spring Symposium*, pages 33–40, Stanford, USA, March 1997.

[Fre94] Gottlob Frege. *Begriffsschrift, English translation in J. van Heijenoort.* Harvard University Press, Cambridge, MA, ed. (1967) from frege to gödel edition, 19794.

[Gai01] Mauro Gaio. Traitements de l'information géographique : Représentations et structures. In *Mémoire d'HDR, Université de Caen*, 2001.

[Gai02] R. Gaizauskas. An information extraction perspective on text mining : Tasks, technologies and prototype applications. In *Euromap Text Mining Seminar*, Sheffield, 2002.

[Gal01] A. Galton. Space, time, and the representation of geographical reality. In *Topoi*, volume 20, pages 173–187, December 2001.

[Gan02] Fabien Gandon. Ontology engineering : A survey and a return on experience, 2002.

[Gen00] D. Genest. *Extension du formalisme des graphes conceptuels pour la recherche d'information.* PhD thesis, Montpellier, 2000.

[GF92] M. R. Genesereth and R. E. Fikes. Knowledge Interchange Format, Version 3.0 Reference Manual. Technical Report Logic-92-1, Stanford University, Stanford, CA, USA, 1992.

[GF94] Michael Gruninger and Mark S. Fox. The role of competency questions in enterprise engineering. In *IFIP WG5.7 Workshop on benchmarking - Theory and Practice*, 1994.

[GF95] M. Grüninger and M. Fox. Methodology for the Design and Evaluation of Ontologies. In *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13, 1995*, 1995.

[GGPPSF07] A. Gangemi, A. Gomez-Perez, V. Presutti, and Suarez-Figueroa. Towards a catalog of owl-based ontology design patterns. In *12 Conference of the Spanish Association for Artificial Intelligence*, 12-16 November 2007.

[GHM+07] Tudor Groza, Siegfried Handschuh, Knud Moeller, Gunnar Grimnes, Leo Sauermann, Enrico Minack, Cedric Mesnage, Mehdi Jazayeri, Gerald Reif, and Rosa Gudjonsdottir. The nepomuk project - on the way to the social semantic desktop. In Tassilo Pellegrini and Sebastian Schaffert, editors, *Proceedings of I-Semantics' 07*, pages pp. 201–211. JUCS, 2007.

[Gin91] Matthew L. Ginsberg. Knowledge interchange format : the kif of death. *AI Mag.*, 12 :57–63, September 1991.

[GKN09] Miha Grčar, Eva Klien, and Blaž Novak. Using term-matching algorithms for the annotation of geo-services. In Bettina Berendt, Dunja Mladenič, Marco Gemmis, Giovanni Semeraro, Myra Spiliopoulou, Gerd Stumme, Vojtěch Svátek, and Filip Železný, editors, *Knowledge Discovery Enhanced with Semantic and Social Information*, volume 220, chapter 8, pages 127–143. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[GLR06] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, April 2006.

[GP99] Asunción Gómez-Pérez. Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Banff Knowledge Acquisition for Knowledge-Based Systems, KAW'99*, volume 2, pages 6.1.1–6.1.18, Banff,

Alberta, Canada", 16-21 October 1999. University of Calgary, Alberta, Canada.

[GPCFL04] Asuncion Gomez-Perez, Oscar Corcho, and Mariano Fernandez-Lopez. *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing)*. Springer, July 2004.

[GPGPB08] Asunción Gómez-Pérez, José Ángel Ramos Gargantilla, Antonio F. Rodríguez Pascual, and Luis Manuel Vilches Blázquez. The ign-e case : Integrating through a hidden ontology. In *SDH*, pages 417–435, 2008.

[GPMM03] A. Gomez-Perez and D. Manzano-Macho. A survey of ontology learning methods and techniques. Deliverable 1.5, OntoWeb Consortium, 2003.

[Gre94] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[Gru93] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition Academic Press Inc. 5(2)*, 5(2), 1993.

[Gry08] E. Gryglicka. Un système d'annotation des entités nommées du type personne pour la résolution de la référence. In *RECITAL 2008, Avignon (France)*, 9-13 juin 2008.

[GSE+08] Mauro Gaio, Christian Sallaberry, Patrick Etcheverry, Christophe Marquesuzaà, and Julien Lesbegueries. A global process to access documents' contents from a geographical point of view. *J. Vis. Lang. Comput.*, 19(1) :3–23, 2008.

[Gua97] Nicola Guarino. Some organizing principles for a unified top-level ontology. In *AAAI 1997 SPRING SYMPOSIUM ON ONTOLOGICAL ENGINEERING (LADSEB-CNR INT. REP. 02/97, V3.0)*. AAAI Press, 1997.

[Gua98] Nicola Guarino. Formal ontology and information systems. In *FOIS'1998*, pages 3–15, Amsterdam, 1998. IOS Press.

[Gui07] Eric Guichard. L'internet et le territoire. *Études de communication*, 30, 2007.

