HAL
open science

Les ontologies comme support à l'interaction et à la personnalisation dans un processus décisionnel. Exploitation de la sémantique pour favoriser l'automatisation cognitive

Sylvie Ranwez

▶ **To cite this version:**

Sylvie Ranwez. Les ontologies comme support à l'interaction et à la personnalisation dans un processus décisionnel. Exploitation de la sémantique pour favoriser l'automatisation cognitive. Recherche d'information [cs.IR]. Université Montpellier II - Sciences et Techniques du Languedoc, 2013. tel-00851259v1

HAL Id: tel-00851259

https://theses.hal.science/tel-00851259v1

Submitted on 13 Aug 2013 (v1), last revised 16 Sep 2013 (v2)

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.

Université Montpellier II
École doctorale I2S
Information Structures et Systèmes

Habilitation à Diriger des Recherches (H.D.R.)
Spécialité : informatique

Les ontologies comme support à l'interaction et à la personnalisation dans un processus décisionnel

Exploitation de la sémantique pour favoriser l'automatisation cognitive

par

Sylvie Ranwez

soutenue le 11 juillet 2013 à 14h devant le jury composé de :

M. Jean Charlet,	Chargé de mission recherche à l'AP-HP & Inserm,	Président
Mme Nathalie Aussenac-Gilles,	Directeur de recherche, CNRS/IRIT, Université Paul Sabatier, Toulouse	Rapporteur
Mme Pascale Kuntz,	Professeur, École Polytechnique de l'Université de Nantes	Rapporteur
M. Guy Melançon,	Professeur, LaBRI, Université de Bordeaux I	Rapporteur
Mme Marianne Huchard,	Professeur, Université de Montpellier II	Examinateur
M. Gilles Kassel,	Professeur, Université de Picardie Jules Verne, Amiens	Examinateur
M. Jacky Montmain,	Professeur, École des mines d'Alès	Examinateur

A Papa

Préambule

Ce document décrit les principaux résultats des travaux de recherche que j'ai menés au sein du centre de recherche LGI2P de l'école des mines d'Alès, depuis mon recrutement en janvier 2001 jusqu'à ce jour. Mes travaux de thèse (septembre 1997 – décembre 2000), également effectués au LGI2P, ne seront que brièvement mentionnés, puisqu'ils ont fait l'objet d'un mémoire.

L'équipe KID (*Knowledge and Image analysis for Décision making*), dirigée par Jacky Montmain et à laquelle je suis rattachée, axe ses recherches sur le développement de modèles, méthodes et techniques pour aider l'opérateur humain confronté à la maîtrise d'un système complexe, qu'il soit technique, social ou économique, en particulier dans un contexte de prise de décision. Dans ce contexte, la préoccupation majeure des membres de l'équipe concerne l'analyse et l'enrichissement *d'unités de connaissances* (ces unités peuvent représenter des segments de texte, des enregistrements audio-visuels, des pixels, etc.) afin de les exploiter efficacement dans un processus décisionnel. Mes travaux concernent plus particulièrement la **modélisation de connaissance à l'aide d'ontologies de domaine** et leur utilisation comme **support à l'interactivité et à la personnalisation** dans diverses applications, avec en ligne de mire, l'automatisation cognitive, i.e. comment créer un contexte propice à la mise en œuvre de capacités cognitives efficientes. Les résultats obtenus sont le fruit de collaboration entre plusieurs membres de l'équipe mais également avec Vincent Ranwez, de Montpellier SupAgro ou plus récemment avec David Sánchez et Montserrat Batet de l'université Rovira i Virgili de Catalogne (Tarragona, Espagne).

Au sein de cette équipe, j'ai eu l'opportunité d'assumer différentes responsabilités concernant l'enseignement et la recherche et, en particulier, j'ai pu encadrer des stagiaires de master, cinq doctorants, collaborer à différents projets industriels et accompagner des créateurs d'entreprise.

De ces différentes facettes du métier d'enseignant-chercheur, sont nées des réflexions enrichissantes qui, se nourrissant mutuellement, ont abouti à des résultats qui me permettent de présenter ce mémoire en vue d'obtenir une habilitation à diriger les recherches.

En préambule, un curriculum vitae détaille mon parcours, ensuite mes travaux de recherche sont exposés au travers de 5 chapitres, le dernier constituant mon projet de recherche pour les années à venir.

Remerciements

J'adresse ma plus profonde reconnaissance à toutes les personnes qui ont participé de près ou de loin à ces travaux : en premier lieu à celles et ceux qui m'ont fait confiance, qui m'ont écoutée, guidée, conseillée, encouragée, mais également à celles et ceux dont j'ai accompagné les premiers pas en recherche ; merci pour la richesse de nos échanges.

Merci aux rapporteurs pour leurs retours enrichissants sur ce manuscrit et à tous les membres du jury qui ont accepté d'évaluer ce travail.

Merci à Jacky Montmain, responsable de l'équipe KID, pour ses conseils et son amitié, à Yannick Vimont, Directeur du centre de recherche LGI2P, et à toute l'équipe de direction de l'école des mines d'Alès. Je remercie également les membres de l'école doctorale I2S de l'UM II.

Ce mémoire synthétise des résultats obtenus en équipe et je remercie en particulier Michel Crampes, qui a guidé mes premiers pas en recherche, et Stefan Janaqi : leurs points de vue parfois antagonistes forcent à une ouverture d'esprit salutaire. Merci aux thésards que j'ai encadrés Fabien Jalabert, Jean Villerd, Mohameth François Sy, Sébastien Harispe et Nicolas Fiorini. Merci à nos collaborateurs espagnols : Montserrat Batet et David Sánchez, pour les discussions fructueuses de ces derniers mois.

Merci aussi à ceux avec qui j'ai collaboré ponctuellement : Isabelle Rondeau, Benjamin Duthil. Merci aux développeurs de l'ombre sans qui rien n'est possible : Pierre Jean, Ansata Data Balde Sy et Nicolas Clairon. Merci aux stagiaires de Master : Stéphanie Baar (merci de m'avoir ouverte à un domaine qui n'était pas le mien et qui m'a permis d'avoir un œil neuf sur mes travaux), Makhtar MBao, Cécile Wolff, Véronique Gachet et Gilbert Okaro.

Que serait la recherche sans le soutien des documentalistes. Nous avons la chance d'avoir un service de documentation "au top", et cela en grande partie grâce à l'efficacité de son grand manitou, Françoise Armand, que je remercie infiniment pour dégoter l'article introuvable, éveiller notre attention sur les nouvelles découvertes, les derniers outils et être l'agitateur propice à toute production intellectuelle tant auprès des permanents, que des thésards et des étudiants…

Merci à nos partenaires qui se sont impliqués dans la validation de nos travaux, en particulier Armelle Regnault et Patrick Augereau de l'Inserm, et Marie-Thérèse Ménager du CEA.

Des projets industriels ont été à l'origine de mes travaux ou ont bénéficié de leurs résultats. Un merci tout particulier à Torsten Leidig (SAP, Karlsruhe), Christophe Carniel (Nétia), David Carteret (Innova software), Yann Lescurat et Romain Marion (Solunea) et les créateurs pour nos discussions fructueuses : Emmanuel Beaufort, Mory Doukouré, Christophe Dupuis, Nicolas Clairon.

Je remercie les membres de la communauté IC (Ingéniérie des Connaissances) pour les échanges lors des conférences ainsi que tous les chercheurs qui ont participé à la réflexion autour des documents virtuels personnalisables (Serge Garlatti en particulier).

Merci à tous mes collègues de l'enseignement qui, depuis des années, cherchent avec moi les meilleurs moyens de faire aimer l'informatique à un public d'ingénieurs

généralistes. Merci en particulier à Anne-Lise Courbis, Annie Liothin, Gérard Dray et Thomas Lambolais. Merci aussi à l'équipe de formation continue diplômante et en particulier à Sylvie Trouilhet (Université Paul Sabatier) : une efficacité redoutable et toujours avec le sourire ! Merci à nos secrétaires…

Merci à tous ceux qui m'ont fait rire ! ☺

Enfin, merci à ma famille en qui je puise beaucoup de ressources ; merci maman pour tes enseignements d'une grande sagesse. Merci Mathilde et Camille pour votre patience lors de nos longs tête à tête avec une machine.

Merci à toi, Vincent, pour tout ce que tu sais… et parce que j'aurais voulu avoir écrit ces lignes pour toi :

«*T'as construit des lacets*
à travers la montagne
Pour me porter dans les sommets
pour que je gagne…"

Amélie-les-Crayons,
"On n'est pas fatigués", Album "Jusqu'à la mer"

Sommaire

PREAMBULE **5**

REMERCIEMENTS **7**

SOMMAIRE **9**

TABLE DES ILLUSTRATIONS **13**

CURRICULUM VITAE ETENDU **15**

INTRODUCTION **31**

I. CHAPITRE I **35**

CONTEXTE ET POSITIONNEMENT : MODELISER LA CONNAISSANCE D'UN DOMAINE POUR MIEUX EXPLOITER LES RESSOURCES DISPONIBLES **35**

I.1. LES ENJEUX DE LA RECHERCHE D'INFORMATION **36**

I.1.1. POURQUOI RECHERCHER, PARTAGER L'INFORMATION ? 37

I.1.2. QUELLE INFORMATION ? 39

I.1.3. QUI PARTAGE L'INFORMATION ? 40

I.1.4. UNE REQUETE CONCEPTUELLE POUR DES RESULTATS PERTINENTS 41

I.2. MODELISATION DE LA CONNAISSANCE D'UN DOMAINE **42**

I.2.1. DONNEE, INFORMATION ET CONNAISSANCE 42

I.2.2. ONTOLOGIES : DEFINITION(S) 43

I.2.3. INGENIERIE ONTOLOGIQUE : CONCEPTION, FORMALISATION, METHODES ET LANGAGES 44

I.2.3.1. Spécification et évaluation *47*

I.2.3.2. Conception et évolution *48*

I.2.3.3. Diffusion : déploiement et mise en place de l'ontologie *49*

I.2.3.4. Utilisation *49*

I.2.3.5. Gestion et planification *49*

I.2.3.6. Parmi les langages de représentation des ontologies un se détache : OWL *50*

I.2.4. REDUCTION D'ONTOLOGIES AUX LIENS DE SPECIALISATION (ONTOLOGIES *LEGERES*) 51

I.3. SYNTHESE **52**

II. CHAPITRE II **55**

LES ONTOLOGIES COMME SUPPORT A L'INDEXATION CONCEPTUELLE **55**

II.1. QUALITES ATTENDUES D'UNE INDEXATION **56**

II.1.1. EXPRESSIVITE ET FIABILITE 56

II.1.2. TECHNICITE 57

II.1.3. EXPLOITABILITE 57

II.2. ENTITES A INDEXER ET CARACTERISTIQUES A PRENDRE EN COMPTE. **58**

II.2.1. CATALOGAGE OU DESCRIPTION D'EDITION 59

II.2.2. L'INDEXATION DU CONTENU POUR SUPPLEER A UNE RECHERCHE CONTEXTUELLE 59

II.3. L'INDEXATION CONCEPTUELLE AU CŒUR DE NOTRE DEMARCHE 60
II.3.1. ATOUTS DE L'INDEXATION CONCEPTUELLE 60
II.3.2. LIMITES IDENTIFIEES DE L'INDEXATION CONCEPTUELLE 61
II.4. L'INDEXATION PAR PROPAGATION : PRINCIPE ET MISE EN ŒUVRE 63
II.4.1. PRINCIPE DE L'INDEXATION PAR PROPAGATION 64
II.4.2. MISE EN ŒUVRE DE L'INDEXATION PAR PROPAGATION 64
II.4.3. RESULTATS ET LIMITE DE L'APPROCHE 66
II.5. L'INDEXATION CONCEPTUELLE PAR PROPAGATION 67
II.6. SYNTHESE 68
III. CHAPITRE III 71
LA RECHERCHE D'INFORMATION BASEE SUR LES ONTOLOGIES 71
III.1. LES PRINCIPES DE BASES DE LA RECHERCHE D'INFORMATION 72
III.1.1. EXPRESSION DE LA REQUETE 74
III.1.2. ORDONNANCEMENT DES RESULTATS 75
III.1.3. EVALUATION DES SYSTEMES DE RECHERCHE D'INFORMATION 76
III.2. LA RECHERCHE D'INFORMATION CONCEPTUELLE 77
III.2.1. UTILISER LA STRUCTURE SEMANTIQUE DES ONTOLOGIES POUR ETENDRE LA REQUETE 77
III.2.2. UTILISER L'ONTOLOGIE POUR LA REFORMULATION 80
III.3. OU APPROCHES CONCEPTUELLES ET LEXICALES PEUVENT SE COMPLETER 81
III.3.1. ANALYSE LEXICALE POUR COMPLETER L'INDEXATION 81
III.3.2. ANALYSE LEXICALE POUR UNE MEILLEURE COMPREHENSION DES RESULTATS 82
III.4. LES MESURES DE SIMILARITE SEMANTIQUE AU CŒUR DE LA RI 84
III.4.1. MESURES DE SIMILARITE SEMANTIQUE PAR INTENTION 84
III.4.2. MESURES DE SIMILARITE SEMANTIQUE PAR EXTENSION 85
III.4.3. PROPRIETES ET CHOIX DES MESURES DE SIMILARITE SEMANTIQUE 85
III.5. FOCUS SUR DES ALGORITHMES UTILISES DANS DIFFERENTES PHASES DE LA RECHERCHE D'INFIORMATION 86
III.5.1. CALCUL DES LCA D'UN ENSEMBLE DE CONCEPTS 86
III.5.2. EXTRACTION DE SOUS-ONTOLOGIES 87
III.6. SYNTHESE 88
IV. CHAPITRE IV 91
VISUALISATION ET INTERACTIONS EXPLOITANT UN MODELE SEMANTIQUE 91
IV.1. CARTES CONCEPTUELLES ET CARTES SEMANTIQUES 92
IV.1.1. OU IL EST QUESTION DE PROJECTION MDS 93
IV.1.2. DEFINITION DE CARTE SEMANTIQUE, EXEMPLES ET UTILISATION 93
IV.1.3. COMMENT ASSISTER LA CREATION DE CARTES SEMANTIQUES 95

IV.2. ASSISTER VISUELLEMENT LA NAVIGATION DANS DE LARGES BASES DE DOCUMENTS ... 95

IV.2.1. GUIDES ET REPERES SEMANTIQUES POUR DECRYPTER LA CARTE ... 96

IV.2.1.1. Repères sémantiques sur la carte ... 96

IV.2.1.2. Spectres sémantiques pour représenter les ressources lors de l'indexation par propagation 98

IV.2.2. LE PRINCIPE "FOCUS AND CONTEXT" ... 99

IV.2.3. UTILISATION DE L'ANALYSE DE CONCEPTS FORMELS POUR LA VISUALISATION ... 100

IV.2.4. VISUALISATION DE PHOTOS SOCIALES SUR UN DIAGRAMME DE HASSE ... 102

IV.2.5. PASSAGE A L'ECHELLE : PERSPECTIVES DE RECHERCHE CONCERNANT LE NIVEAU D'ABSTRACTION 103

IV.3. EXPLICITER LES RESULTATS DE LA RECHERCHE D'INFORMATION ... 104

IV.3.1. PRESENTATION GLOBALE DE L'INTERFACE ET DES OUTILS DE PARAMETRISATION ... 105

IV.3.2. CONSTRUCTION DE LA CARTE SEMANTIQUE SYNTHETISANT LES RESULTATS ... 106

IV.3.3. REPRESENTATION DES ELEMENTS DU CORPUS DANS LA CARTE ... 107

IV.3.4. IDENTIFICATION DES PASSAGES PERTINENTS ... 109

IV.4. SYNTHESE ET PERSPECTIVES ... 109

V. CHAPITRE V ... 111

CONCLUSION, PERSPECTIVES ET PROJET DE RECHERCHE ... 111

V.1. PERSPECTIVES DE RECHERCHE ... 112

V.1.1. INDEXER RAPIDEMENT UN GRAND NOMBRE DE RESSOURCES ... 112

V.1.2. PROPOSER DE NOUVEAUX MODES DE RECHERCHE D'INFORMATION ... 113

V.1.2.1. Rechercher des informations complémentaires ... 113

V.1.2.2. Approfondir les mesures de distance/similarité sémantique pour les exploiter plus efficacement en fonction du contexte applicatif ... 114

V.1.3. INSTRUMENTER L'AUTOMATISATION COGNITIVE : VISUALISATION, INTERACTION ET PERSONNALISATION AU CŒUR DE LA DEMARCHE ... 115