[Gui08] E. Guichard. Internet, cartes, territoire et culture. *Communication et Langages*, 158 :77–92, 2008.

[GV04] P. George and F. Verger. *Dictionnaire de la géographie - 8ème édition PUF - Quadrige 2004*. Saint-Just-la Pendue, Belin, 2004.

[Her05] Nathalie Hernandez. *Ontologies de domaine pour la modélisation du contexte en Recherche d'information*. PhD thesis, Université Paul Sabatier - Toulouse III, 12 2005.

[HIC07] Guillermo Nudelman Hess, Cirano Iochpe, and Silvana Castano. Towards a geographic ontology reference model for matching purposes. In Lúbia Vinhas and Antônio Carlos da Rocha Costa, editors, *GeoInfo*, pages 35–47. INPE, 2007.

[Hil06] Linda L. Hill. *Georeferencing : The Geographic Associations of Information*. MIT Press, Cambridge, MA, paperback edition, 2006.

[Hin90] Donald Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, ACL '90, pages 268–275, Stroudsburg, PA, USA, 1990. Association for Computational Linguistics.

[HMW+93] R. de Hoog, R. Martil, B. Wielinga, R. Taylor, C. Bright, and W. Velde. The Common KADS model set. Technical report, University of Amsterdam, Lloyd's Register, December 1993.

[IB09] Antoine Isaac and Thierry Bouchet. Rameau et skos. *Dissemination paper in Arabesques*, 54 :13–14, April-June 2009.

[Ins07] Inspire. Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire). Journal officiel de l'union européenne, directive inspire, 2007.

[ISI+00] A. Inaba, T. Supnithi, M. Ikeda, R. Mizoguchi, and J. Toyoda. An overview of learning goal ontology. In *ECAI2000 Workshop on Analysis and Modelling of CollaborativeLearning Interactions*, pages 23–30, 2000.

[JKT97] Christian Jacquemin, Judith Klavans, and Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *ACL 97*, pages 24–31, 1997.

[Jol04] T. Joliveau. Géomatique et gestion environnementale du territoire. recherche sur un usagegéographique des sig, 2004.

[JPR+02] C.B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies : An overview of the SPIRIT project. In *25th ACM SIGIR*, 2002.

[KAG09] Mouna Kamel and Nathalie Aussenac-Gilles. Construction automatique d ?ontologies à partir de spécifications de bases de données (regular paper). In Fabien Gandon, editor, *Journées Francophones d'Ingénierie des Connaissances (IC), Hammamet (Tunisie), 25/05/2009-29/05/2009*, pages 85–96, http ://www.pug.fr, mai 2009. Presses Universitaires de Grenoble.

[KBG08] Eric Kergosien, Marie Noelle Bessagnet, and Mauro GAIO. Semantic cartography : towards helping experts in their indexation task. In *EKAW'2008, 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns*, page Poster, Acitrezza, Catania, Italy, 29th September-3rd October 2008.

[KBGre] Eric Kergosien, Marie-Noelle Bessagnet, and Mauro Gaio. Exploitation d'une cartographie sémantique à des fins de validation : application à

l'indexation experte de corpus documentaires. *Revue Documentation et bibliothèques*, 57(1), 2011, (à paraitre).

[Ker06] Eric Kergosien. Mise en place d'un modèle sémantique pour l'indexation et la recherche dans la base d'images de la médiathèque. masters, Université de Pau des Pays de l'Adour, June 2006. Rapport de stage de Master 2 Recherche, Indexation et Structuration de la Connaissance.

[Ker08] Eric Kergosien. Des documents vers la connaissance d'un territoire. In *Des documents vers la connaissance d'un territoire Majecstic' 08*, page XX, Marseille France, 10 2008. 8 pages, conférences jeune chercheur.

[Ker10] E. Kergosien. L'invité du mois : Représentation sémantique d'un territoire à partir de rameau. In *BnF - Lettre d'information - Actualités du catalogue : produits et services bibliographiques*, Paris, FRANCE, Février 2010.

[KKB+re] Mouna Kamel, Eric Kergosien, Marie-Noelle Bessagnet, Christian Sallaberry, Nathalie Aussenac-Gilles, Mauro GAIO, Marion Laignelet, and Van Tien Nguyen. Exploitation de différents types de corpus pour la construction d'ontologie : application à la géographie. *Revue des Techniques et Science Informatiques, TSI*, 2011, à paraître.

[KKS+09] Eric Kergosien, Mouna Kamel, Christian Sallaberry, Marie-Noëlle Bessagnet, Nathalie Aussenac Gilles, and Mauro Gaio. Construction et enrichissement automatique d'ontologie à partir de ressources externes. In *JFO'09 JFO'09 : 3es Journées Francophones sur les Ontologies*, pages 1–10, Poitiers France, 12 2009.

[KLW95] Michael Kifer, Georg Lausen, and James Wu. Logical foundations of object-oriented and frame-based languages. *J. ACM*, 42 :741–843, July 1995.