V.2. POSITIONNEMENT AU SEIN DE L'ECOLE DES MINES D'ALES ET DE L'INSTITUT MINES-TELECOM ... 116

V.2.1. PERSPECTIVES CONCERNANT LA TRANSMISSION DES SAVOIRS ... 116

V.2.2. INNOVER, CREER AU SERVICE DU DEVELOPPEMENT ECONOMIQUE ET SOCIAL. ... 117

V.3. UN DOMAINE D'APPLICATION PRIVILEGIE : LE DOMAINE BIOMEDICAL ET LA SANTE ... 117

REFERENCES BIBLIOGRAPHIQUES ... 119

Table des illustrations

Figure 1 : Extrait d'une représentation graphique de la hiérarchie de concepts de l'ontologie de la musique. . 45
Figure 2 : Cycle de vie d'une ontologie 46
Figure 3 : Imbrication des différents niveaux de langage d'OWL 51
Figure 4 : Catalogage et indexation, deux moyens d'accéder à l'information (Baar, 2004) 59
Figure 5 : Interface d'indexation des segments vidéo sous forme de triplets dans le projet Prompter. 64
Figure 6 : Interface de l'outil MBox-Composer. 65
Figure 7 : Paysage de référence obtenu par projection du support d'indexation 66
Figure 8 : Schéma global d'un SRI conceptuel utilisant une ontologie de domaine 74
Figure 9 : Utiliser une ontologie pour éviter les *silences* dans un système basé sur une ontologie de domaine (extrait de la Gene Ontology – A) et un corpus indexé, interrogé par un ensemble de concepts (exemple de requête et de ressource indexée, ici, un gène – B). 78
Figure 10 : Schéma global de l'approche CoLexIR 83
Figure 11 : Interface de CoLexIR 84
Figure 12 : Carte sémantique représentant les titres musicaux avec des repères sémantiques 94
Figure 13 : Interface de navigation dans une collection de titres musicaux. 97
Figure 14 : Deux chemins représentant des *playlists* dans MBox-*Composer*. 98
Figure 15 : Visualisation de la propagation des poids des descripteurs grâce à un filtre sémantique. 99
Figure 16 : Interface de l'outil I^2DEE. 100
Figure 17 : Représentation d'un treillis de concepts à l'aide d'une projection MDS. 101
Figure 18 : Interaction entre une vue locale et une vue globale d'une base de brevets. 102
Figure 19 : Interface d'indexation de photos sociales. 103
Figure 20 : Interface d'OBIRS permettant d'exprimer une requête 105
Figure 21 : Interface d'OBIRS présentant les résultats d'une requête 106
Figure 22 : Interface de l'outil OBIRS, porté sur la plateforme collaborative de l'AvieSan ITMO Cancer 107
Figure 23 : Détail de l'interface d'OBIRS - Cadre d'explication de la sélection d'un élément du corpus 108
Figure 24 : Interface de présentation des résultats de CoLexIR 109
Figure 25 : Chaîne de traitement de la donnée à la décision 112

Curriculum vitae étendu

Sylvie Ranwez (née Chabert)

Née le 30 mars 1973, Nationalité française, Mariée, 2 enfants

Coordonnées professionnelles : LGI2P - Parc Scientifique G.Besse, F-30 035 Nîmes cedex 1

Tél. : 04 66 38 70 44, Fax : 04 66 38 70 74
Email : sylvie.ranwez@mines-ales.fr
URL : http://www.lgi2p.mines-ales.fr/~ranwezs/

Poste actuel

Chargée de recherche à l'école des mines d'Alès (EMA), dans l'équipe "*Knowledge and Image analysis for Decision making*" (KID) du Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P).

Domaine de recherche

Depuis plusieurs années, les chercheurs du centre de recherche LGI2P et plus particulièrement ceux de l'équipe de recherche KID (*Knowledge and Image analysis for Decision making*) à laquelle je suis rattachée, axent leurs activités de recherche autour de la maîtrise des systèmes complexes et de l'*automatisation cognitive*. Autrement dit, comment augmenter, dans un contexte de prise de décision, la réactivité et la fiabilité des opérateurs humains, en favorisant l'exécution de certaines tâches de façon automatique. Le contrôle de l'interaction homme/machine devient un élément clé de cette problématique. Mes recherches s'inscrivent pleinement dans cet axe, avec pour objectif d'**apporter à l'utilisateur une information pertinente et facile à interpréter, pour le guider dans sa prise de décision**. Pour cela, j'**exploite la sémantique** contenue dans les documents en mettant en œuvre les **ontologies**.

En effet, les ontologies qui formalisent une représentation de la connaissance d'un domaine, peuvent être utilisées comme guide sémantique lors de la navigation dans des bases de documents numériques toujours plus grandes et hétérogènes. Actuellement elles sont au cœur de certaines fonctionnalités des systèmes d'information (recherche d'information, filtrage, analyse de contenu, etc.) où elles fournissent un support pour le calcul de distances sémantiques, pour évaluer la pertinence de certaines ressources ou comme guide pour la visualisation de ces ressources.

Mes travaux de recherche ont pour objectif l'ajout et l'exploitation d'un niveau sémantique dans des applications dédiées à :

- la recherche d'information ;
- l'indexation et la navigation sémantique dans de grandes bases de données multimédia ;
- la visualisation de cartes de connaissances, cartes conceptuelles ;
- la composition de documents virtuels.

Formation et diplômes

2001,-5,-9	Qualification aux fonctions de maître de conférences, section informatique : CNU 27
1997-2000	Doctorat en informatique, université Montpellier II (UM II, sciences et techniques) – mention très honorable *"Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur"*
1996-1997	DEA en Informatique – mention AB *"Composition et Recomposition Narrative dans un contexte Multimédia"*

1994-1997	Diplôme d'Ingénieur en Informatique, École pour les Etudes et la Recherche en Informatique et Electronique (EERIE) – Option Intelligence Artificielle
1993-1994	DUT en Informatique, université Montpellier II
1991-1993	DUT en Mesures Physiques, université Montpellier II
1991	Baccalauréat série C – mention AB, académie de Montpellier

Thèse de doctorat

Titre de la thèse : *"Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur"*

École doctorale I2S (Information Structures Systèmes) de l'université Montpellier II – *spécialité informatique* – soutenue le 21 décembre 2000, préparée au Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) de l'école des mines d'Alès, site de Nîmes.

Jury composé de :

Mme Violaine Prince,	Professeur, Université Montpellier II,	Présidente
Mme Cécile Roisin,	Maître de conférences, Université de Grenoble II,	Rapporteur
M. Guy Gouardères,	Professeur, Université de Pau,	Rapporteur
M. Torsten Leidig,	Diplom-Informatiker, SAP, Corporate Research, Karlsruhe,	Examinateur
M. Michel Crampes,	Maître assistant, École des mines d'Alès,	Encadrant
M. Marc Nanard,	Professeur, Conservatoire National des Arts et Métiers,	Directeur de thèse

Expérience professionnelle en lien avec l'enseignement et la recherche

1er avril 2003 maintenant	Chargée de recherche au LGI2P, École des mines d'Alès, Site de Nîmes (*temps partiel 80%*)
1er nov. 2001 1er avril 2003	Chargée de Recherche au LGI2P (50%) et Responsable pédagogique de l'Institut EERIE (50%), École des mines d'Alès, Site de Nîmes
21 jan. 2001 21 oct. 2001	Ingénieur de recherche au LGI2P, École des mines d'Alès, Site de Nîmes
1er oct. 1998 30 déc. 2000	Deuxième et troisième année de thèse, au LGI2P, École des mines d'Alès
1er sept. 1997 30 sept. 1998	Première année de thèse, Digital Equipment − CEC Karlsruhe, Allemagne
1er févr. 30 juin 1997	Stage de recherche (DEA Informatique) au LGI2P, École des mines d'Alès, Site de Nîmes

Synthèse des travaux de recherche

La possibilité de stockage quasi illimité qu'offrent les supports numériques et l'extension de l'Internet engendrent une mutation de notre rapport à *l'Information*. Celle-ci ne prend vraiment tout son sens que dans un contexte particulier, face à une demande précise. Si pendant de nombreuses années l'opérateur humain utilisant un moteur de recherche relativement performant pouvait rapidement trier les données collectées pour satisfaire son besoin, il n'en est plus de même aujourd'hui, où dans nombre de domaines, des outils puissants doivent accompagner cette **recherche d'information** afin de la rendre vraiment efficace.

Mes recherches s'axent autour de la mise en place de ces outils, en utilisant des techniques de **modélisation de connaissance** associées à des techniques de **visualisation**.

Lors de ma thèse, je m'intéressais à la mise en place de nouveaux modes de transfert d'information, en particulier dans les environnements de formation à distance. Dans ce contexte, nous cherchions à **composer automatiquement des cursus personnalisés** à partir de *Briques d'Information* disponibles dans une base de données indexées. Pour cela j'ai travaillé sur la composition de documents virtuels personnalisables. J'ai élargi ce champ d'application, pour m'intéresser aujourd'hui à de nombreux domaines où la quantité d'information et son hétérogénéité freinent son exploitation (biomédical, innovation, gestion de collectifs). Guidée par un besoin concret exprimé par nos partenaires industriels et académiques, je m'efforce de trouver des **méthodes basées sur les modèles de connaissances, et plus particulièrement les ontologies, pour retrouver l'information pertinente, la filtrer, l'analyser et l'intégrer à un processus décisionnel**.

Dans ce contexte, des techniques visuelles adaptées peuvent s'avérer plus intuitives que des techniques classiques (moteur de recherche, listes de résultats, requêtes dans des bases de données) pour exploiter au mieux de nombreuses ressources d'information. Pour cela **les cartes de connaissances** présentées à l'utilisateur doivent être sémantiquement explicites. Nous nous sommes donc intéressés à la représentation sur un même paysage des données manipulées (documents textuels, photos, titres musicaux…) et de la connaissance du domaine correspondant. Nous combinons ainsi modélisation de connaissance et visualisation de cartes de connaissances pour favoriser :

- l'indexation de nombreuses ressources multimédia dans une collection ;
- la recherche d'information ;
- la navigation dans de larges corpus ;
- l'interaction homme-machine ;
- la composition de documents.

Dans ce cadre, nous utilisons les ontologies comme espace conceptuel pour éviter les silences lors de certaines requêtes (grâce à la spécialisation ou à la généralisation des termes de la requête), estimer la pertinence d'une réponse (en utilisant la pondération des concepts et la notion de distance/similarité sémantique) ou restreindre la zone de recherche (extraction d'une sous-ontologie).

Un médiateur unique pour accéder à des données hétérogènes distribuées.

La première thèse que j'ai co-encadrée était celle de Fabien Jalabert (soutenue en novembre 2007) et portait sur **l'intégration et la visualisation de données hétérogènes** dans le domaine biomédical. Cette thèse s'inscrivait dans un axe stratégique des écoles des Mines autour de la biologie : GEM-BIO[1]. Plusieurs partenaires biologistes déploraient le fait qu'ils doivent consulter simultanément de nombreuses bases de données spécifiques pour les analyser et les croiser. Cela entrainait un grand nombre de fenêtres ouvertes en même temps et une surcharge cognitive. L'objectif de la thèse de Fabien était de proposer une interface de médiation entre ces différentes applications et un point d'entrée unique vers les différentes bases. Cette interface synthétisait un premier niveau d'analyse et des regroupements entre certaines ressources. Un système a été développé : I2DEE (I*nteractive and Integrated Data Exploration Environment)* composé de deux modules principaux (Jalabert, Ranwez, Derozier, & Crampes, 2006). Le premier concernait l'intégration de données provenant de différentes bases (gènes, publications scientifiques, etc.). L'intégration était réalisée en utilisant UMLS[2] comme pivot sémantique. Le deuxième module concernait la visualisation de ces données dans une même interface. Cette thèse, très ambitieuse, a proposé des solutions intéressantes et ouvert de nombreuses perspectives. Mais elle a aussi soulevé certaines limites qui ont nourri nos réflexions futures : en particulier la surcharge cognitive de l'opérateur lors du passage à l'échelle et la nécessité de disposer les éléments en fonction de la sémantique qui leur est associée. Ces limites, nous y étions également confrontés dans le cadre de projets industriels.

Organiser les ressources en fonction de la sémantique et du contexte d'usage.

Je contribuais également à cette époque, à plusieurs projets exploitant des données multimédia[3] au travers d'interfaces graphiques. Le projet MBox-Composer, avait pour objectif la composition automatique de *playlists* musicales pour des chaines de magasins, des concessions automobiles ou des radios. Cette composition nécessitait une indexation particulière (autre que par les simples titres, auteurs et interprètes) et nous avons proposé un environnement pour indexer rapidement un grand nombre de titres musicaux : un paysage musical était présenté à l'utilisateur, dans lequel il suffisait de *glisser* et *déposer* un nouveau titre pour qu'il s'indexe automatiquement, par *propagation*. Cette indexation sur des cartes, pour être pleinement efficace (que tous les champs de description d'une ressource soient rapidement renseignés), nécessitait des cartes claires et faciles à interpréter pour l'utilisateur. Ces cartes devaient, pour avoir du sens, être accompagnées **d'indicateurs caractérisant la sémantique des différents éléments présents sur la carte** et respectant des proximités sémantiques. Nous avons proposé différentes vues sur les ressources en fonction de différents contextes : indexation, navigation, composition… La conception de ces cartes pouvait être difficile et nous avons réfléchi à la possibilité de les concevoir à partir de modèles, mis en liaison avec la connaissance du domaine. C'est ce que nous avons appelé l'approche *Domain-View-Control*. L'article "*Concept Maps for Designing Adaptive Knowledge Maps*" présente cette méthode (Crampes, Ranwez, Villerd, et al., 2006). Disposant d'une représentation de la connaissance du domaine et d'un méta-modèle de la visualisation dans notre environnement (MolAge), l'utilisateur peut créer un 'patron' de la visualisation en indiquant les endroits où il souhaite voir apparaître un type d'information. Le paysage final est alors présenté en respectant ce patron et les spécifications fixées par l'utilisateur.

La deuxième thèse que j'ai co-encadrée était celle de Jean Villerd (soutenue en novembre 2008). Elle avait pour objectif la **représentation visuelle de connaissances à partir d'une organisation formelle de ces connaissances**. En associant projection multidimensionnelle (MDS pour *MultiDimensional Scaling*) et analyse de concepts formels (FCA pour *Formal Concept Analysis*), Jean a proposé un environnement interactif où une représentation conceptuelle du domaine (son application concernait une base de brevets) interagissait avec une représentation des ressources (les

[1] Groupement des Ecoles des Mines
[2] Unified Medical Language System - http://www.nlm.nih.gov/research/umls/
[3] Ces différents projets seront détaillés dans la partie "Transfert de technologie".

brevets) pour permettre à l'utilisateur de filtrer progressivement les ressources et d'identifier celles qui sont pertinentes dans son contexte. Par ces travaux de recherche, Jean a apporté des réponses au manque de formalisation des données qui prévaut souvent dans les environnements de visualisation. Mais il a également proposé des solutions dans la communauté FCA en particulier concernant les données manquantes et des données indexées sur des dimensions mixtes (binaires, nominales ou continues). Cette recherche a conduit, notamment, à la publication de "*Using Concept Lattices for Visual Navigation Assistance in Large Databases: Application to a Patent Database*" (Villerd, Ranwez, Crampes, & Carteret, 2007).

Nous avons appliqué la **visualisation de cartes de connaissance** dans différents contextes : indexation de titres musicaux et composition automatique de *playlists* (projets MBox-composer et SAVIC), navigation dans un corpus de documents scientifiques (projet ToxNuc-e), navigation et recherche d'information dans une base de brevets industriels (projet INova), navigation dans des données hétérogènes issues des sciences du vivant. Dans ces différents domaines, nous nous sommes souvent heurtés à la difficulté de représenter visuellement les ontologies de domaine en complément des données composant le paysage. En effet, aucune alternative à la représentation traditionnelle sous-forme d'arborescence ne permet de présenter les concepts en fonction de leur proximité sémantique. Nous avons donc proposé dans "*Ontological ISA-Distance Measure for Information Visualisation on Conceptual Maps*", une **mesure de distance sémantique** basée sur la hiérarchie de concepts *is-a*. Cette mesure respecte les propriétés attendues des distances (positivité, symétrie et inégalité triangulaire). C'est cette mesure qui a été utilisée, couplée avec des techniques de projection MDS, dans la thèse de Jean Villerd pour la visualisation et la recherche d'information dans une base de brevets industriels

Recherche d'information conceptuelle et visualisation

La troisième thèse que j'ai co-encadrée, celle de Mohameth François Sy (soutenue en décembre 2012), a abordé la problématique de la **recherche d'information conceptuelle** et de la **visualisation des résultats**. Soucieux de fournir à un opérateur humain confronté à une prise de décision, une information la plus pertinente possible et une justification de sa pertinence, nous avons proposé un calcul d'adéquation entre une requête et un document indexé conceptuellement. Ce calcul permet d'estimer la contribution des différents concepts à la sélection d'un document. En lien avec des collègues de l'Inserm, nous l'avons intégré à un environnement de recherche de gènes (http://www.ontotoolkit.mines-ales.fr/ObirsClient/). L'approche choisie dans cette thèse s'inscrit dans la volonté de favoriser l'automatisation cognitive et les travaux ont été menés en prêtant une attention toute particulière à la perception des résultats par l'utilisateur. Nous avons également proposé une méthode de reformulation qui permet d'affiner les résultats d'une requête pour mieux répondre aux attentes de l'utilisateur.

Propriétés des mesures de similarité sémantique et cadre unificateur

Au cours de ces travaux, nous avons été confronté à l'impact de la mesure choisie pour **estimer la similarité sémantique** de deux concepts (et par extension la similarité sémantique d'entités indexées par des concepts) sur les résultats de la recherche d'information. Ces calculs doivent être optimisés (les ontologies actuelles peuvent contenir des dizaines de milliers de concepts), et leur étude approfondie est actuellement réalisée dans le cadre de la thèse de Sébastien Harispe (débutée en avril 2011). Un cadre unificateur de ces mesures a été proposé qui permet la définition de nouvelles méthodes de calcul. Une librairie pour analyser et comparer les principales mesures existantes a été développée (SML[4]), accompagnée par des benchmark pour les tester et choisir la plus appropriée dans un contexte donné.

Recherche d'informations complémentaires

Plus récemment, et dans la lignée des travaux présentés ci-dessus, une nouvelle thèse a commencé avec Nicolas Fiorini (débutée en octobre 2012), qui s'intéresse à la **recherche d'informations complémentaires** : la pertinence de plusieurs documents par rapport à une requête, n'étant plus fonction uniquement de leurs pertinences individuelles, mais de leur complémentarité vis-à-vis de cette requête. Cette recherche ne vise pas la diversification des résultats, mais à une analyse de leur complémentarité, c'est-à-dire que les résultats sélectionnés doivent être à la fois proches mais distincts par rapport à la requête. Une formalisation est en cours et les premiers développements (extension du moteur OBIRS) ont débuté.

Liens avec d'autres institutions de recherche

Dans le cadre de mes activités de recherche, j'ai également participé à des **collaborations avec d'autres laboratoires de recherche**. Dans le cadre de projet contractualisé, comme ce sera détaillé par la suite, nous avons collaboré avec des chercheurs de l'Inserm. Avec le laboratoire ISEM de l'université Montpellier II, j'ai participé à la réalisation du site OrthoMam (http://www.orthomam.univ-montp2.fr/) qui permet de rechercher des gènes pertinents pour la reconstruction de l'arbre de l'évolution de mammifères (étude de leur phylogénie). Actuellement, une collaboration est en cours avec des chercheurs de l'université Rovira i Virgili de Tarragone (Espagne). Nous avons accueilli durant l'automne 2012 Montserrat Batet-Sanroma, avec qui nous avons réfléchi à une mesure de distance sémantique entre

[4] http://www.semantic-measures-library.org/sml/

concepts issus de plusieurs ontologies. Cette collaboration se poursuit puisqu'au printemps 2013, c'est David Sánchez qui rejoint notre équipe pour trois mois.

Il faut noter que mes travaux ont toujours été menés avec le souci de l'utilisation finale des méthodes et outils conçus. Cela m'a souvent amenée à échanger avec mes collègues des sciences humaines et sociales. Ainsi, de janvier 2004 à juin 2006, j'ai participé et co-animé des ateliers/séminaires sur le "Web-sémantique et son impact dans les nouveaux modes d'apprentissage", dans le département des Sciences de l'Information et la Communication de l'Université Paul Valéry (Montpellier III), dirigé par Alex Mucchielli. De ces échanges sont nés des collaborations, notamment avec Isabelle Rondeau, avec qui nous avons co-écrit une publication. Plus récemment, c'est également ce même souci qui m'a poussée à co-encadrer des mémoires de Master CTN qui s'intéressaient aux usages des nouvelles technologies et à leur impact sur nos modes de vie.

Animation de la communauté scientifique

Comités de programme et relecture

Relecteur pour la revue STE (sciences et techniques éducatives) en 2002.
Colloque National sur la Recherche en Informatique et ses Applications, Université de Ziguinchor, Sénégal, 25 au 27 avril 2013.
Conférence IC2013 : Ingénierie des connaissances, Lille, 3-5 juillet, 2013.
Conférence IC2012 : Ingénierie des Connaissances, Paris, 25-29 juin 2012.
Conférence IC2011 : Ingénierie des Connaissances, Chambéry, 16-20 mai 2011.
Conférence IHM 2011, Nice - Sophia Antipolis, 24-27 octobre 2011.
Conférence IC2010 : Ingénierie des Connaissances, Nîmes, 8-11 juin 2010.
Conférence DVP'02 : Documents Virtuels Personnalisables 2002, Brest, 10-11 Juillet 2002.

Comités d'organisation

Conférence IC2010, Nîmes, 8-11 juin 2010 (gestion du contenu du site Web).

Session : "Systèmes Dynamiques de Gestion de la Connaissance et du Multimédia". Dans le cadre des Journées sur l'Ingénierie de Systèmes et les NTIC – NimesTIC'2000 , Nîmes, 11-13 septembre 2000.

Workshop : "Documents Virtuels Personnalisables : de la Définition à l'Utilisation" Dans le cadre de la11ème conférence francophone IHM'99 (Interaction Homme-Machine), Montpellier, 22-26 Novembre 1999.

Expertise scientifique, évaluation de projet

European Cooperation in the field of Scientific and Technical Research – COST[5], action TU0801 : "Semantic enrichment of 3D city models for sustainable urban development".

Comités de thèse

Benjamin Duthil (2010 et 2011). Thématique : Analyse de textes, segmentation et détection d'opinion.

Encadrements (1 post-doc, 5 thèses dont 2 en cours, 5 Master II dont 1 en cours)

Post-Doctorant

☑ François-Elie Calvier (février 2012 – février 2013) (*co-encadrement à 25% avec Gérard Dray, 25% et Michel Plantié, 50%*)

Apprentissage basé sur les ontologies pour la classification multi-classes de documents scientifiques du domaine biomédical.

3 thèses soutenues et 2 en cours *(par ordre chronologique)*

☑ **Fabien Jalabert** (thèse soutenue le 5 novembre 2007)
Cartographie des connaissances et ingénierie ontologique - application aux sciences du vivant

Directeur de thèse : Michel Crampes (*35%*), encadrants de proximité : Sylvie Ranwez (*35%*), Vincent Derozier (*30%*).

Thématique : Nous nous sommes intéressés à la représentation sur une même carte de connaissances, de données hétérogènes (données d'expression génique, documents scientifiques, bases de connaissance générales, ou centrées sur *Plasmodium falciparum*, *Gene Ontology*, etc.) de façon à identifier des proximités sémantiques et favoriser, par exemple, une recherche documentaire ciblée sur certains gènes dérégulés par le

[5] http://www.sbf.admin.ch/htm/themen/international/cost_fr.html

paludisme. Nous avons également travaillé sur l'aspect "*médiateur ontologique*" entre différents modèles issus de communautés différentes (chimistes, biologistes, etc.).

Depuis sa thèse : Ingénieur en développement J2EE dans une SSII depuis septembre 2007. Participation à divers projets : système de gestion de combat chez un leader du naval militaire, planification de ressources en *conferencing*, e-commerce, ERP (Enterprise Resource Planning) pour une société d'aménagement dans le domaine de l'irrigation, de la production et la distribution d'eau sanitaire.

☑ **Jean Villerd** (thèse soutenue le 19 novembre 2008)
Représentations visuelles adaptatives de connaissances associant projection multidimensionnelle (MDS) et analyse formelle de concepts (FCA)

Directeur de thèse : Michel Crampes (*50%*), encadrante de proximité : Sylvie Ranwez (*50%*).

Thématique : La problématique centrale de cette thèse était la navigation dans de larges bases de connaissance et le passage à l'échelle : les solutions proposées devaient en effet favoriser une navigation conceptuelle à travers les données, grâce à une visualisation sous forme de carte de connaissances. Nous avons proposé une navigation conceptuelle sur un treillis de Galois pour filtrer l'information et proposer des documents pertinents à l'utilisateur. La visualisation sous forme de cartes nous a poussés à nous intéresser aux propriétés des mesures de similarité sémantique et nous avons proposé une mesure de distance pour favoriser une projection pertinente des données.

Depuis sa thèse : post-doctorant de 2008 à 2010 dans l'équipe Orpailleur (Loria - INRIA Nancy Grand Est) : fouille de données médicales (pharmacovigilance) fondée sur l'Analyse de concepts formels. Depuis le 1er octobre 2010, ingénieur de recherche INRA au sein du Laboratoire Agronomie et Environnement (unité mixte INRA/Institut National Polytechnique de Lorraine). Modélisation et analyse de données environnementales pour le développement d'indicateurs agri-environnementaux (phytosanitaires et biodiversité).

☑ **Mohameth François Sy** (thèse soutenue le 11décembre 2012)
Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents

Directeur de thèse : Michel Crampes (*25%*), encadrants de proximité : Sylvie Ranwez (*50%*), Vincent Ranwez (*25%*).

Dans le cadre de cette thèse, nous nous sommes intéressés à la recherche d'information exploitant une indexation conceptuelle, i.e. utilisant des ontologies. Le domaine d'application que nous avons privilégié concerne le domaine biomédical. L'espace conceptuel fourni par une ontologie de domaine est utilisé, d'une part, comme espace d'indexation des documents et requêtes et d'autre part comme une source de métriques à travers des mesures de similarité sémantique, pour des mesures de pertinence (adéquation document/requête). Dans une démarche d'automatisation cognitive, un soin particulier a été apporté à la présentation des résultats à l'utilisateur et à l'interaction homme-machine.

Depuis sa thèse : post-doctorant de 2013 à 2014 dans l'équipe Metah (Modèles et Technologies pour l'Apprentissage Humain) au LIG (Laboratoire d'Informatique de Grenoble) : Référentiel ontologique Open Data de compétences et connaissances

↺ **Sébastien Harispe** (avril 2011 – soutenance prévue en mars 2014)
Algorithmes d'Optimisation pour la Recherche d'Information basée sur des Ontologies de Grande Taille

Directeur de thèse : Jacky Montmain (*20%*), encadrants de proximité : Stefan Janaqi (*40%*), Sylvie Ranwez (*40%*).

Cette thèse s'intéresse à l'optimisation d'algorithmes pour le traitement d'ontologies de grande taille en particulier dans le cadre de la recherche d'information conceptuelle. La première partie de la thèse s'est focalisée sur la définition d'un cadre unificateur qui permet notamment de classer différentes mesures de distances sémantiques en fonction de leurs caractéristiques. Une librairie a été développée pour faciliter la sélection de la mesure la plus appropriée à un contexte applicatif donné en les comparant sur plusieurs jeux de données.

↺ **Nicolas Fiorini** (octobre 2012 – soutenance prévue en septembre 2015)

Utilisation d'ontologies comme support à la recherche d'informations complémentaires
Directeurs de thèse : Jacky Montmain (*25%*), Vincent Ranwez (*25%*), encadrant de proximité : Sylvie Ranwez (*50%*).

De nombreuses innovations scientifiques et technologiques découlent de la capacité à identifier et croiser des technologies, des méthodes et des connaissances issues de domaines connexes, et donc différentes, mais complémentaires en regard d'une problématique donnée. Des informations sont dites complémentaires si leur pertinence globale est augmentée par le fait de les considérer ensemble. Cette thèse propose d'investiguer la complémentarité d'information en RI qui se rapproche de la notion de "*query result diversification*".

5 Master II *(dont un en cours)* **et 1 stage de licence** *(par ordre chronologique)*

☑ **Stéphanie Baar** (2004) (*taux d'encadrement 100%*) – Stage de fin d'étude du DESS Auteur Rédacteur Multimédia (ARM[6]), spécialité sciences de l'information et de la communication et informatique.
"*Préconisations pour la création d'un module d'indexation dans un outil de développement d'applications Web"*. (Baar, 2004)

☑ **Makhtar Mbao** (2007) (*taux d'encadrement 60%*) – Stage de master II recherche en informatique (université Montpellier II).
"Distance Sémantique et Carte Conceptuelle" (Mbao, 2007)

☑ **Cécile Wolff** (2009 – 2010) (*taux d'encadrement 50%*) – Mémoire de master II Communication et Technologie Numérique (CTN[7]).
"*Les Technologies de l'Information et de la Communication contribuent-elles à l'essor d'une société de l'information en adéquation avec les enjeux sociaux du développement durable. Entre mythe et réalité, quelle appropriation de l'internet par les 16-25 ans en France ?"* (Wolff, 2010)

☑ **Véronique Gachet** (2011) (*taux d'encadrement 50%*) – Stage de master II CTN
Formalisation de l'ontologie du domaine de la toxicologie nucléaire environnementale.
Dans le cadre du projet ToxNuc et de l'accompagnement de la communauté scientifique. Stage effectué au CEA.

☑ **Pierre Fontana** (2012) (*taux d'encadrement 30%*) – Stage de licence
Constitution d'une base de données de publications scientifiques liées au Cancer, dans le cadre de la plateforme collaborative AvieSan Cancer.

↺ **Gilbert Okaro** (2012 – 2013) (*taux d'encadrement 50%*) – Mémoire de master II Communication et Technologie Numérique (CTN). *"Développement du télétravail en France : les technologies numériques au service du développement socio-économique des territoires ruraux."*

Publications

Dans les références des articles co-écrits avec des étudiants que j'ai encadrés, leur nom est souligné.

1 chapitre d'ouvrage	**How ontology based information retrieval systems may benefit from lexical text analysis.** *Sylvie Ranwez, Benjamin Duthil, Mohameth François Sy, Jacky Montmain, Patrick Augereau, Vincent Ranwez.* In "New Trends of Research in Ontologies and Lexical Resources", chapter 11, pp. 209-230, Series: Theory and Applications of Natural Language Processing, Oltramari, Alessandro; Vossen, Piek; Qin, Lu; Hovy, Eduard (Eds.), Springer, ISBN 978-3-642-31781-1, February 2013.
↺ 3 revues internationales en soumission	**An information theoretic approach to improve semantic similarity assessments across multiple ontologies.** *Montserrat Batet, Sébastien Harispe, Sylvie Ranwez, David Sánchez, Vincent Ranwez.* Submitted to Information Sciences **A unifying theoretical framework for the study and definition of semantic measures - application to bio-ontologies.** *Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi and Jacky Montmain.* Submitted to BMC bioinformatics. **Efficient conceptual relevance feedback using semantic neighborhood connectivity.** *Mohameth François Sy, Sylvie Ranwez, Jacky Montmain, Vincent Ranwez.* Submitted to TKDE
8 revues internationales	**Subontology Extraction Using Hyponym and Hypernym Closure on is-a Directed Acyclic Graphs.** *Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi.* In IEEE Transactions on Knowledge and Data Engineering (**TKDE**), volume 24, Issue 12, IEEE computer Society Digital Library, ISSN: 1041-4347, pp. 2288-2300, December 2012. **An O(n.m) algorithm for calculating the closure of lca-type operators.** *Vincent Ranwez, Stefan Janaqi, Sylvie Ranwez.* In **Ars Combinatoria**, Volume 104, pp. 107-128, 2012. **User Centered and Ontology Based Information Retrieval System for Life Sciences**.*Mohameth-François Sy, Sylvie Ranwez, Jacky Montmain, Armelle Regnault, Michel Crampes, Vincent Ranwez.* In **BMC Bioinformatics**, 13(Suppl 1):S4, 2012. **Visualizing Social Photos on a Hasse Diagram for Eliciting Relations and Indexing New Photos**.*Michel Crampes, Jeremy de Oliveira-Kumar, Sylvie Ranwez, Jean Villerd.* In IEEE Transactions on Visualization and Computer Graphics (**TVCG**), Edited by Kwan-Liu Ma, Torsten Möller, Hanspeter Pfister, Sheelagh Carpendale, Jean-Daniel Fekete, Volume 15, Issue 6, ISSN: 1077-2626, pp. 985-992, November/December 2009.

[6]co-habilité par l'Université Paul Valéry, Montpellier III et l'Ecole des Mines d'Alès.

[7]co-habilité par le Celsa – université Paris Sorbonne et l'Ecole des Mines d'Alès.

Using Concept Lattices for Visual Navigation Assistance in Large Databases: Application to a Patent Database.*Jean Villerd, Sylvie Ranwez, Michel Crampes, David Carteret.* In **Journal of General Systems**, Special Issue on "Concept Lattices and their Applications", Volume 38, Issue 4, pp. 405-425 May 2009.

OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. *Vincent Ranwez, Frederic Delsuc, Sylvie Ranwez, Khalid Belkir, Marie-Ka Tilak, Emmanuel JP Douzery.* In **BMC Evolutionary Biology**, Volume 7, Issue 241, Novembre, 2007.

Concept Maps for Designing Adaptive Knowledge Maps. *Michel Crampes, Sylvie Ranwez, Jean Villerd, Filip Velickovski, Christopher Mooney, Andrew Emery, Nicolas Mille.* In **Information Visualization,** Special Issue on Concept Maps, Volume 5, Issue 3, Guest Editors: S.-O. Tergan, T. Keller and R. Burkhard, Palgrave - Macmillan, 2006.

Formalization to Improve Life-Long Learning. *Sylvie Ranwez, Torsten Leidig and Michel Crampes.* In **Journal of Interactive Learning Research**, Special double issue on "Intelligent Systems/Tools in Training and Life-Long Learning", AACE, Vol. 11, N°3/4, pp. 389-409, 2000.

3 revues nationales

Extraction de sous-ontologies autonomes par fermeture des opérateurs hyponymie et hyperonymie. *Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi.*Les ontologies à l'épreuve des usages, Ladjel Ballatreche, Gilles Kassel et Philippe Thiran eds, **Technique et Science Informatiques**, Volume 31, N°1, pp. 11-38, Hermès/Lavoisier, 2012.