[Kok06] M. Kokla. Guidelines on geographic ontology integration. In *ISPRS Technical Commission II Symposium*, Vienna, Austria, 12-14 July, 2006.

[KPH05] Aditya Kalyanpur, Bijan Parsia, and James A. Hendler. A tool for working with web ontologies. *Int. J. Semantic Web Inf. Syst.*, 1(1) :36–49, 2005.

[Lau91] D. Laur. *Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple*. PhD thesis, Université de Toulouse II, 1991.

[Lav07] Benoit Lavoie. *Notion d'ontologie et construction d'ontologie à partir de textes*. PhD thesis, Université du Qhébec à Montréal, 2007. Programme de Doctorat en Informatique Cognitive.

[LDG90] B. Lenat Douglas and R. V. Guha. *Building Large Knowledge-Based Systems : Representation and Inference in the (CYC) Project*. Addison-Wesley, Reading, Massachusetts, 1990.

[Lei04] Jochen L. Leidner. Toponym resolution in text (abstract only) : "which sheffield is it ?". In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 602–602, New York, NY, USA, 2004. ACM.

[Les07] Julien Lesbegueries. *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé*. PhD thesis, Université de Pau et des Pays de l'Adour, 11 2007.

[LF04] Larson and Frontiera. Ranking and representation for geographic information retrieval. *Workshop on Geographic Information Retrieval - SIGIR*, 2004. http://www.geo.unizh.ch/~rsp/gir/abstracts/larson.pdf.

[LGN07] P Loustau, Mauro Gaio, and T. Nodenot. Des déplacements à l'itinéraire, du syntagme au discours. In *SAGEO*, pages 1–15, 2007.

[LGPSS99] M. F. Lopez, A. Gomez-Perez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using methontology and the ontologydesign environment. *IEEE Intelligent Systems and their Application*, 14(1) :37–46, 1999.

[LKIR94] C. A. Lindley, V. R. Kumar, R. Irrgang, and J. R. Robertsonn. An evaluation of information retrieval methods and semantic network processing for automatic link generation in hypermedia systems. In *Second International Interactive Multimedia Symposium*, Perth, Western Australia, 23-28 January 1994.

[LL03] Jacques Levy and Michel (dir.) Lussault. *Dictionnaire de la géographie et de l'espace des sociétés*. Éd. Hachette, Saint-Just-la Pendue, Belin, 2003.

[LM01] O. Lassila and Deborah L. McGuinness. The role of frame-based representation on the semantic web. Technical Report KSL-01-02, Stanford University, Stanford, 2001.

[LMP01] S. Lardon, P. Maurel, and V. Piveteau. *Représentations spatiales et développement territorial*. Hermès sciences publications, Paris, France, 2001.

[LMR93] R. Laurini and F. Milleret-Raffort. *Les bases de données en géomatique*. Paris, ed hermès edition, 1993.

[LNG08] Pierre Loustau, Thierry Nodenot, and Mauro Gaio. Spatial decision support in the pedagogical area : from travel stories to geocoded itineraries. In Europia, editor, *ICHSL2008, 6th International Conference on Human System Learning, Toulouse (France)*, page 8 pages (format IEEE), mai 2008.

[Lou08] P. Loustau. *Interprétation automatique d'itinéraires dans des récits de voyages*. PhD thesis, Université de Pau et des Pays de l'Adour, 2008.

[LPLGS07] Annig Le Parc Lacayrelle, Mauro Gaio, and Christian Sallaberry. La composante temps dans l'information géographique textuelle. *Revue Document Numérique*, 10(2) :129–148, 2007.

[LSG06] Julien Lesbegueries, Christian Sallaberry, and Mauro Gaio. Associating spatial patterns to text-units for summarizing geographic information. In ACM, editor, *Proceedings of ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop*, pages 40–43, Seattle États-Unis, 08 2006.

[Lyn60] K. Lynch. Cognitive maps in rats and menthe image of the city. *MIT Press*, 1960.

[Mah96] Kavi Mahesh. Ontology development for machine translation : Ideology and methodology. Technical Report MCCS-96-292, CRL, New Mexico State University, 1996.

[Mal03] N. Malandain. *La relation Texte/Image, Essai de modélisation dans un corpus géographique.* PhD thesis, Université de Caen, 2003.

[MAM10] Ammar MECHOUCHE, Nathalie ABADIE, and Sébastien MUSTIÈRE. Alignment-based measure of the distance between potentially common parts of lightweight ontologies. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, and Isabel Cruz, editors, *International Workshop on Ontology Matching, 9th International Semantic Web Conference (ISWC*

10), http ://om2010.ontologymatching.org/, 25 november 2010.

[Mar82] M. Marie. Un territoire sans nom. page 46, Paris, France, 1982.

[MBD00] Enrico Motta, Simon Shum Buckingham, and John Domingue. Ontology-driven document enrichment : principles, tools and applications. *Int. J. Hum.-Comput. Stud.*, 52 :1071–1109, June 2000.