Qualités d'une indexation portée par XML et une ontologie au regard d'un standard. *Michel Crampes, Sylvie Ranwez, Michel Plantié et Christophe Vaudry.* **Sciences et techniques éducatives**, Hors série 2003, "Ressources numériques, XML et éducation", Eric Bruillard et Brigitte de La Passardière eds., Hermès/Lavoisier, pp.105-134, 2003.

Instanciation d'Ontologies Pondérées et Calcul de Rôles Pédagogiques – Principe et mise en œuvre. *Sylvie Ranwez et Michel Crampes.* Sciences et techniques éducatives, Volume 9, N°3, Hermès/Lavoisier, pp.341-370, 2002.

10 conférences internationales

Visualising Social Photos on a Hasse Diagram for Eliciting Relations and New Photo Indexing. *Michel Crampes, Jeremy de Oliveira-Kumar, Sylvie Ranwez and Jean Villerd.*In proceedings of InfoVis 2009, the **IEEE Information Visualization** Conference. Atlantic City, New Jersey, October 11-16, 2009.

Using Concept Lattice for Visual Navigation Assistance and Attribute Selection. *Jean Villerd, Sylvie Ranwez, Michel Crampes.* In Supplementary Proceedings of **ICCS 2008**, the 16th International Conference on Conceptual Structures, Peter Eklund and Ollivier Haemmerlé edition, pp. 41-48, Toulouse, France, July 7-11, 2008.

Using Concept Lattices for Visual Navigation Assistance in Large Databases: Application to a Patent Database.*Jean Villerd, Sylvie Ranwez, Michel Crampes, David Carteret.* In **CLA 2007**, the fifth International Conference on Concept Lattices and Their Applications. Jean Diatta, Peter Eklund and Michel Liquière editors, pp. 88-99, Montpellier, France, October 24-26, 2007.

Cross-linking Music and Pictures through Moods. *Michel Crampes, Olivier Gout, Nicolas Mille, Jean Villerd, Sylvie Ranwez.*In proceedings of Multimedia Systems and Applications, **MSA'07**, within the scope of the 2007 World Congress in Computer Science Computer Engineering, and Applied Computing, WORLDCOMP'07, Monte Carlo Resort, Las Vegas, Nevada, USA, June 25-28, 2007.

Automatic Playlist Composition in a Dynamic Music Landscape. *Michel Crampes, Jean Villerd, Andrew Emery, Sylvie Ranwez.***ACM International Conference Proceeding Series**, Vol. 259, Proceedings of the 2007 International Workshop On Semantically Aware Document Processing And Indexing, **SADPI'07**, In Cooperation with ACM SIGWEB, Montpellier, France, May 21-22, 2007. ISBN:978-1-15159-668-4.

Ontological ISA-Distance Measure for Information Visualisation on Conceptual Maps. *Sylvie Ranwez, Vincent Ranwez, Jean Villerd, Michel Crampes.* In Proceedings of On the Move to Meaningful Internet Systems 2006: **OTM 2006**, Workshops on Ontology content and evaluation in Enterprise, Lecture Notes in Computer Science, publisher: Springer Berlin / Heidelberg, Volume 4278/2006, ISBN 978-3-540-48273-4, pp.1050-1061, 2006.

i²dee : An Integrated and Interactive Data Exploration Environment Used for Ontology Design.*Fabien Jalabert, Sylvie Ranwez, Vincent Derozier and Michel Crampes.* In Managing Knowledge in a World of Networks, proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (**EKAW 2006**), Lecture Notes in Computer Science, publisher: Springer Berlin / Heidelberg, Volume 4248/2006, ISBN 978-3-540-46363-4, pp. 256-271, 2006.

An Integrated Visual Approach for Music Indexing and Dynamic Playlist Composition. *Michel Crampes, Sylvie Ranwez, Filip Velickovski, Christopher Mooney and Nicolas Mille.***MMCN 2006**, Thirteenth Annual Multimedia Computing and Networking, January 18-19, 2006, San Jose, California.

Ontology-Supported and Ontology-Driven Conceptual Navigation on the World Wide Web. *Michel Crampes and Sylvie Ranwez.***HT'00**, the 11th ACM Conference on Hypertext San Antonio, Texas.

	Adaptive Narrative Abstraction. *Michel Crampes, Jean-Paul Veuillez and Sylvie Ranwez.* Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia, **Hypertext'98**, Pittsburgh, Pennsylvania, June 20-24,1998
↺ 2 conférences internationales en soumission	**From Theoretical Framework To Generic Semantic Measures Library.** *Sébastien Harispe, Stefan Janaqi, Sylvie Ranwez and Jacky Montmain.* ODBase 2013. **Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems.** *Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain*. ODBase 2013.
Autres : conférence internationale sans actes	**ORTHOMAM: a database of candidate orthologous makers for mammalian phylogenetics**.*Vincent Ranwez, Frédéric Delsuc, Sylvie Ranwez, Marika Tilak-Jean, Emmanuel Douzery.*In Evolution 2007, the joint annual meeting of the Society for the Study of Evolution (SSE), the Society of Systematic Biologists (SSB), and the American Society of Naturalists (ASN). The Christchurch Convention Centre, Christchurch, New Zealand, June 16-20, 2007.
14 conférences nationales	**Mesures sémantiques basées sur la notion de projection RDF pour les systèmes de recommandation.** *Sébastien Harispe, Sylvie Ranwez, Stefan Janaqiet Jacky Montmain*. 24e journées francophones d'ingénierie des connaissances - IC 2013, Lille, 3-5 juillet 2013 **Ontology based machine learning for semantic multiclass classification.** *François-Elie Calvier, Michel Plantié, Gérard Dray, Sylvie Ranwez.* Terminology & Ontology : Theories and applications, TOTH 2013, Chambéry, France, 6-7 juin 2013 **OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine**. *Mohameth François Sy, Sylvie Ranwez, Jacky Montmain, Vincent Ranwez.* Actes de la neuvième édition de la COnférence en Recherche d'Information et Applications, **CORIA 2012** dans le cadre de la semaine du document numérique et de la recherche d'information, Michel Beigbeder, Véronique Eglin, Nicolas Ragot et Mathias Géry eds., pp. 135-150, Bordeaux, France, 21-23 mars 2012. **Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information**. *Sylvie Ranwez, Vincent Ranwez, Mohameth-François Sy, Jacky Montmain et Michel Crampes.* dans les actes des 21es Journées Francophones d'Ingénierie des Connaissances **IC 2010**, Editeur : Sylvie Desprès, Presse des Mines, ISBN 978-2-911256-25-7, pp. 247-258, École des mines d'Alès - site de Nîmes, Nîmes, 9-11 juin 2010. **Extraction de sous-ontologies autonomes par fermeture des opérateurs hyponymie et hyperonymie**. *Vincent Ranwez, Sylvie Ranwez, Stefan Janaqi.* dans les actes de la 3ème édition des journées francophones sur les ontologies **JFO 2009**, Editeurs : Ladjel Ballatreche, Gilles Kassel, Philippe Thiran, ACM 978-1-60558-842-1, pp. 45-55, LISI/ENSMA, Futuroscope, Poitiers, 3-4 décembre 2009. **Indexation de photos sociales par propagation sur une hiérarchie de concepts**. *Michel Crampes, Jeremy De Oliveira-Kumar, Sylvie Ranwez, Jean Villerd.***IC 2009**, 20es Journées Francophones d'Ingénierie des Connaissances, Hammamet, Tunisie, 25 au 29 mai 2009. **Cartographie sémantique auto-organisée d'un référentiel de connaissances partagé**. *Michel Crampes, Sylvie Ranwez, Jean Villerd.***IC 2008**, 19èmes journées francophones d'Ingénierie des Connaissances, Nancy, France, 18 au 20 juin 2008. **Navigation sur des cartes de connaissances supportées par un treillis de Galois.** *Jean Villerd, Sylvie Ranwez, Michel Crampes.* **CARTO 2.0**, "Où en êtes-vous de la mise en scène de vos informations ?" pp. 105-113, Champs Sur Marne, France, 3 avril 2008. **Cartes conceptuelles pour l'ingénierie de cartes de connaissances adaptatives**. *Michel Crampes, Jean Villerd, Filip Velickovski, Nicolas Mille et Sylvie Ranwez.* **IC 2006**, 17èmes journées francophones d'Ingénierie des connaissances, dans le cadre de la Semaine De la Connaissance, NANTES, France, 28-30 juin 2006. **I²DEE : intégrer et visualiser des données biologiques pour concevoir une ressource termino-ontologique**. *Fabien Jalabert, Sylvie Ranwez, Michel Crampes et Vincent Derozier.***IC 2006**, 17èmes journées francophones d'Ingénierie des connaissances, dans le cadre de la Semaine De la Connaissance, Nantes, France, 28-30 juin 2006. **Approche multidimensionnelle du texte pour le balisage des ressources pédagogiques**. *Isabelle Rondeau, Sylvie Ranwez et Michel Crampes.* Journée thématique : Web sémantique pour le e-Learning - AFIA 2005, Coordonnateurs : Rose Dieng-Kuntz, Monique Grandbastien, Danièle Hérin, Nice, France, 31 mai, 2005. **Initiative mixte dans les DVP : de la Pertinence à l'Adaptation**. *Christophe Vaudry, Sylvie Ranwez, Astrid Poulon et Michel Crampes.* **DVP 2002**, Documents Virtuels Personnalisables, Coordonnateurs : Serge Garlatti et Michel Crampes, ENST Bretagne, Brest, France, pp. 141-154, 10-11 juillet, 2002. **Algorithme pour la génération de Documents Personnalisés Structurés - Optimisation par fusion des classes d'utilisateurs.** *Vincent Ranwez et Sylvie Ranwez.* **NîmesTIC'2001**, La relation Homme-Système : Complexe ?, Ed. École des mines d'Alès, Nîmes, France, pp. 147-152, 12-14 décembre, 2001. **Méta-description en XML de documents vidéo.** *Sylvie Ranwez et Michel Crampes.* présenté à **ISKO-France'99**, Lyon, Fance, 21-22 Octobre, 1999.

7 ateliers nationaux et internationaux	**User Centered and Ontology Based Information Retrieval System for Life Sciences**. *Sylvie Ranwez, Vincent Ranwez, Mohameth François Sy, Jacky Montmain and Michel Crampes.* Semantic Web Applications and Tools for Life Sciences, SWAT4LS, Berlin, Germany, December 10th, 2010. **I2DEE : Intégration et visualisation de données hétérogènes pour accéder à l'information**. *Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier.* Ontologie, Grille et Intégration Sémantique pour la Biologie, réunion satellite à la conférence JOBIM 2006, Bordeaux, 4 juillet 2006. **Comparaison interactive de regroupements - Une application aux données d'expressions de Plasmodium Falciparum**. *Fabien Jalabert, Michel Crampes, Vincent Derozier, Sylvie Ranwez.***RIA 2006**, 2èmes Rencontres Inter-Associations, Laboratoire ERIC, Université Lumière Lyon 2, 20-21 Mars 2006. **Médiation ontologique pour l'ingénierie des connaissances**. *Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier.* Ontologie, Grille et Intégration Sémantique pour la Biologie, réunion satellite à la conférence **JOBIM 2005**, Lyon, 4 juillet 2005. **L'analogie comme support à la composition adaptative - Comment utiliser les ontologies pour servir de base à l'adaptation de la rhétorique.** *Sylvie Ranwez.* Journées communes CERIC-LGI2P sur l'intelligence collective et le travail collaboratif, Site EERIE, École des mines d'Alès, 29-30 septembre 2004. **Description and Construction of Pedagogical Material using an Ontology based DTD**. *Sylvie Ranwez, Michel Crampes and Torsten Leidig.* Workshop on Ontologies for Intelligent Educational Systems, Ninth International Conference on Artificial Intelligence in Education, AI-ED'99, Le Mans, France, July 19-23, 1999. **Conceptual Documents and Hypertext Documents are two Different Forms of Virtual Documents**. *Sylvie Ranwez and Michel Crampes.* Proceedings of the Workshop on Virtual Documents, Hypertext Functionnality and the Web, Eighth International World Wide Web Conference, Toronto, Canada, May 10-15, 1999.
7 posters et démonstrations	**OBIRS, un outil de requêtage conceptuel.** *Mohameth-François Sy, Sylvie Ranwez, Vincent Ranwez, Jacky Montmain et Michel Crampes.* 22es Journées Francophones d'Ingénierie des Connaissances IC 2011, Chambéry, 16-20 Mai 2011. **Cartographies multiples d'un collectif de chercheurs.** *Jean Villerd, Imane Anoir, Michel Crampes, Andrew Emery, Sylvie Ranwez, Jean-Michel Penalva.* Rencontres Intelligence Collective (RIC 2006), École des mines d'Alès, Nimes, 22-24 mai 2006. **Architecture dynamique, distribuée et à composants pour les traitements de corpus.** *Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier.* Journée d'étude de l'ATALA, ENST Paris, 12 février 2005. **Construction d'une carte de connaissance dans le domaine des sciences du vivant.** *Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier.* JOBIM 2005, Lyon, France, Juillet 2005. **Médiation et environnement intégré d'ingénierie ontologique.** *Fabien Jalabert, Sylvie Ranwez, Michel Crampes, Vincent Derozier.* IC 2005, 16e journées francophones d'Ingénierie des connaissances, Nice, 31 mai - 3 juin 2005. **Forum du Groupe Connaissance et Complexité du LGI2P : Environnements Multimédias Adaptatifs.** *Christophe Vaudry, Sylvie Ranwez, Michel Crampes, Jacky Montmain.* Panoramas des laboratoires et organismes, actes de IHM-HCI 2001, Lille, France, Volume II, pp. 273-274, Septembre 2001. **Forum de l'Equipe Multimédia et Interfaces Homme-Machines Adaptatives du LGI2P**. *Christophe Vaudry, Michel Crampes et Sylvie Ranwez.* Presenté lors de la 11ème conférence francophone IHM'99 (Interaction Homme-Machine), Montpellier, France, 22-26 Novembre 1999.

Transfert de technologie

En 2010 et 2011, j'ai été coordonnatrice pour deux propositions de projet soumises par un consortium public-privé, en réponse à l'appel ContInt (Contenus et Interactions) de l'Agence Nationale de la Recherche. Ces propositions n'ont malheureusement pas été retenues pour un financement de l'ANR, cependant, elles sont mentionnées car elles ont mobilisé des énergies et été à la source de plusieurs collaborations et travaux (thèse de N. Fiorini, par exemple).

Par ailleurs, je suis partenaire de trois projets d'accompagnement de communautés scientifiques via l'outillage de plateformes collaboratives : ToxNuc3, SensoMines et AvieSan/Inserm. Pour ces trois projets nous proposons à nos partenaires académiques (respectivement CEA, Groupe Carnot M.I.N.E.S. et Inserm) une plateforme collaborative intuitive pour la gestion de leur collectif et surtout pour la gestion des documents déposés sur cette plateforme. En effet, ces plateformes doivent gérer un grand nombre de documents scientifiques ou administratifs. Nous réfléchissons à la gestion de ces documents et à leur exploitation (navigation, indexation, recherche d'information) via un modèle de connaissance du domaine (ontologie) associé à une interface visuelle. Il nous semble difficile d'envisager de

construire de manière totalement manuelle les ontologies. Aussi nous travaillons sur l'extraction automatique de candidats-termes à partir des corpus, et à leur mise en relation grâce à des techniques de *clustering*.

Au sein des écoles des Mines, les activités de recherche sont souvent portées par des projets contractualisés avec des industriels ou financés dans le cadre de projets nationaux ou européens. Si cette organisation peut paraître contraignante, elle permet de mener une recherche en lien direct avec les préoccupations des utilisateurs finaux et de disposer de suffisamment de retour sur utilisation et de retours d'experts pour orienter et conforter nos approches. Je me suis investie dans plusieurs projets pour lesquels voici quelques informations complémentaires.

CIRCE – appel à projet ContInt 2011 de l'ANR (*projet non sélectionné*) – Coordinatrice de cette proposition.

Recherche et croisement d'informations complémentaires pour favoriser l'innovation – Complementary Information Retrieval and Crossing Environment for innovation

Résumé : Un grand nombre d'innovations technologiques découlent du croisement de techniques, de méthodologies, de connaissances, entre plusieurs domaines. La recherche de ces ressources peut être considérée comme une recherche d'informations complémentaires pour atteindre un objectif scientifique précis. Le projet CIRCE avait pour but la **formalisation de l'innovation et de la complémentarité en recherche d'information et la conception des modèles et systèmes associés**. Les éléments d'information recherchés peuvent être des documents scientifiques ou bien des personnes (décrites suivant leurs compétences, par exemple). Cependant il est difficile de contrôler le processus de RI dans ce contexte, en particulier parce que la pertinence et la complétude des résultats sont difficiles à évaluer, les résultats valant par leur complémentarité, l'évaluation de leur pertinence ne peut plus se faire de manière indépendante, document par document. Même si ce projet n'a pas été financé par l'ANR, cet axe de recherche est celui retenu pour la thèse de Nicolas Fiorini, sur financement propre.