[Mäd02] Alexander Mädche. *Ontology learning for the semantic Web.* Kluwer international series in engineering and computer science ; 665. Kluwer, Boston [u.a.], 2002.

[MDH05] Hayley Mizen, Catherine Dolbear, and Glen Hart. Ontology ontogeny : Understanding how an ontology is created and developed. In M. Andrea Rodríguez, Isabel F. Cruz, Max J. Egenhofer, and Sergei Levashkin, editors, *GeoS*, volume 3799 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2005.

[MESB08] Christophe Marquesuzaà, Patrick Etcheverry, Christian Sallaberry, and Mustapha Baziz. Accessing Heritage Documents according to Space Criteria within Digital Libraries. *Journal of Digital Information Management*, 6(1) :102–117, 02 2008.

[Min74] Marvin Minsky. A framework for representing knowledge. Technical report, Cambridge, MA, USA, 1974.

[Min08] A.-L. Minard. Recherche et analyse de ressources terminologiques liées à la topographie. masters, Université Lille 3, 2003 2008. Rapport de stage de Master 1, Traitement automatique des langues.

[MIS97] R. Mizoguchi, M. Ikeda, and K. Sinitsa. Roles of shared ontology. In *Roles of Shared Ontology in AIED Research –Intelligence, Conceptualization, Standardization, and Reusability, AIED-97*, pages 537–544, 1997.

[Miz98] R. Mizoguchi. A step towards ontological engineerin. In *12th National Conference on AI of JSAI*, pages p. 24–31, 1998.

[MKSK00] Riichiro Mizoguchi, Kouji Kozaki, Toshinobu Sano, and Yoshinobu Kitamura. Construction and deployment of a plant ontology. In Rose Dieng and Olivier Corby, editors, *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, volume 1937 of *Lecture Notes in Computer Science*, pages 61–75. Springer Berlin / Heidelberg, 2000. 10.1007/3-540-39967-4_9.

[Mén93] Dominique Ménillet. Thésaurus et indexation. *BBF*, (5) :44–46, 1993.

[MN95] Kavi Mahesh and Sergei Nirenburg. A situated ontology for practical nlp. In *In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, 1995.

[MS01] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16 :72–79, March 2001.

[MST03] Claude Motte, Isabelle Séguy, and Christine Théré. *Communes d'hier, communes d'aujourd'hui : les communes de la France métropolitaine, 1801-2001, dictionnaire d'histoire administrative.* (Classiques de l'économie et de la population, études et enquêtes historiques), Paris, Institut national d'études démographiques, 2003.

[MT04] Philippe Muller and Xavier Tannier. Une méthode pour l'annotation de relations temporelles dans des textes et son évaluation. In *Actes de la 11ème Conférence annuelle de Traitement Automatique des Langues Naturelles*, pages 319–328, Fès, Maroc, April 2004.

[MUC98] MUC-7. Defense advanced research projects agency. In *Seventh Message Understanding Conferences (MUC-7)*, 1998.

[MV07] Claude Motte and Marie-Christine Vouloir. Le site cassini.ehess.fr : un instrument d'observation pour une analyse du peuplement. *Bulletin du Comité français de cartographie*, (191) :68–84, mars 2007.

[MW00] I. Mani and G. Wilson. Temporal granularity and temporal tagging of text. In *AAAI-2000 Workshop on Spatial and Temporal Granularity*, AAAI-2000, Austin, 2000.

[NFF+91] Robert Neches, Richard Fikes, Timothy W. Finin, Thomas R. Gruber, Ramesh S. Patil, Ted E. Senator, and William R. Swartout. *AI Magazine*, pages 36–56, 1991.

[NFM00] N.F. Noy, R.W. Fergerson, and M.A. Musen. The knowledge model of protégé-2000 : Combining interoperability and flexibility. *Lecture Notes in Computer Science*, 1937 :69–82, 2000.

[NH04] K. Neshatian and M. R. Hejazi. Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies. In *2nd Workshop on Information Technology and its Disciplines*, Workshop ITD'04, pages 43–48, 2004.

[Nie03] Ewa Zofia Nieszkowska. Quelle indexation pour une bibliothèque spécialisée ? le cas de la bibliothèque de l'institut français d'architecture. masters, 2003. Mémoire d'étude, diplôme de conservateur des bibliothèques.

[NN06] Claire Nedellec and Adeline Nazarenko. Ontologies and information extraction. *CoRR*, abs/cs/0609137, 2006.

[Nor93] Donald A. Norman. *Things That Make Us Smart : Defending Human Attributes in the Age of the Machine*. Addison Wesley Publishing Company, 1993.

[NSD+01] Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, 16 :60–71, 2001.

[NV06] Roberto Navigli and Paola Velardi. Enriching a formal ontology with a thesaurus : an application in the cultural heritage domain. In *Proceedings of* 2^{nd} *Workshop on Ontology Learning and Population (OLP), in the* 44^{th} *Annual Meeting of the Association for Computational Linguistics joint with the* 21^{st} *International Conference on Computational Linguistics (COLING-ACL 2006)*, pages 1–9, Sydney, Australia, 2006.

[ODWA86] Robert V ONeill, D L DeAngelis, J B Waide, and T F H Allen. *A Hierarchical Concept of Ecosystems*, volume 23. Princeton University Press, 1986.

[Ouz06] Mourad Ouziri. Accessing data in the semantic web : An intelligent data integration and navigation approaches. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Artificial Intelligence Applications and Innovations*, volume 204 of *IFIP International Federation for Information Processing*, pages 119–128. Springer Boston, 2006. 10.1007/0-387-34224-9_14.

[Par70] T. Parsons. An analysis of mass terms and amount terms. *Foundations of languages*, 6 :363–388, 1970.

[Pas04] Marius Pasca. Acquisition of categorized named entities for web search. In *Proceedings of the thirteenth ACM international conference on In-*

formation and knowledge management, CIKM '04, pages 137–145, New York, NY, USA, 2004. ACM.

[PH02] Jack Park and Sam Hunting, editors. *XML Topic Maps : Creating and Using Topic Maps for the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[Pio92] Xavier Piolle. Proximité géographique et lien social, de nouvelles formes de territorialités ? *L'espace géographique*, 4 :349–358, 1992.

[PM04] Helena Sofia Pinto and Jo&#x00e3 ;o P. Martins. Ontologies : How can they be built ? *Knowl. Inf. Syst.*, 6 :441–464, July 2004.

[Poi03] Thierry Poibeau. *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier, 2003. ISBN 2-7462-0610-2.

[PS09] Karen Pinel-Sauvagnat. Propagation-based structured text retrieval. In O.M. Tamer and L. Ling, editors, *Encyclopedia of Database Systems*, pages 2197–2201. Springer, http ://www.springerlink.com, mai 2009. Sur invitation.

[PSA07] M. Perry, A. Sheth, and I.B. Arpinar. *Geospatial and Temporal Semantic Analytics*, pages 1–14. Encyclopedia of Geoinformatics, Hassan A. Karimi (Ed), Idea-Group Inc., à paraître en 2007.

[Qui68] M. R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.

[RLB+04] Catherine Roussey, Robert Laurini, Caroline Beaulieu, Yohann Tardy, and Monique Zimmerman. Le projet Towntology : un retour d'expérience pour la construction d'une ontologie urbaine. *Revue Internationale de Géomatique*, 14(2) :217–237, February 2004.

[RR06] François Role and Guillaume Rousse. Construction incrémentale d'une ontologie par analyse du texte et de la structure des documents. *Document numérique*, 9(1) :77–92, 2006.

[SAG99] Patrick Séguéla and Nathalie Aussenac-Gilles. Extraction de relations sémantiques entre termes et enrichissement de modèles conceptuel . In *Ingénierie des Connaissances - IC'99 , Palaiseau (F), 14/06/99-18/06/99*, pages 10–20, Paris (F), juin 1999. Ecole Polytechnique.

[Sag06] B. Sagot. *Analyse automatique du français : lexiques, formalismes, analyseurs, Thèse de doctorat en informatique*. PhD thesis, Université Paris VII, 2006.

[Sal09] Maryse Salles. Ontologies pour l'aide à la décision publique et prise en compte des doxas. In *Actes IC2009 IC2009*, page 109, HAMMAMET Tunisia, 03 2009. Financement Région Midi-Pyrénées.

[SAS02] York Sure, Jürgen Angele, and Steffen Staab. Ontoedit : Guiding ontology development by methodology and inferencing. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002*

Confederated International Conferences DOA, CoopIS and ODBASE 2002, pages 1205–1222, London, UK, UK, 2002. Springer-Verlag.

[Sav05] Jacques Savoy. Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française. In *CORIA*, pages 9–24, 2005.

[SBAG02] Sylvie Szulman, Brigitte Biébow, and Nathalie Aussenac-Gilles. Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE . *TAL*, 43(1) :103–128, 2002.

[SBF98] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering : Principles and methods. *Data & Knowledge Engineering*, 25(1-2) :161 – 197, 1998.

[SBLG07] Christian Sallaberry, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio. Une approche d'extraction et de recherche d'information spatiale dans les documents textuels - évaluation. In *CORIA*, pages 53–64. Université de Saint-Étienne, 2007.

[Sch94] J. Scheibling. *Qu'est-ce que la géographie ?* Éd. Hachette, Paris, France, 1994.

[SCM03] Ulrike Sattler, Diego Calvanese, and Ralf Molitor. The description logic handbook. chapter Relationships with other formalisms, pages 137–177. Cambridge University Press, New York, NY, USA, 2003.

[SE05] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*, Lecture Notes in Computer Science, chapter 5, pages 146–171. 2005.

[SGD04] Lina Fatima Soualmia, Christine Golbreich, and Stéfan Jacques Darmoni. Representing the mesh in owl : Towards a semi-automatic migration. In Udo Hahn, editor, *KR-MED*, volume 102 of *CEUR Workshop Proceedings*, pages 81–87. CEUR-WS.org, 2004.