Partenaires : Centre de recherche commun **Armines/LGI2P** de l'école des mines d'Alès, Inova Software, CNRS DIST (Direction de l'Information Scientifique et Technique), Inserm DISC (Département de l'information scientifique et de la communication), LSIIT (Laboratoire des Sciences de l'Image, de l''Informatique et de la Télédétection – université de Strasbourg, CNRS), LGECO (Insa de Strasbourg), Praxiling université Montpellier III, ISEM université Montpellier II.

DISTIL – appel à projet ContInt 2010 de l'ANR (*projet non sélectionné*) – Coordinatrice de cette proposition.

Indexation et Fouille de Documents sur des Paysages Interactifs – Document Indexing and Search Through Interactive Layout

Résumé : les médias numériques ont crû et évolué sous l'impulsion d'une offre technologique innovante en matière d'échange et d'interaction, réappropriée par une intelligence venant de la base. Ce sont désormais les *utilisateurs des TIC*, organisés en réseaux, qui sont les principaux instigateurs et producteurs de contenus. Les réseaux permettent, en effet, non seulement d'administrer de façon globale et extensive différentes communautés, mais aussi de reformuler l'agencement des compétences et leur répartition dans ces mêmes communautés. Le projet DISTIL s'inscrivait dans cette dynamique et avait pour but la conception de **méthodes et de technologies innovantes pour la gestion de corpus documentaires,** en particulier dans des réseaux scientifiques, en utilisant des modèles et des techniques d'interaction axés sur la sémantique.

Partenaires : Centre de recherche commun **Armines/LGI2P** de l'école des mines d'Alès, Inova Software, Inserm DISC (Département de l'Information Scientifique et de la Communication) en lien avec le CNRS-CCSD, LIRMM/université de Montpellier II, ISEM/université Montpellier II, Sanofi-Pasteur, Merial, Total.

CEA ToxNuc, ToxNuc3 et ToxNuc-e (1999-2013) – Partenaire.

Accompagnement de la Direction des Sciences du Vivant du CEA dans le développement de la communauté scientifique ToxNuc qui s'intéresse à la toxicologie nucléaire (environnementale), notamment au travers de la mise en place d'un système d'information collectif. L'objectif principal est la conception et le développement d'un système de partage de données, d'informations, et de connaissances, évolutif qui inclut des cartes de connaissances sur le collectif et sur les documents.

Partenaires : CEA, Centre de recherche LGI2P de l'école des mines d'Alès.

Plateformes collaboratives AvieSan[8] (2008-2013) – Partenaire, référent pour la plateforme IHP.

Mise en place d'un système d'information collaboratif dans le cadre de la restructuration des instituts thématiques multi organismes (ITMO) impliqués dans la recherche médicale en France, en particulier l'institut Technologies pour la Santé (ITS – https://its.aviesan.fr/), l'institut Immunologie, Hématologie, Pneumologie (IHP –https://ihp.aviesan.fr/) et l'institut Cancer (https://itcancer.aviesan.fr/). L'objectif est la conception et le développement d'un système de partage de données, d'informations, et de connaissances, évolutif.

[8] L'Alliance pour les sciences de la vie et de la santé (www.aviesan.fr/)

Partenaires : Inserm, Centre de recherche LGI2P de l'école des mines d'Alès.

Carnot (2008-2009) – Partenaire.

Mise en place d'un système d'information collectif pour le partage de données au sein de programme Carnot, en particulier – en pilote, sur le programme SensoMines. (http://www.communaute.carnot-mines.eu/). L'objectif est la conception et le développement d'un système de partage de données, d'informations, et de connaissances.

Partenaires : Institut Carnot M.I.N.E.S, Centre de recherche LGI2P de l'école des mines d'Alès.

Solunea/ Patterns (2008-2009) – Porteur de projet.

Assister l'auteur lors de la conception d'une formation, à partir de patrons pédagogiques et de scénarios.

Partenaires : Entreprise Solunea, Centre de recherche LGI2P de l'école des mines d'Alès.

Owanis consulting – Docalis (2007) – Porteur de projet.

Conception d'un environnement de composition semi-automatisée de cahier des charges liés à des outils d'aide à la décision et de "profilage".

Partenaires : Entreprise Owanis Consulting, Centre de recherche LGI2P de l'école des mines d'Alès.

INova (2006-2007) – Partenaire.

Assistance à la navigation et à la recherche d'information dans une base de données de brevets industriels.

Partenaires : Entreprise I-Nova, Centre de recherche LGI2P de l'école des mines d'Alès.

Nétia RIAM – SAVIC (2005-2006) **suite du projet MBox** (financement Région Languedoc Roussillon, 2003-2004) – Partenaire.

Indexation, Navigation dans une base de données de titres musicaux et Composition automatique de programmes musicaux.

Partenaires : Entreprise Nétia, Centre de recherche LGI2P de l'école des mines d'Alès.

L'école des mines d'Alès accompagne la création de nombreuses entreprises, via son incubateur (l'un des plus anciens de France). Les projets accompagnés sont suivis par des enseignants-chercheurs des laboratoires de recherche sur la thématique concernée. En ce qui me concerne, j'ai accompagné les projets suivants :

EKeet (2010-2012)

Conception et développement d'un produit innovant dans le domaine du Web sémantique (soumis à confidentialité). http://www.elkorado.com/

Soreha (2007-2008)

Mise en place d'un module de gestion de base de données d'échantillons biologiques. Depuis, le projet des créateurs a évolué et l'entreprise créée s'est spécialisée dans le développement d'applications mobiles. http://www.soreha.fr/

Hestia (2006-2007)

Réalisation d'un ordinateur simplifié pour personnes âgées : MAGUI. Premier prix du concours Lépine, 11 mai 2008.

Activités liées à l'enseignement

Au cours de ma carrière à l'EMA (école des mines d'Alès), je me suis souvent impliquée dans les réflexions concernant l'évolution des enseignements. En particulier lors du montage de la formation par alternance InfRes (Informatique et Réseaux) de l'EMA (j'étais alors responsable pédagogique de la formation de l'Institut EERIE qui a évolué pour devenir la formation INFRES), puis lors de la mutation de la formation initiale de l'EMA : alors que nous recrutions nos étudiants en fin de math-sup (Bac+1), depuis la rentrée 2010 nous les recrutons en fin de math spé (Bac+2). Cette réflexion portait à la fois sur le contenu des enseignements et sur leur forme : diminution de l'encadrement en présentiel pour laisser plus d'autonomie aux étudiants et favoriser le travail personnel. J'étais coordinatrice du groupe de réflexion concernant les enseignements liés à l'informatique. Cette mutation est d'ailleurs toujours à l'étude et je fais partie en 2013 du groupe de réflexion sur les innovations pédagogiques au sein de l'école. Enfin, lors de la création du département EMACS (Engineering and Management of Complex System – rentrée 2012), j'ai également participé à la définition des programmes et à l'organisation du département en modules.

Par ailleurs, dans la coordination de certains modules, j'ai été amenée à coordonner des équipes d'enseignants :

- pour le module GAGL (Génie Automatique et Génie Logiciel) des départements GSI (Génie des Systèmes d'Information) et ISP (Ingénierie des Systèmes de Production) – de 2003 à 2009 : coordination entre différents intervenants et mise en place d'un projet commun couvrant les différentes phases du génie logiciel.

- Pour le module d'initiation à l'algorithmique et à la programmation – 1999 à 2013 : coordination de 3 à 5 enseignants qui interviennent en parallèle auprès des étudiants de première année.

Concernant mes interventions en présentiel auprès des étudiants(en moyenne 130h eq.TD par an) les différents cours, TD, TP que j'ai dispensés sont détaillés dans le synoptique suivant (en fond bleuté, des enseignements qui n'ont plus lieu). A ces heures, il faut rajouter des encadrements pédagogiques divers détaillés plus bas.

Matière	Formation	Nb h eq.TD	Commentaires
Ontologies dans les systèmes d'information et Web Sémantique (2011-2013)	Formation initiale EMA – département EMACS (option Aide à la Décision)	15h / an	Dernière année de formation de l'EMA (10 à 15 étudiants).
Ontologies pour une recherche d'information efficace (2008 - 2012)	Master2 pro Communication et technologies numériques, EMA/Celsa	3h / an	Public d'origines très variées (journalisme, master de communication, sciences humaines et sociales…).
Initiation à l'Algorithmique et à la programmation (2003-2013) Pour certaines années, **coordination des enseignants**	Formation initiale ENSTIMA – 1ères années	60h / an	Public généraliste et débutant.
Initiation à l'Algorithmique et à la programmation (2010-2012)	Formation TIC&Santé	12h / an	Public très hétérogène visant une formation technique spécialisée pour le domaine bio-médical.
Initiation à l'algorithmique et à la programmation (2007-2013)	Formation Continue Diplomante à Distance	50h / an	Formation à distance de l'EMA qui englobe une séance de TP en présentiel, des séances de TD le soir via l'outil Breeze et un accompagnement sur un forum. Les étudiants sont issus du GEM (Groupement des Écoles de Mines) et intègrent différentes écoles en 3ème année (Albi, Douai, Alès)
Architecture des ordinateurs(2010-2012)	Formation Continue Diplomante à Distance	10h/an	Idem ci-dessus.
Génie Automatique et Génie Logiciel *Analyse du besoin, cahier des charges, UML, programmation JAVA*(2003-2009) ***Coordination du module***	3ème année ENSTIMA : option GSI (Génie des Systèmes d'Information), et ISP (Ingénierie des systèmes de Production)	30h / an	Un projet fil rouge était choisi chaque année, pour amener les étudiants du cahier des charges au développement d'une solution informatique et automatique (pilotage d'une maquette).
Système UNIX (1999-2006)	3ème année option informatique et automatique du cycle ingénieur de l'EMA	20h/an	Si les supports de cours sont toujours ceux que j'avais créés, je n'assure plus cet enseignement depuis 2006.
PROLOG (1998-1999)	3ème année option informatique du cycle ingénieur de l'EMA	8h	Initiation à Prolog, environ 20 étudiants
Ingénierie Multimédia (1999-2000)	DESS ARM - Auteur Rédacteur Multimédia	8/an	32 étudiants
Algorithmique et structure de données (1999-2001)	Première année EMA , Mastère NTIC et formation continue	30h par formation	De 14 à 40 élèves.
Bureautique (1999-2001)	1ère année EMA (1999-2000)	20h	Utilisation des logiciels Word et Excel et développement de macros en Visual Basic

En complément de ces enseignements, je participe chaque année à différents exercices pédagogiques qui sont proposés à nos étudiants :

- Synthèse bibliographique : accompagnement d'un groupe de 3 étudiants sur un sujet de recherche ciblé pendant 3 mois avec un suivi hebdomadaire.
- Tutorat de projets de fin d'étude : référent académique pour les stagiaires qui sont en stage en entreprise.
- Encadrement de missions de terrain (accompagnement d'un groupe de 3 étudiants qui sont immergés en entreprise pendant 5 semaines sur un sujet donné en lien avec une étude de marché, le développement d'une activité nouvelle ou encore l'optimisation d'une chaine de production) et de mini-missions (encadrement de 3 étudiants pendant 1 semaine dans un laboratoire de recherche sur des sujets proposés par les équipes de recherche). Ces exercices sont spécifiques à la dimension entrepreneuriale de la formation initiale et au label "ingénieur-entrepreneur" qui a été déposé par l'école en 1999.
- Participation à divers jurys et correction de rapports (Projets de Fin d'Etude – PFE, stages ouvriers, stages techniciens, etc.).
- Projets InfRes : réalisation de prototypes logiciels répondant à une demande des équipes de recherche du laboratoire.
- Mémoire de Recherche Appliqué du master CTN : problématisation apportant des éléments de compréhension d'une situation professionnelle, établissement d'un diagnostic et élaboration de préconisations.
- Encadrement de stages Master CTN ou de DESS.
- Tutorat académique des élèves de formation par apprentissage : chaque étudiant qui entre en formation par apprentissage Informatique et Réseaux (InfRes) est accompagné pendant la durée de la formation (3 ans) par un maître d'apprentissage qui est son référent en entreprise et un tuteur académique qui s'assure du bon déroulement de la formation, visite l'apprenti régulièrement sur son lieu de travail et participe à divers jurys, correction de rapports, etc.

Charges collectives

J'ai été élue, en 2007, représentante du personnel à la Commission Consultative Paritaire qui est la commission où se discutent les avancements des différents agents de l'école sous statut EPA (Etablissement Public Administratif). Soucieuse de participer activement aux différentes activités du centre de recherche, je m'implique à différents niveaux :

- Je ne reviens pas sur l'élaboration de cursus pédagogiques qui ont ouvert récemment à l'école des mines d'Alès (Master CTN, InfRes) car elle a été citée plus haut.
- Je ne reviens pas, non plus, sur mon implication dans les projets de l'incubateur (accompagnement et participation à des jurys de sélection).
- J'ai également participé ponctuellement à l'organisation logistique de rencontres scientifiques qui ont eu lieu sur le Site de Nîmes de l'EMA (Conférences Nîmes-TIC, Conférence RIC 2006, LFA 2007). J'ai également collaboré à l'organisation du PréForum AFIS et de la Finale du concours Robafis 2008.
- Enfin je contribue à l'organisation des conférences "Communication Sciences et Société" (C2S) que le centre de recherche LGI2P et l'EMA organise au profit des étudiants du département EMACS mais qui sont également ouvertes au grand public et plus particulièrement aux industriels et académiques de la région. Les thématiques abordés sont diverses. Je me suis plus particulièrement investie dans celle qui avait pour thème : "Le Web sémantique peut-il changer nos habitudes ? ".

Promouvoir les métiers scientifiques n'est pas toujours une tâche facile, d'autant plus lorsqu'ils se déclinent au féminin. J'ai, à deux reprises, animé des ateliers-rencontres auprès de lycéens lors des "Journée de découverte des métiers scientifiques et techniques au féminin" organisées par ConnaiSciences, réseau des cultures scientifiques en Languedoc-Roussillon, et soutenu par la Délégation Régionale aux Droits des Femmes et à l'Égalité. Dans le même ordre d'idée, je participe régulièrement à des ateliers-rencontres organisées dans notre établissement pour promouvoir les métiers de la recherche auprès des classes préparatoires et lycées de la région.

Enfin, j'ai participé à la conception (structure et graphisme) du site Web du LGI2P qui a été longtemps en ligne et qui vient d'être fondu dans le nouveau site de l'EMA.

Synthèse

Mes travaux de recherche, qui font l'objet de régulières publications dans des revues internationales, ont toujours été menés avec un grand souci de prendre en compte l'utilisateur final confronté à un système complexe. Au cœur de notre démarche, le contrôle de l'interaction homme/machine nécessite **l'analyse et l'exploitation sémantique des ressources** utilisées dans un **processus décisionnel**. Je suis convaincue que **recherche d'information**, **filtrage**, **visualisation** et **personnalisation** peuvent bénéficier des approches conceptuelles. C'est pourquoi je me suis

appliquée à proposer des **méthodes basées sur les ontologies de domaine** pour fournir à l'utilisateur une information pertinente et facile à interpréter, favorisant ainsi l'automatisation de tâches à haute valeur cognitive (processus d'apprentissage et de décision).

Les différentes actions que j'ai menées en tant que chargée de recherche au centre de recherche LGI2P et les responsabilités qui m'ont été confiées ont confirmé ma capacité à coordonner des personnes et à favoriser les échanges. C'est la raison pour laquelle, aujourd'hui, je souhaite mettre à profit ces qualités dans l'encadrement de travaux de recherche et que je suis candidate à l'habilitation à diriger les recherches.

Introduction

"Mais à quoi servent les ontologies ?" C'est une question qui vient spontanément dans la bouche des étudiants de Master après leur avoir présenté les différentes définitions des ontologies et leur avoir donné quelques exemples. Cette question est légitime, car si le monde académique vante depuis de nombreuses années les ontologies et leur capacité à supporter des solutions automatiques pour le traitement de la sémantique, l'inférence de nouvelles connaissances et, de façon plus générale pour l'ingénierie des connaissances, les applications associées sont longtemps restées à l'étude dans les laboratoires. Les raisons en sont multiples, nous le verrons par la suite, mais la principale est sans nul doute que pour disposer d'applications qui utilisent les ontologies, il faut disposer… d'ontologies ! De plus, pour bénéficier pleinement de leur apport dans les applications, les ontologies de domaine doivent être relativement *complètes*, c'est-à-dire couvrir de manière satisfaisante l'ensemble du domaine étudié. Or le processus de conception d'une ontologie de domaine requiert une longue phase de collecte de la connaissance auprès des experts (éventuellement aidés par des traitements d'analyse lexicale de corpus du domaine), associée avec une définition claire et non ambigüe des différents concepts, et de leur formalisation. Cette phase est d'autant plus longue qu'elle nécessite que ces experts s'accordent entre eux sur la définition des concepts qu'ils utilisent. Or dans certains domaines, la connaissance peut être répartie entre différentes disciplines dont les membres n'ont pas forcément la même vision de leur domaine et les mêmes terminologies, ce qui implique de longs débats avant d'obtenir un consensus. C'est particulièrement vrai dans le domaine des sciences de la vie, où se côtoient des phénomènes biologiques, chimiques, physiques, etc.

Cependant, poussées par l'augmentation des données numériques accessibles et par le besoin d'outils efficaces pour les analyser et les exploiter, certaines communautés se sont dotées, au fil des années, d'un modèle structuré de la connaissance de leur domaine. On peut citer en exemple la Gene Ontology (GO) qui recense l'ensemble des concepts qui peuvent être utilisés pour annoter les gènes, ou, dans des domaines plus généraux, WordNet qui tend à proposer une représentation générale du monde qui nous entoure. On entre alors dans un cercle vertueux : les ontologies étant disponibles, le transfert de différentes applications des laboratoires vers l'industrie, devient possible ce qui suscite de nouvelles idées dans d'autres domaines où il devient motivant de concevoir de nouvelles ontologies, ...

Les travaux décrits dans ce mémoire ont été conduits durant cette période de gestation, de naissance, puis de développement des applications reposant sur un modèle sémantique de la connaissance. Cet engouement a été d'autant plus fort qu'il fut poussé par la naissance du Web sémantique et des technologies qui lui sont associées et leur appropriation dans des applications industrielles. Au commencement de ma thèse (1997), la notion d'ontologie était connue d'un *cercle d'initiés* et la communauté scientifique dans son ensemble ignorait souvent jusqu'à son existence. Aujourd'hui, une dizaine d'année plus tard, des entreprises viennent nous solliciter parce qu'elles veulent "mettre des ontologies dans leurs applications pour les rendre plus performantes". Les langages de représentation des ontologies ont évolué, se sont standardisés, et de nombreuses équipes de recherche, de par le monde, imaginent des solutions nouvelles de raisonnement, d'inférence, de traitement. La prise de conscience de l'importance des aspects sémantiques a donc évolué dans le bon sens, mais une large action de diffusion reste encore nécessaire afin de démythifier les ontologies, quelles sont les applications qui peuvent en bénéficier et quelles sont les contraintes attenantes.

Ce mémoire apporte quelques éléments en ce sens et présente nos contributions à différentes phases du traitement de l'information dans un processus décisionnel, les ontologies servant de support à l'exploitation de ressources et à la personnalisation des résultats. Cependant, avant toute chose, il est nécessaire de distinguer les applications qui utilisent les axiomes contenus dans les ontologies et qui permettent de mettre en œuvre des techniques de raisonnement automatique et d'inférence (dérivation logique : logique de description ou terminologique, F-logic, etc.) des applications qui utilisent la structure formée par les concepts mis en relation les uns avec les autres (approches topologiques qui se rapprochent de la théorie des graphes). La hiérarchie de concepts supportant l'ontologie constitue alors un espace sémantique dans lequel les liens d'hyperonymie et d'hyponymie permettent, par exemple, de définir des mesures de similarité sémantique entre concepts. Les travaux présentés dans ce mémoire se revendiquent de ce second type d'approches.

Les travaux de recherche du laboratoire se positionnent résolument dans une démarche *d'automatisation cognitive* où les méthodes et techniques conçues sont au service d'un opérateur humain confronté à une situation de prise de décision, parfois dans un contexte de crise. Il est alors particulièrement important de trouver rapidement la bonne information. L'objectif des recherches présentées dans ce mémoire est clairement de lui faciliter l'analyse d'une situation, par exemple en l'assistant dans la recherche d'informations liées à son contexte. Il est alors nécessaire de *retrouver* les documents pertinents, qui doivent au préalable être sémantiquement *indexés*, et de lui proposer une *visualisation* de ces documents qui facilite leur analyse et l'aide dans sa prise de décision finale.

Une dizaine d'années en arrière, dans mes travaux de thèse, j'utilisais les ontologies comme support à la composition de documents à partir de *briques d'information*. Pour proposer une composition de ces nouveaux documents qui fasse sens dans un contexte précis, il était nécessaire d'identifier les briques pertinentes dans ce contexte et j'utilisais pour cela une indexation sémantique associée à chaque brique. Ainsi, même si en depuis le domaine d'application de mes travaux a évolué, trois mots semblent être au cœur de ma démarche scientifique : indexer, retrouver, visualiser. Ces trois phases sont des composantes incontournables des différentes applications sur lesquelles il nous a été donné de travailler. Elles feront l'objet des trois chapitres principaux de ce mémoire et seront présentées dans cet ordre, même si celui-ci ne reflète pas forcément l'ordre chronologique des publications y afférant. Les solutions imaginées privilégient la personnalisation et l'interaction avec l'utilisateur. Avant toute chose, il convient de préciser comment nous envisageons les ontologies, quelles sont les caractéristiques que nous utilisons et le contexte général de nos recherches. C'est ce que fait le premier chapitre.

De la nécessité d'indexer

Au cours de nos travaux, l'indexation a souvent représenté une composante incontournable mais délicate. Que ce soit pour améliorer la sélection de ressources pertinentes, ou pour mettre en relation les briques d'information les unes avec les autres afin de les organiser pour reproduire une logique du discours, les métadonnées sémantiques constituent un support efficace. Par le lien qu'elles établissent entre les ressources et la connaissance du domaine il devient possible d'appliquer certaines règles de composition, par exemple. Cependant disposer de ces métadonnées ne va pas de soi et cette phase d'indexation apparaît souvent comme un verrou difficile à lever. Si de nombreuses interrogations avaient été soulevées au cours de ma thèse, l'année suivante durant laquelle j'ai été ingénieur de recherche au LGI2P n'a fait que les confirmer. Dans le cadre d'un projet de recherche (Prompter – 1999-2000) avec un fournisseur de solutions logicielles pour la gestion et la diffusion de contenus audio-visuel (Nétia), nous avons proposé des solutions pour la génération de résumés de rencontres sportives. Pour ce faire, il était nécessaire d'indexer manuellement des segments de vidéos, en utilisant une ontologie de domaine (ici, celle concernant le football). L'ontologie a été construite, un environnement spécifique a été développé, mais la tâche d'indexation restait lourde et fastidieuse. Par la suite, en tant que chargée de recherche, j'ai poursuivi cette réflexion autour de

l'indexation et de l'outillage possible pour la rendre plus *légère* pour l'utilisateur. Le deuxième chapitre de ce mémoire positionne notre approche par rapport aux approches classiques dans ce domaine et détaille nos contributions. En particulier, nous avons expérimenté l'*indexation par propagation* à partir d'un ensemble d'entités[9] déjà indexées et représentées physiquement sur une carte de connaissance. En positionnant une nouvelle entité sur cette carte, elle peut s'indexer de façon automatique par propagation des indexations de ses plus proches voisins. La tâche d'indexation perd alors de son côté fastidieux pour devenir plus ludique et conviviale. Cependant cette approche présente encore des limites qui font que ce champ de recherche reste actif au sein de notre équipe.

La recherche d'information basée sur une ontologie de domaine

Disposant d'un ensemble d'entités indexées à l'aide de concepts issus d'une ontologie de domaine, des techniques de requêtage performantes en matière de *rappel* et de *précision* peuvent être mises en place. La littérature scientifique propose différentes approches dans ce sens. Cependant, comme nous l'avons précisé, nos travaux s'inscrivent dans un processus décisionnel et c'est sans doute ce qui constitue leur originalité : il ne s'agit pas tant de fournir un résultat à une requête d'utilisateur, il faut également être à même de lui justifier ce résultat. Nous nous attachons donc tout autant au calcul de pertinence qui permet d'ordonner un ensemble d'entités en fonction d'une requête qu'aux moyens de présenter et justifier cet ensemble de résultats. L'approche que nous avons choisie concerne la mise en application d'opérateurs d'agrégation issus du domaine de l'aide à la décision dans le calcul de pertinence. Nous utilisons également cette famille d'opérateurs lors de la reformulation de requêtes dans le cas d'un processus itératif d'interrogation de la base de données. Nous détaillons ces deux phases de la recherche d'information dans le troisième chapitre.

Visualisation et interaction se basant sur des ontologies de domaine

Que ce soit dans la phase d'indexation souvent fastidieuse ou lorsqu'il est confronté à de trop nombreux résultats à une de ses requêtes, l'opérateur humain a besoin de dispositifs techniques facilitant son appropriation des informations disponibles. Nous avons choisi de privilégier la visualisation et la manipulation de ces informations sur des cartes de connaissances interactives. Ces cartes utilisent les ontologies de domaine comme guide sémantique. La proximité physique des entités représentées tient compte de leur proximité sémantique, différentes informations peuvent être mises en relief (certaines relations, des regroupements, etc.). Des filtres particuliers permettre de personnaliser la présentation de l'information. Nos contributions dans ce domaine seront détaillées dans le quatrième chapitre.

Enfin un dernier chapitre présentera mes perspectives en matière de recherche et les pistes que j'envisage d'explorer à la lumière des résultats obtenus pendant ces dix années. Cependant cette liste n'est pas figée et je souhaite que la lecture de ces quelques pages suscite chez le lecteur attentif des interrogations et des discussions fructueuses qui pourraient se concrétiser par des collaborations futures…

[9] Nous utilisons le terme *entité* pour désigner ce que nous manipulons, représentons, etc. Ces entités peuvent représenter des documents, des enregistrements, des vidéos, des photos, des gènes, des personnes… en fonction des applications dédiées.

CHAPITRE I

«My companion smiled an enigmatical smile.

"That's just his little peculiarity," he said. "A good many people have wanted to know how he finds things out." »

Arthur Conan Doyle, "A Study in Scarlet"

Contexte et positionnement : modéliser la connaissance d'un domaine pour mieux exploiter les ressources disponibles

---

I.1. LES ENJEUX DE LA RECHERCHE D'INFORMATION 36

I.1.1. POURQUOI RECHERCHER, PARTAGER L'INFORMATION ? 37

I.1.2. QUELLE INFORMATION ? 39

I.1.3. QUI PARTAGE L'INFORMATION ? 40

I.1.4. UNE REQUETE CONCEPTUELLE POUR DES RESULTATS PERTINENTS 41

I.2. MODELISATION DE LA CONNAISSANCE D'UN DOMAINE 42

I.2.1. DONNEE, INFORMATION ET CONNAISSANCE 42

I.2.2. ONTOLOGIES : DEFINITION(S) 43

I.2.3. INGENIERIE ONTOLOGIQUE : CONCEPTION, FORMALISATION, METHODES ET LANGAGES 44

I.2.3.1. Spécification et évaluation 47

I.2.3.2. Conception et évolution 48

I.2.3.3. Diffusion : déploiement et mise en place de l'ontologie 49

I.2.3.4. Utilisation 49

I.2.3.5. Gestion et planification 49

I.2.3.6. Parmi les langages de représentation des ontologies un se détache : OWL 50

I.2.4. REDUCTION D'ONTOLOGIES AUX LIENS DE SPECIALISATION (ONTOLOGIES *LEGERES*) 51

I.3. SYNTHESE 52

Comme nous l'avons souligné en introduction de ce mémoire, le terme 'ontologie' commence à entrer dans le vocabulaire commun. On *en a entendu parler*, on sait qu'*il existe des applications qui les utilisent*, il paraît même que "*certains risquent leur chiffre d'affaire dessus*"… Ainsi, certains industriels, conscients de l'apport que pourraient représenter la prise en compte d'un modèle sémantique dans leurs applications, nous consultent pour les accompagner dans leur démarche.

Cependant, assez rapidement, il faut se rendre à l'évidence : la vision qu'on peut avoir des ontologies est très sommaire, assez imprécise et souvent chargée d'espérances, sinon vaines du moins fortement contraintes. C'est pourquoi, malgré les définitions que l'on peut trouver dans la littérature scientifique et sur la toile (Internet), il est important de clarifier dans ce mémoire ce que sont les ontologies de domaine, comment elles peuvent être utilisées dans des applications liées à l'extraction et à la gestion de connaissances et quelles en sont les limites actuelles. C'est l'objectif de ce chapitre introductif qui dresse le contexte de notre étude et l'illustre en confrontant les points de vue émanant des diverses applications sur lesquelles nous avons travaillé. Le parti pris ici n'est pas de faire une description exhaustive des ontologies, mais plutôt de présenter ce que nous en retenons dans notre approche.

Ce chapitre est en partie inspiré de supports de cours que j'ai écrits pour les étudiants du master CTN[10] et les étudiants de dernière année de la formation d'ingénieur généraliste[11] de l'École des mines d'Alès, ainsi que de rapports rédigés dans le cadre de projets industriels ou de propositions de projets soumises à l'ANR[12] pour lesquelles j'étais coordinateur. L'accent est porté ici sur le lien entre la connaissance d'un domaine et son exploitation un processus décisionnel, en particulier pour sélectionner les informations pertinentes.

I.1. Les enjeux de la recherche d'information[13]

La problématique de la recherche d'information est sans doute née avec les premiers écrits et au fil des siècles, des solutions ont été trouvées pour classer, cataloguer, archiver les documents et les retrouver rapidement. Certaines de ces solutions sont d'ailleurs encore à l'œuvre dans la plupart des bibliothèques ou dans les organismes de gestion d'archives. Cependant la révolution numérique les remet en partie en cause. Les informations (numériques) accessibles sont nombreuses, hétérogènes, distribuées, ce qui impose autant de contraintes à un système de gestion de connaissances qui se voudrait performant. Avec cet essor des informations accessibles et le développement de nouvelles techniques de communication, la recherche d'information constitue un enjeu majeur dans différents secteurs industriels et académiques, en particulier pour les acteurs impliqués dans la recherche scientifique, la veille technologique, l'innovation industrielle, la prise de décision, la gouvernance d'instituts publics de recherche, etc. Tout se passe comme si la performance et la capacité d'innovation des différentes équipes dépendait directement de leur capacité à trouver la bonne information, le plus rapidement possible.

Qu'il s'agisse de retrouver une information que l'on sait stockée dans une base de données – *retrouvage* pour J. Maniez (Maniez, 2002), ou de rechercher de nouvelles informations qui se rapportent à une problématique (via des moteurs de recherche sur Internet, par exemple) la qualité de la réponse est intimement liée à la formulation de la requête. Les ontologies peuvent servir de guide sémantique à l'expression de cette requête et de support actionnable lors du processus d'appariement entre cette requête et les documents disponibles. Cependant tout environnement de requêtage doit être conçu en fonction d'objectifs clairement établis en amont de sa conception. Ces objectifs pourront être fixés en fonction des réponses aux questions suivantes.

[10] Communication et Technologie Numérique, co-habilité par la Sorbonne et l'Ecole des Mines d'Alès

[11] Durant les deux dernières années de leur cursus, les étudiants ingénieur généralistes de l'école des Mines d'Alès sont accueillis dans des départements ; le site de Nîmes accueille les départements de Génie des Systèmes d'Information (GSI) et d'Ingénierie des Systèmes de Production (ISP).

[12] Plus précisément l'appel à projet ContInt (contenus et interactions) avec le projet DISTIL soumis en 2010 et le projet CIRCE soumis en 2011, projets non retenus pour financement mais dont certaines pistes ont été explorées.

[13] Cette section résulte en grande partie d'une réflexion menée avec Stéphanie Baar lors de son stage de fin d'étude de master Auteur Rédacteur Multimédia – ARM, que j'ai encadré (Baar, 2004).

I.1.1. Pourquoi rechercher, partager l'information ?

Les solutions logicielles vont être différentes en fonction de l'objectif de la recherche d'information. On peut être amené à rechercher une information dans différents contextes et avec plusieurs finalités, comme nous avons pu le constater au cours de ces dix dernières années où nous avons été confrontés à plusieurs demandes émanant d'industriels ou d'académiques. On peut les regrouper de la façon suivante :

- Veille technologique et innovation industrielle – *préoccupation majeure de la société Inova software*[14] *avec qui nous avons collaboré dans le cadre de la thèse de Jean Villerd (champ d'application sur la gestion d'une collection de brevets) et dans les deux projets soumis à l'ANR*. L'objectif consiste à réaliser un état de l'art des méthodes et techniques existantes, à connaître ses concurrents, leurs produits, les techniques qu'ils emploient, etc. Dans ce cas, la recherche concerne principalement des documents textuels (publications scientifiques, brevets, etc.). On peut intégrer au processus de requêtage une analyse du texte contenu dans les documents. Dans le cas où cette recherche documentaire serait transversale, il faudrait envisager de disposer de différentes ontologies de domaine ; c'est le cadre que nous nous proposions d'étudier dans le projet CIRCE[15] qui visait à trouver des informations complémentaires dans des domaines connexes et qui constitue une partie du sujet de thèse de Nicolas Fiorini.
- Recherche d'information pour une communauté scientifique – *il s'agit d'une demande fréquente des partenaires que nous accompagnons dans la gestion de leur communauté scientifique et que nous outillons via des plateformes collaboratives (le CEA avec le programme ToxNuc*[16] *ou l'Inserm avec les plateformes de plusieurs ITMO*[17] *de l'AvieSan*[18]*)*. Il s'agit, dans ce cas, de rechercher des entités qui peuvent être mises en relation les unes avec les autres. Pour la gestion de collectifs, par exemple, il peut être intéressant de trouver des personnes ou des laboratoires qui possèdent des compétences particulières (besoin exprimé par le CNRS-DIST[19] dans le projet ANR CIRCE). On peut aussi avoir besoin d'outils dédiés à la fouille de données spécifiques à un domaine. Par exemple dans les sciences du vivant, certains chercheurs peuvent être amenés à identifier des gènes qui interviennent dans les mêmes fonctions biologiques ou une même pathologie. Dans ce cas, la pertinence des résultats repose entièrement sur la qualité de l'indexation qui a été faite de ces gènes (avec des concepts de la Gene Ontologie) et du calcul de pertinence qui est appliqué.
- Prise de décision – *préoccupation majeure de la plupart des industriels, comme nous avons pu le constater par exemple lors du montage de la plateforme CRII en réponse à un appel à projet dans le cadre du grand emprunt, avec Sanofi, IBM et BioRad en particulier*. L'objectif, dans ce cas est plus large que la seule collecte d'information. Il s'agit de recueillir le plus de données possible sur un produit, un processus, une méthode afin de *valuer* un ensemble de critères et de croiser ces résultats pour faire émerger une recommandation en fonction de différentes stratégies ou d'argumenter des choix. Ici, s'il y

[14] http://www.inova-software.com/

[15] Réponse à l'appel à projet ContInt 2011 de l'ANR (*projet non sélectionné*)

[16] Voir http://www.toxnuc.org/ puis http://www.toxnuc-e.org/

[17] Instituts multi-organismes, en particulier les ITMO Cancer, IHP (Immunologie, Hématologie et Pneumologie) et ITS (Technologies pour la Santé).