[SGLL07] C. Sallaberry, M. Gaio, J. Lesbegueries, and P. Loustau. *A Semantic Approach for Geospatial Information Extraction from Unstructured Documents*, page 93. Scharl, A. and Tochtermann, K., 2007.

[SH01] Frank Schilder and Christopher Habel. From temporal expressions to temporal information : semantic tagging of news messages. In *Proceedings of the workshop on Temporal and spatial information processing - Volume 13*, TASIP '01, pages 9 :1–9 :8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[SLL+04] Dagobert Soergel, Boris Lauser, Anita C. Liang, Frehiwot Fisseha, Johannes Keizer, and Stephen Katz. Reengineering thesauri for new applications : The agrovoc example. *J. Digit. Inf.*, 4(4), 2004.

[SM00] S. Staab and A. Mädche. Axioms are Objects, too - Ontology Engineering beyond the Modeling of Concepts and Relations. In V.R. Benjamins,

A. Gomez-Perez, and N. Guarino, editors, *Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI 2000, Berlin, Germany*, August 21 – 22, 2000.

[SM01] B. Smith and D. Mark. Geographical categories : An ontological investigation. *International Journal of Geographical Information Science*, 15(7) :591–612, 2001.

[SM03] David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, pages 45–49, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Sow84] J. F. Sowa. *Conceptual structures : information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.

[Sow95] John F. Sowa. Top-level ontological categories. *Int. J. Hum.-Comput. Stud.*, 43 :669–685, December 1995.

[SR09] Brigitte Safar and Chantal Reynaud. Alignement d'ontologies basé sur des ressources complémentaires illustration sur le système taxomap. *Technique et Science Informatiques*, 28(10) :1211–1232, 2009.

[SRKR97] B. Swartout, P. Ramesh, K. Knight, and T. Russ. Toward Distributed Use of Large-Scale Ontologies. *AAAI Symposium on Ontological Engineering*, 1997.

[SSSS01] Steffen Staab, H.-P. Schnurr, Rudi Studer, and York Sure. Knowledge processes and ontologies. *IEEE Intelligent Systems*, 16(1), 2001.

[Sto02] Dejan Stosic. *"par" et "à travers" dans l'expression des relations spatiales : comparaison entre le français et le serbo-croate*. PhD thesis, Université Toulouse le Mirail - Toulouse II, 12 2002.

[Sve86] Elaine Svenonius. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5) :331–340, September 1986.

[SVV01] Christoph Schlieder, Thomas Vögele, and Ubbo Visser. Qualitative spatial representation for information retrieval by gazetteers. In *COSIT*, pages 336–351, 2001.

[TAAM+10] Michel Tchotsoua, Moussa Aboubakar, Guy-Florent Ankogui Mpoko, Alfred Bertin Bangara, Eric Fotsing, Boniface Ganota, Agard Koyoumtan, Arabi Mouhaman, Bedjaoué Moupeng, and Jérôme Picard. Contribution de la géomatique à la gestion des territoires villageois des savanes d'Afrique centrale. In P. BOUMARD L. SEINY-BOUKAR, editor, *Actes du colloque « Savanes africaines en développement : innover pour durer » Savanes africaines en développement : innover pour durer*, page 9 p., Garoua Cameroon, 2010. Cirad.

[Tal00] L. Talmy. *Towards a cognitive semantics*. MIT Press, Cambridge (MA), 2000.

[Tex05] R. Texier. Taxinomies, thésaurus et ontologies. *EliKya, intelligence des organisations*, pages p.1–3, 2005.

[Tol48] E.C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55 :189–208, 1948.

[Top01] TopicMaps.org. Xml topic maps (xtm) 1.0. Technical report, TopicMaps.org, 2001.

[Tro04] Raphaël Troncy. *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels*. PhD thesis, Université Joseph-Fourier - Grenoble I, mars 2004. Prof. Yves CHIARAMELLA, Université Joseph Fourier (Président)
M. Jean CHARLET, Assistance Publique-Hôpitaux de Paris (Rapporteur)
Prof. Jacques LE MAITRE, Université du Sud Toulon Var (Rapporteur)
Prof. Asuncion GOMEZ-PEREZ, Université Polytechnique de Madrid
(Examinateur)
M. Yannick PRIE, Université Lyon 1 (Examinateur)
M. Bruno BACHIMONT, Université Technologique de Compiègne - Institut National de l'Audiovisuel (Directeur de Thèse)
M. Jérôme EUZENAT, INRIA Rhône-Alpes (Directeur de Thèse).

[Uit01] Henricus Theodorus Johannes Antonius Uitermark. *Ontology-Based Geographic Data Set Integration*. PhD thesis, Enschede, September 2001.

[UK95] Mike Uschold and Martin King. Towards a methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing*, 80(July) :275–280, 1995.