[18] Plusieurs plateformes ont été mises en place pour répondre aux besoins des différents instituts : IHP (**Immunologie, Hématologie et Pneumologie**) https://ihp.aviesan.fr/, ITS (**Technologies pour la Santé**) https://its.aviesan.fr/, ITCancer https://itcancer.aviesan.fr/

[19] Centre National de la Recherche Scientifique – Direction de l'Information Scientifique et Technique

a eu indexation, elle doit tenir compte des jugements de valeurs qui peuvent être exprimés de différentes façons (notes, appréciation, etc.).

- Loisirs, culture – *principalement évoquée lors des collaborations avec la société Nétia*[20] *ou la société Atonis*[21]. Plusieurs des applications que nous avons conçues manipulaient des données multimédia (photos, musiques, vidéo) pour composer de nouveaux documents textuels ou faire des montages multimédia (albums, *playlists*, etc.). Dans ce dernier cas, l'indexation aura des spécificités par rapport à celle d'un document textuel. La caractérisation du contenu peut être plus subjective et dépendra très fortement de l'utilisation finale.

En réponse à ces différents objectifs, il faut proposer des solutions adaptées qui correspondent à des stratégies de recherche particulières. On peut notamment citer :

- recherche factuelle : recherche dans une base de données ;
- recherche documentaire : retrouver des documents à partir de leurs références ;
- recherche bibliographique : retrouver les références complètes à partir de documents ;
- recherche contextuelle : rechercher directement dans l'*information primaire* (l'information brute, telle que délivrée dans le document) des occurrences de mots ou groupes de mots (par des techniques de Traitement Automatique de la Langue Naturelle – TALN) ;
- recherche par navigation (de proche en proche) ; cette navigation peut être guidée par un modèle de la connaissance du domaine (e.g. suggestion d'articles similaires).

Si l'on considère des documents textuels, les ordinateurs ne pourront probablement jamais "comprendre" un texte. Aussi, ils ne pourront certainement jamais pallier complètement les non-dits car cela suppose une connaissance pragmatique. On cherche cependant à combler ce manque en concevant des modèles de la connaissance du domaine aussi précis que possible et en les associant au besoin avec une mémoire des expériences. "Comprendre un texte pour une machine, consiste alors à mettre en correspondance les informations sur les objets, les évènements, les faits décrits par ce texte, avec un modèle préétabli dont une représentation existe en machine. En fonction du degré de finesse du modèle, de la richesse des informations qu'il prend en compte, sa *compréhension* sera plus ou moins élaborée." (Lefèvre, 2000). Cette mise en correspondance se fait au moyen d'une indexation qui consiste à associer à l'*information primaire* une *information secondaire* (méta-information ou métadonnées) (Lefèvre, 2000). Les métadonnées "électroniques" sont des informations sur des objets numériques (image, texte, son) qui décrivent aussi bien le contenu de ces objets que leur gestion, leur structure, leur contexte ou les conditions d'accès (Richy & Després, 2007). En fonction du type de recherche que l'on souhaite mettre en œuvre, les informations secondaires utiles ne seront pas du même ordre. Ces informations permettent la réutilisation, le partage, l'échange de documents.

Pour effectuer une **recherche documentaire**, par exemple, l'utilisation de différents types de classification a été éprouvée. La classification décimale de *Dewey* (Dewey, 1876) est la plus ancienne. Elle a été complétée et perfectionnée par la classification décimale universelle (CDU) (Otlet & La Fontaine, 1896). Plus proche de nous (1995), le Dublin Core[22] est un schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques et d'établir leurs relations avec d'autres ressources. Etant établi comme norme internationale, il est employé par différents organismes inter-gouvernementaux (l'Organisation Mondiale de la Santé, par exemple).

Cependant le contenu des documents n'est pas forcément explicité dans ces différentes classifications or il peut être intéressant d'avoir plus de détail sur des partis pris, des solutions proposées, etc. On veut, dans ce cas, appliquer des méthodes de **recherche contextuelle**. Pour cela

[20] http://www.netia.net/

[21] http://www.atonis.com/, société accompagnée lors de leur création par l'incubateur de l'école des Mines d'Alès.

[22] http://dublincore.org/

on peut caractériser les concepts contenus dans un texte (unités conceptuelles) qui seront organisés suivant une syntaxe particulière (règles de combinaison de ces concepts) pour former l'indexation. Les professionnels de l'information, bibliothécaires et documentalistes, utilisent des langages documentaires, vocabulaires normalisés et structurés pour l'analyse documentaire et la description structurale et sémantique du document. Les langages documentaires sont toujours construits par rapport à un but précis. Ainsi, si les grandes classifications ont une visée universelle, les langages documentaires (que l'on peut assimiler à des thésaurus) sont limités à un domaine précis. Un thésaurus (e.g. RAMEAU[23]) est une liste alphabétique de concepts (termes préférentiels) et de renvois (termes non préférentiels) couvrant un domaine particulier (vocabulaire spécialisé) et présentant trois types de relations sémantiques : la relation d'équivalence, la relation hiérarchique et la relation associative. Ces thésaurus, on parle aussi de liste d'autorité, sont encore très utilisés dans les bibliothèques et certaines bases de données (e.g. Pascal[24]). L'interrogation peut être effectuée par mots séparés ou combinés (Chevillotte, 2005). On se rapproche de plus en plus de la notion d'ontologie.

I.1.2. Quelle information ?

Après s'être intéressé à l'objectif final de l'environnement de requêtage à concevoir, il est important de caractériser les types d'entités qui vont être manipulées par le système. De leur nature dépendront les traitements qui pourront leur être appliqués.

- **Médium** : les informations *primaires* à manipuler peuvent être stockées dans des textes (souvent), des enregistrements audio (e.g. *podcast*), musicaux, des vidéos ou des photos. Chaque médium peut nécessiter des traitements particuliers. Nous avons vu que sur le texte il est possible d'appliquer des techniques de TALN. Pour l'indexation de titres musicaux, nous avions envisagé de coupler l'indexation sémantique que nous proposions avec des données physiques extraites automatiquement par des techniques d'analyse du signal (tonalité, rythme, etc.). Certains chercheurs de l'équipe KID travaillent en analyse d'image sur la reconnaissance de visages ou de formes, la détection de contours ou de points d'intérêt, la recherche d'objets dans une image. Il est envisageable d'utiliser ces techniques d'analyse pour fournir une assistance lors de l'indexation.
- **Contenu** : si le contenu peut être directement analysé comme nous l'avons évoqué au point précédent, il peut également être représenté dans des métadonnées spécifiques (indexation) ou encore ne pas être rendu exploitable par une machine. Si l'on ne dispose pas d'une indexation, comment analyser automatiquement les entités manipulées (en un temps raisonnable) ? Cependant, même si l'indexation peut être assistée par des techniques d'analyse appropriées aux informations traitées, un gap sémantique reste à franchir entre l'analyse de l'entité et sa caractérisation au niveau conceptuel. Ainsi pour disposer d'une indexation conceptuelle performante, l'intervention humaine, même si elle peut être allégée, reste indispensable. En effet en l'absence de règles d'indexation préétablies un contrôle humain est nécessaire, la qualité des résultats produits par les systèmes qui utilisent par la suite l'indexation reposant presque intégralement sur la qualité de cette indexation.
- **Structure** : l'information peut être très structurée (article scientifique, rapports d'activité, CV), semi-structurée (blogs, forums) ou composée d'informations éparses (photos représentant plusieurs points de vue d'une même entité, par exemple). La structure des documents peut servir de guide à leur analyse. On peut, par exemple repérer dans l'ontologie de domaine les termes qui correspondent à certaines sections du document. En

[23] http://rameau.bnf.fr/informations/rameauenbref.htm (01 février 2013)

[24] Base de données de l'INIST (Institut de l'Information Scientifique et Technique) du CNRS http://www.inist.fr/spip.php?article22 (01 février 2013)

ce qui concerne l'information peu ou pas structurée, il est toujours possible d'appliquer une recherche plein texte en fonction de mots-clés, mais cette recherche atteint vite ses limites (en raison des problèmes rencontrés liés à l'homonymie, la synonymie, etc.).

- **Lieu** : enfin, les traitements envisagés seront très dépendants de la localisation des informations. Dans certains cas elle est très localisée (dans une base de données précise, sur un site particulier), parfois, au contraire elle est diffuse sur la toile.

De ces différentes caractéristiques dépendra l'indexation. Le principe de base de la recherche d'information consiste ensuite à apparier cette indexation avec une requête. Si ces différentes caractéristiques nécessitent parfois des traitements particuliers, il faut noter qu'en choisissant une indexation conceptuelle, i.e. associant des concepts de l'ontologie de domaine aux entités à indexer, on peut généraliser les traitements et favoriser le croisement même entre des données hétérogènes, dès lors qu'elles ont été indexées avec une même ontologie. Dans les travaux de thèse de Fabien Jalabert, c'est en effet l'ontologie de domaine (ou plutôt devrions-nous dire, les ontologies de domaine puisque nous utilisions UMLS et les modèles de connaissances associés) qui servait de support à la mise en relation des publications scientifiques issues du domaine biomédical (PubMed) et des gènes.

I.1.3. Qui partage l'information ?

Identifier et caractériser clairement les utilisateurs ou les groupes d'utilisateurs d'un environnement de gestion des connaissances permet de fournir des solutions adaptées. Il peut s'agir de personnes isolées qui font une recherche ponctuelle. Dans ce cas il faut déterminer si l'outil de requêtage à développer est destiné au grand public ou à des spécialistes du domaine, les solutions à mettre en œuvre étant différentes. On peut également vouloir outiller une communauté d'utilisateurs. Celle-ci peut être clairement identifiée par la liste nominative de ses membres (partenaires d'un projet, dépositaires d'un login/mot de passe sur un réseau particulier, etc.), ou au contraire être uniquement caractérisée par un centre d'intérêt commun (l'ensemble des internautes qui s'intéressent à la cuisine lozérienne) ; on parle alors de communauté virtuelle. En caractérisant ces groupes d'utilisateurs, on cherche à savoir si ces personnes partagent la même "culture", parlent le même langage et ont des habitudes de pratique communes afin de fournir des outils adaptés à cette communauté.

Nous avons été amenés à répondre aux attentes de ces deux types de communautés. Dans le cadre des plateformes collaboratives que nous proposons via la solution K-Hub[25], un compte est créé pour chaque membre de la communauté (e.g. ToxNuc, ITS ou IHP…) sur décision de la gouvernance du collectif. Pour y accéder, il doit s'identifier avec un login et un mot de passe. On connait donc assez précisément le profil des personnes inscrites et leurs attentes par rapport aux outils développés. Néanmoins, la conception de ces outils peut s'avérer être une tâche délicate car, même au sein de ces collectifs identifiés, on recense parfois des profils de chercheurs très différents qui partagent, certes, une thématique commune mais peuvent néanmoins avoir des besoins particuliers et des vocabulaires spécifiques. Ainsi dans le programme ToxNuc, on recense des physiciens, des biologistes, des professionnels de la santé, et la gouvernance de programmes nationaux, certains désignant la même chose avec des termes différents ou employant le même terme pour désigner des choses différentes. Pour favoriser les échanges et lever toute ambiguïté, les ontologies de domaine sont particulièrement adaptées. En ce qui concerne la conception et le développement de l'environnement OBIRS[26] (*Ontological Based Information Retrieval System*) nous avons imaginé une solution générique qui puisse être appliquée à n'importe quel domaine, pour peu que celui-ci dispose d'une ontologie associée et d'une base de données indexées avec les concepts de cette ontologie. L'application qui est accessible en ligne est dédiée à la recherche de gènes indexés

[25] Knowledge-Hub, http://www.khub.mines-ales.fr/

[26] http://www.ontotoolkit.mines-ales.fr/ObirsClient/

avec les concepts de la Gene Ontologie (GO). Nous ciblons donc l'ensemble des personnes intéressées directement ou indirectement par la génomique (biologistes, médecins, généticiens…). Cet outil a également été décliné[27] pour la fouille de publications scientifiques liées au cancer dans le cadre de l'ITMO Cancer de l'AviSan. Nous avons également proposé une extension de ce système (*CoLexIR*), conjuguant une recherche conceptuelle et une analyse lexicale, pour l'identification de passages pertinents dans des articles scientifiques (nous reviendrons dessus dans les chapitres III et IV).

I.1.4. Une requête conceptuelle pour des résultats pertinents

Dans tout système de recherche d'information, la pertinence des résultats obtenus est intimement liée à la formulation de la requête. Lors d'une recherche booléenne par simple mots-clés, ceux-ci doivent être écrits en parfaite conformité orthographique pour que les documents pertinents aient une chance d'être identifiés. Cependant, même si ces mots-clés sont écrits correctement, le processus d'appariement entre la requête et les documents se heurte à certains problèmes :

- de multilinguisme : en anglais US et en anglais de Grande-Bretagne, le centre ne s'écrit pas pareil (*center* et *centre*) ;
- de synonymie : vélo, bicyclette, cycle, deux-roues…
- d'antonymie : faut-il rechercher "distance" ou "similarité" ?
- de polysémie : le terme "rouge", par exemple va être compris différemment en fonction de son contexte, i.e. si on parle de vin, de signalisation routière, de peinture ou de politique.

Les ontologies étant composées de concepts désambiguïsés, elles permettent de pallier ces limites et d'exprimer une requête plus contrainte et donc plus *précise*. De plus, la connaissance étant modélisée dans cette ontologie, la recherche d'information peut bénéficier des relations entre les différents concepts qui y sont représentés. Leur hiérarchisation permet d'adapter le niveau de recherche souhaité. En utilisant les liens d'hyperonymie, on peut généraliser ou spécialiser sa requête en la reformulant à l'aide de termes plus ou moins génériques ou de termes apparentés (i.e. de sens voisin de celui recherché). Les termes sont post-coordonnés, c'est-à-dire que leur croisement est effectué au moment de la recherche (pas d'ordre préétabli). Dans une application dédiée aux plantes, par exemple, si on recherche un document qui traite des propriétés du thym commun et qu'aucun document ne correspond, on peut généraliser à la famille de plantes (Labiacées ou Lamiacées) et ainsi on pourra retourner des documents qui traitent du thym sauvage ou du thym serpolet. On évite ainsi les silences et l'utilisateur aura une information qui, même si elle ne correspond pas à cent pour-cent à sa requête a de grandes probabilités de l'intéresser. Ce sont ces propriétés qui font des ontologies un support particulièrement efficace de la recherche d'information.

Pour évaluer la qualité des résultats et de façon globale le système de requêtage, on peut rechercher le degré de pertinence de la réponse (à-propos, adéquation d'un document par rapport à une requête) ou l'efficacité du système (temps de réponse). Plusieurs mesures ont été imaginées afin d'évaluer de tels systèmes :

- Le **rappel** : proportion de documents pertinents retrouvés par rapport au nombre total de documents pertinents présents ;

- Le **silence** : proportion de documents pertinents non retrouvés ;

- La **précision** : proportion de documents pertinents par rapport à l'ensemble des documents proposés à l'utilisateur ;

- Le **bruit** : proportion de documents non pertinents parmi ceux proposés à l'utilisateur.

[27] http://obirs-cancer.mines-ales.fr/ObirsClient/public/Obirs/

Cependant dans le cas d'une recherche d'information répartie (par exemple sur le Web), ces mesures ne peuvent pas être appliquées et il faut alors mettre en place d'autres mesures : par exemple en utilisant une distance sémantique entre texte et requête. Là encore les ontologies peuvent servir de support à ce calcul de distance.

I.2. Modélisation de la connaissance d'un domaine

Le terme *ontologie* est apparu durant les dernières décennies dans le domaine informatique. Il fait référence à la notion *d'Ontologie* utilisée depuis des siècles en philosophie (c.f. plus bas). L'ontologie, en informatique, faisant appel aux notions de données numériques, de traitement de l'information et de façon plus globale à l'ingénierie des connaissances, il paraît important de préciser le sens de ces termes.

I.2.1. Donnée, information et connaissance

La **donnée** est "l'enregistrement, dans un code convenu, d'une observation, d'un objet ou d'un phénomène, d'une image, d'un son, d'un texte" (Afnor). Sitôt qu'une donnée est mise en contexte, associée avec d'autres données, elle devient porteuse de sens, on parle alors **d'information**.