[Usc96] Mike Uschold. Building ontologies : Towards a unified methodology. In *In 16th Annual Conf. of the British Computer Society Specialist Group on Expert Systems*, pages 16–18, 1996.

[UTC04] E.L. Usery, G. Timson, and M. Coletti. Multidimensional representation of geographic features. In *International Journal of Geographic Information Science, in review*, pages 1–8, 2004. http://carto-research.er.usgs.gov/multi-dimension/pdf/usery.996.pdf.

[Van86] Claude Vandeloise. *L'espace en français*. aux Editions du Seuil, Paris, 1986.

[Van87] C. Vandeloise. La préposition á et le principe d'anticipation. *Langue Française*, 76 :77–110, 1987.

[VFM01] Paola Velardi, Paolo Fabriani, and Michele Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pages 270–284, 2001.

[VHSW97] G. Van Heijst, A. Th. Schreiber, and B. J. Wielinga. Using explicit ontologies in kbs development. *Int. J. Hum.-Comput. Stud.*, 46 :183–292, March 1997.

[VL01] J. Virbel and C. Luc. Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, XXIII(1) :103–123, 2001.

[VOS03] Raphael Volz, Daniel Oberle, and Rudi Studer. Views for light-weight web ontologies. In *Proceedings of the 2003 ACM symposium on Applied computing*, SAC '03, pages 1168–1173, New York, NY, USA, 2003. ACM.

[VW07] Thomas Vander Wal. Folksonomy coinage and definition, 2007.

[WB06] Antoine Widlöcher and Frédérik Bilhaut. La plate-forme LinguaStream. In *Colloque international des étudiants chercheurs en didactique des langues et en linguistique*, Grenoble, France, juillet 2006.

[WB07] Antoine Widlöcher and Frédérik Bilhaut. La plate-forme linguastream. In *Autour des langues et du langage : perspective pluridisciplinaire*, pages 447–454, Grenoble, France, 2007. Presses Universitaires de Grenoble. Publication faisant suite au colloque international des étudiants chercheurs en didactique des langues et en linguistique ayant eu lieu en juillet 2006.

[WC03] J. C. Weis and J. Charlet. Construction d'ontologie à partir de textes : application à un réseau de périnatalité. In *7èmes Journées Ingénierie des Connaissances*, pages 85–100, Laval, France, 1-3 juillet 2003. Dieng-Kuntz R., Presses universitaires de Grenoble.

[WN07] Davy Weissenbacher and Adeline Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Proceedings of Traitement Automatique des Langues Naturelles*, pages 145–155, France, 06 2007. ATALA.

[Woo75] William A. Woods. What's in a Link : Foundations for Semantic Networks. In D.G. Bobrow and A. Collins, editors, *Representation and Understanding*. Academic Press, 1975.

[WP94] Allison Woodruff and Christian Plaunt. GIPSY : Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9) :645–655, 1994.

[WSWS01] Bob J. Wielinga, A. Th. Schreiber, Jan Wielemaker, and Jacobijn Sandberg. From thesaurus to ontology. In *K-CAP*, pages 194–201. ACM, 2001.

[Zar04] H. Zargayouna. Contexte et sémantique pour une indexation de documents semi-structrés. In *Premiere COnference en Recherche d'Information et Applications (CORIA'O4)*, mars 2004.

[Zwe94] P. Zweigenbaum. Menelas : an access system for medical records using natural language. *Comput Methods Programs Biomed*, 45(1-2) :117–20, 1994.

Résumé

Dans les bibliothèques et les médiathèques, une caractéristique importante des fonds documentaires mis à disposition est qu'ils contiennent d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire, et il est primordial pour ces centres de valoriser ces spécificités territoriales pour répondre à des objectifs d'information et d'éducation.

Dans ce contexte, nous faisons l'hypothèse qu'en utilisant un point de vue géographique pour modéliser un ensemble de ressources terminologiques utilisées pour indexer un fond documentaire, il est possible de faire émerger une représentation du territoire qui y est implicitement décrite. Concernant la modélisation de la connaissance en géomatique [106], de nombreux travaux s'appliquent à construire une représentation sémantique structurée géographique de domaines cibles. Cependant, il ne semble pas exister d'approche permettant de construire une représentation d'un territoire à partir de fonds documentaires annotés. Nous proposons donc une méthodologie complète et automatisée permettant de construire une couche conceptuelle de type ontologie d'un territoire, sur la base d'un fonds documentaire indexé par des experts. Nous positionnons nos travaux dans l'extraction et la structuration de la connaissance que nous appliquons dans le domaine de la géomatique en nous appuyant notamment sur des techniques provenant du Traitement Automatique du Langage Naturel.

Nous entendons ici par territoire un ensemble de lieux que l'on peut mettre en relation selon un ensemble de thèmes en fonction d'une période donnée.

Ainsi, nous présentons un complément original s'appuyant sur le travail d'indexation réalisé par les experts documentalistes sur un fonds documentaire pour faire émerger une ontologie d'un territoire implicitement décrit dans les documents. Une contribution importante de notre travail concerne l'enrichissement de façon incrémentale de la représentation d'un territoire. Nous proposons pour cela une chaine de TALN qui permet de marquer dans des documents textes annotés un ensemble d'informations spatiales, temporelles et thématiques qui nous sert de base pour l'enrichissement de la représentation d'un territoire. Une perspective à ces travaux est de pouvoir valider notre approche sur plusieurs fonds documentaires d'origines diverses. L'intérêt sera de proposer une méthode qui, sur la base des représentations de territoires obtenues, permettrait d'identifier et de représenter les spécificités de chaque fonds documentaire.

Mots-clés: construction d'ontologies à partir de ressources structurées, indexation, fonds documentaire, territoire, Traitement Automatique du Langage Naturel, vocabu-

106. Discipline ayant pour objet la gestion des données à référence spatiale par l'intégration au moyen de l'informatique des savoirs et des technologies reliées à leur acquisition, leur stockage, leur traitement et leur diffusion, et principalement : la topométrie, la cartographie, la géodésie, la photogrammétrie et la télédétection (Joliveau, 2004)

laire contrôlé, information géographique

Abstract

Within libraries and media centers, an important feature of documents available in such centers is that they contain abundant references to history, geography, Heritage, in fact the territory. It is essential for these centers to develop these territorial specificities to answer objectives of information and education.

In this context, we make the hypothesis that by using a geographical point of view to model a set of resources terminology used to index a documentary, it is possible to bring out a representation of the territory which is implicitly described. Concerning the modeling of knowledge in geomatics , many research works propose a methodology in order to build a geographic structured semantic representation of target domains. However, it seems that there is no approach proposing to build a representation of a territory from a corpus and descriptives notices. We propose a complete methodology for building an ontology of a territory, on the basis of a document collection indexed by experts. We position our work in extracting and structuring knowledge that we apply to the field of geomatics including techniques from NLP.

We mean by territory a set of places that we can relate on a set of topics based on a period.

Thus, we present an original complement based on the indexing work done by experts librarians on a documentary to bring out an ontology of a territory implicitly described in the documents. An important contribution of our work is to enrich step by step the representation of territory. We propose an automated linguistic processing that allows to mark in text documents a set of spatial, temporal and thematic information which serves as the basis for the representation of a ontology. Perspective to this work is to validate our approach on several corpus of various origins. The interest will be to propose a method, based on representations of territories obtained, would identify and represent the characteristics of each documentary.

Keywords: Ontology Construction from Structured Textual Data, indexing, corpus, territory, Natural Language Processing, controlled vocabulary, geographic information

Résumé

Dans les bibliothèques et les médiathèques, une caractéristique importante des fonds documentaires mis à disposition est qu'ils contiennent d'abondantes références à l'histoire, à la géographie, au patrimoine, en somme au territoire, et il est primordial pour ces centres de valoriser ces spécificités territoriales pour répondre à des objectifs d'information et d'éducation.

Dans ce contexte, nous faisons l'hypothèse qu'en utilisant un point de vue géographique pour modéliser un ensemble de ressources terminologiques utilisées pour indexer un fond documentaire, il est possible de faire émerger une représentation du territoire qui y est implicitement décrite. Concernant la modélisation de la connaissance en géomatique, de nombreux travaux s'appliquent à construire une représentation sémantique structurée géographique de domaines cibles. Cependant, il ne semble pas exister d'approche permettant de construire une représentation d'un territoire à partir de fonds documentaires annotés. Nous proposons donc une méthodologie complète et automatisée permettant de construire une couche conceptuelle de type ontologie d'un territoire, sur la base d'un fonds documentaire indexé par des experts. Nous positionnons nos travaux dans l'extraction et la structuration de la connaissance que nous appliquons dans le domaine de la géomatique en nous appuyant notamment sur des techniques provenant du Traitement Automatique du Langage Naturel.

Nous entendons ici par territoire un ensemble de lieux que l'on peut mettre en relation selon un ensemble de thèmes en fonction d'une période donnée.

Ainsi, nous présentons un complément original s'appuyant sur le travail d'indexation réalisé par les experts documentalistes sur un fonds documentaire pour faire émerger une ontologie d'un territoire implicitement décrit dans les documents. Une contribution importante de notre travail concerne l'enrichissement de façon incrémentale de la représentation d'un territoire. Nous proposons pour cela une chaine de TALN qui permet de marquer dans des documents textes annotés un ensemble d'informations spatiales, temporelles et thématiques qui nous sert de base pour l'enrichissement de la représentation d'un territoire. Une perspective à ces travaux est de pouvoir valider notre approche sur plusieurs fonds documentaires d'origines diverses. L'intérêt sera de proposer une méthode qui, sur la base des représentations de territoires obtenues, permettrait d'identifier et de représenter les spécificités de chaque fonds documentaire.

Mots clés : construction d'ontologies à partir de ressources structurées, indexation, fonds documentaire, territoire, Traitement Automatique du Langage Naturel, vocabulaire contrôlé, information géographique.