« Assimiler Information et **Connaissance** serait une approximation lourde de confusion […] le terme connaissance étant beaucoup plus vaste que celui d'information, qui ne réunit que l'ensemble des connaissances mises en forme, c'est à dire explicitées » (Maniez, 2002). Il existe de nombreuses définitions de la connaissance (Jalabert, 2007), nous retiendrons celle de Nathalie Hernandez (Hernandez, 2005) pour qui "*l'information devient connaissance à partir du moment où elle sert de fondement à une inférence, au déclenchement d'un processus* […] *Les connaissances sont des informations actives.*"

L'**Ingénierie des Connaissances** (IC)**,** au sens de notion acquise sur un domaine, s'intéresse au codage de l'information et à son exploitation dans un contexte donné. Le codage des informations consiste à les enregistrer sous une forme précise, non ambiguë (pour éviter les problèmes de multilinguisme, de synonymie, etc.), normalisée et donc reproductible. L'objectif est d'améliorer l'exploitation de cette information par un opérateur humain ou par une machine, afin d'optimiser un processus applicatif (aide à la décision, analyse de risques, partage de connaissances, coordination…). Par exemple dans le domaine médical, les applications vont permettre de favoriser la communication entre spécialistes, entre établissements, avec le patient.

Pour John Sowa[28], père des graphes conceptuels (Sowa, 1984), "*l'*Ingénierie des Connaissances *est l'application des ontologies et de la logique à la construction de modèles programmables, dans un domaine, pour une application, un contexte, particuliers. Les dispositifs calculables, le domaine, et les enjeux en font une branche de l'ingénierie contrairement aux mathématiques (qui permettent des raisonnements sans domaine d'application, sur des ensembles non calculables (infinis) et sans but précis, autre que la satisfaction esthétique d'une abstraction élégante). Les sciences empiriques ont un domaine d'application et font des prévisions sur ce domaine mais se contentent souvent de satisfaire un simple besoin de connaissance. L'ingénierie, elle, utilise les sciences et les mathématiques pour résoudre des problèmes pratiques, en satisfaisant au mieux des contraintes budgétaires et de dates limites. Ainsi l'*IC *peut être définie comme une branche de l'ingénierie qui analyse la connaissance sur un sujet et la transforme en modèles calculables pour résoudre certains problèmes.*" Plus récemment, l'IC a été définie comme une branche de l'intelligence artificielle qui s'intéresse à l'acquisition, la modélisation, et le stockage des connaissances, leur consultation et le raisonnement automatique qui peut leur être associé (Hernandez, 2005).

[28] Propos traduits d'un mail diffusé sur la liste GC en 2002.

L'**ingénieur des connaissances**, quant à lui, est à la fois spécialiste du domaine et spécialiste des environnements informatisés liés à la gestion des connaissances. Il va intervenir dans les étapes amont de la recherche d'information, et en particulier durant l'indexation. Son action se situe principalement dans :

- le choix (et éventuellement la création) de la terminologie de référence en fonction de l'application, du domaine, du contexte, et des terminologies existantes.
- l'établissement de liens entre les termes de cette terminologie et les documents.

I.2.2. Ontologies : définition(s)

En philosophie, l'Ontologie[29] (*ontos* : être, *logos* : langage, raison) est l'étude de l'être en tant qu'être, de "l'essence" des entités qui nous entourent. Est-ce que les concepts existent en dehors de notre esprit ? Peut-on classifier les entités du monde ? Comment distinguer essence et existence, l'essence étant "quelque chose" en devenir, l'existence étant "quelque chose" du monde réel ? Autant de questions philosophiques auxquelles la science de *l'être* tente de répondre. L'IC s'est inspirée de ce courant philosophique lorsqu'elle a tenté de définir et de modéliser une certaine *connaissance* d'un domaine, pour pouvoir ensuite l'exploiter automatiquement via des systèmes informatisés de résolution de problèmes. En informatique, une définition semble faire l'unanimité : "**une ontologie est une spécification explicite (formelle) d'une conceptualisation (partagée)**" – (Gruber 1993) et page 1 de (Staab and Studer 2009). Elle est déclinée, complétée, enrichie et interprétée dans différents contextes, par différents auteurs de disciplines variées. Faire une synthèse de toutes ses variantes sortirait de l'objectif de ce chapitre introductif. Le lecteur peut s'en faire une idée plus précise en se reportant à la littérature (Abel et al., 2005; Chabert-Ranwez, 2000; Gomez-Pérez, Fernandez-Lopez, & Corcho, 2004; Hernandez, 2005; Jalabert, 2007). Pour une vision très détaillée des ontologies et de leur utilisation dans différents contextes on pourra se référer au "*Handbook on ontologies*" (Staab & Studer, 2009). Nous retiendrons ici que "*ce qui existe*" pour un système informatique est "*ce qui peut être représenté et manipulé par des programmes*".

La **conceptualisation** se réfère à un modèle abstrait d'un certain phénomène du monde reposant sur l'identification des concepts pertinents de ce phénomène (Hernandez, 2005). La spécification devant être analysée automatiquement, elle se doit d'être **formelle**. Enfin l'aspect **partage** est important à souligner puisque l'ontologie doit avant tout être un trait d'union entre les acteurs humains et les systèmes impliqués dans un même contexte (Chen & Mizoguchi, 1999). La connaissance modélisée dans une ontologie doit donc faire référence pour une communauté. Ce dernier point est central dans nos travaux puisque c'est ce lien de médiation que nous exploitons dans les ontologies, puisqu'il sous-tend l'interaction entre un (groupe d')utilisateur(s) et le système.

Différents types d'ontologies peuvent être considérés (ontologies de *tâche*, ontologies de *domaine*, ontologies *terminologiques*, etc.). De même (Hernandez, 2005) et (Gomez-Pérez et al., 2004) distinguent les ontologies "*légères*" (contenant essentiellement une taxonomie) et les ontologies "*lourdes*" (intégrant en plus des axiomes dans leur formalisation). Dans la suite nous allons conserver une approche très générale et nous verrons que dans nos travaux nous utilisons principalement la hiérarchie de concepts et le graphe qui peut lui être associé. Nous pouvons donc dire que nous utilisons principalement des ontologies dites *légères* même si elles peuvent contenir des milliers de concepts et de relations et, en conséquence, ne sont pas si légères que ça…

De façon très simplifiée, nous dirons qu'en informatique, une ontologie peut être vue comme une organisation hiérarchique des concepts pertinents associée à un ensemble de relations entre ces concepts. Elle doit être interprétable par une machine et si possible réutilisable. Elle doit également offrir des capacités logiques pour le raisonnement et les inférences (règles et axiomes). En plus du lien de spécialisation (noté *is-a* dans la suite) qui est obligatoirement présent pour former la

[29] Par convention on l'écrira, dans ce sens là, avec une majuscule.

hiérarchie de concepts, on peut spécifier dans une ontologie d'autres types de relations (certaines étant très dépendantes du domaine considéré) et des contraintes sur celles-ci, comme par exemple :

- des liens de composition : une voiture est composée d'un châssis, de roues… ; un composé chimique est constitué d'éléments ; un organe est composé de tissus, etc.
- des contraintes sur les définitions des concepts : une personne est une mère uniquement si elle est de sexe féminin et a au moins un enfant ;
- des contraintes d'intégrité : une personne possède un seul numéro de sécurité sociale ;
- des fonctions de calcul : pour bénéficier d'allocations familiales, il faut que le(s) enfant(s) soient âgés de moins de 18 ans ;
- des propriétés sur certaines relations : transitivité, symétrie, etc. : "partie-de" est une relation transitive, par exemple, "est marié à" est une relation symétrique…
- des propriétés par défaut : une voiture a quatre roues
- des relations inverses : la relation "est parent de" implique une relation inverse "est enfant de", de même "fait partie de" implique la relation inverse "inclut"
- des règles spécifiques au domaine considéré : en biologie, pour chaque récepteur qui active une fonction moléculaire, si cette fonction joue un rôle dans le fonctionnement de l'organisme, alors le récepteur joue le même rôle.

Pour Riichiro Mizoguchi (qui a beaucoup travaillé dans les systèmes d'apprentissage automatique et donc les ontologies pédagogiques), l'ingénierie ontologique est la recherche de concepts généraux, réutilisables, partageables et durables pour construire un modèle de connaissances capable d'aider des personnes à résoudre des problèmes (Mizoguchi & Bourdeau, 2004). Une ontologie est alors une représentation déclarative de connaissances ce qui permet au système de modifier son comportement en modifiant sa connaissance. L'ordinateur peut être un médiateur pour la dissémination de la connaissance entre les personnes travaillant dans des domaines variés.

Pour le projet M-Box *composer* avec la société Nétia, une ontologie de la musique a été conçue (nous y reviendrons plus loin). La Figure 1 présente un extrait de la hiérarchie de concepts conçue à l'époque (environ 400 concepts composaient la version entière).

I.2.3. Ingénierie ontologique : conception, formalisation, méthodes et langages

L'**ingénierie ontologique** concerne toutes les activités liées au processus de conception et de développement des ontologies, leur cycle de vie, les méthodes, les méthodologies, les outils et les langages.

Pendant longtemps, si l'on trouvait plusieurs méthodologies de conception dans la littérature, e.g. ascendante, descendante, centrifuge (Abel et al., 2005; Chabert-Ranwez, 2000), dans la pratique c'était souvent par tâtonnements que l'on procédait à la conception des ontologies de domaine, en partant de l'application à concevoir. Depuis, de nombreux travaux concernant l'ingénierie ontologique ont permis de préciser une méthodologie de conception. Sur le schéma du cycle de vie d'une ontologie (Abel et al., 2005) – Figure 2, on perçoit clairement l'amorce de la conception : la détection des besoins. Puis, un premier noyau est conçu par les spécialistes du domaine qui recensent les principaux concepts de leur domaine et les relations qui les unissent. Ce noyau est partagé (il peut aussi être intégré à une application), et une évolution progressive va permettre d'enrichir le modèle ontologique (par mise à jour, compléments, spécialisation, etc.).

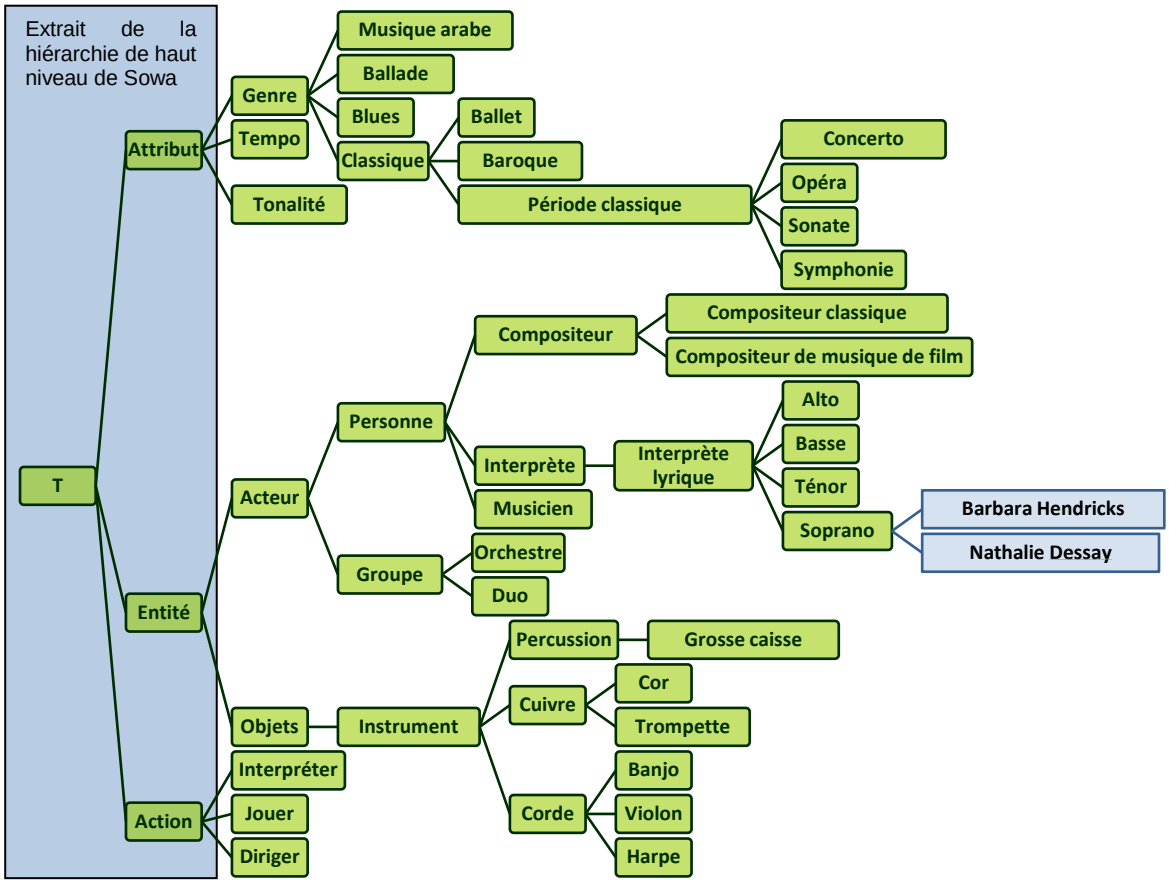


Figure 1 : Extrait d'une représentation graphique de la hiérarchie de concepts de l'ontologie de la musique.

Ici, seule la relation is-a est représentée. Les concepts les plus génériques sont à gauche et on peut remarquer deux instances à droite sur la figure.

Pour arriver le plus rapidement possible à une ontologie utilisée et actionnable, il faut se poser différentes questions :

- à qui va servir l'ontologie ? : un collectif de spécialistes, un environnement informatisé, des acteurs non spécialistes, etc.
- à quoi va-t-elle servir ? : communiquer entre humains, échanger des données entre applications, créer un système de requêtage, inférer des connaissances, etc.

Les réponses à ces deux questions vont jouer un rôle majeur dans la conception de l'ontologie et surtout dans le degré de formalisation attendu. En effet, même si certains préconisent l'utilisation d'une ontologie de haut niveau de laquelle dériveraient des ontologies spécifiques (c'est le cas de Nicolas Guarino et de John Sowa, par exemple), le type d'entités représentées est étroitement lié à l'application et c'est donc souvent en fonction d'une tâche particulière que l'ontologie va être conçue, à partir d'un mélange d'introspection et de créativité personnelle (Chabert-Ranwez, 2000). Un des meilleurs moyens de concevoir une ontologie de domaine est donc de travailler à partir de scénarios d'utilisation (Abel et al., 2005), tout en essayant de rester aussi générique que possible afin de conserver la propriété de réutilisabilité. Cette réflexion fait écho aux travaux de Tamara Munzner, qui insiste, lors de la conception et de la validation d'environnement de visualisation, sur la nécessité

de caractériser le domaine auquel le système va être appliqué avant de s'intéresser aux données à représenter. Différentes méthodes sont proposées : interviews d'experts, analyse des tâches des utilisateurs, etc. (Munzner, 2009). A partir de l'analyse des tâches, plusieurs niveaux de granularité sont distingués.

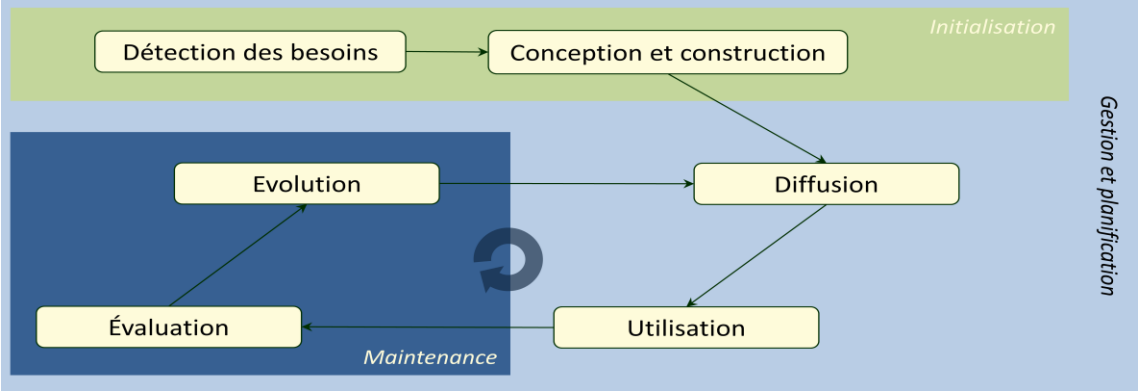


Figure 2 : Cycle de vie d'une ontologie

Citons à titre d'exemple une des ontologies les plus référencées dans le domaine des sciences du vivant : la Gene Ontology (GO). A l'origine, cette ontologie (qui, malgré son nom, n'en était pas formellement une) était un moyen d'unifier le vocabulaire d'un domaine où différentes disciplines se croisent : biologistes, médecins, physiciens, chimistes, etc. avec des vocabulaires souvent différents pour désigner les mêmes choses. A partir du moment où elle a servi de support pour des traitements automatisés, il a fallu désambiguïser, compléter, amender ce qui avait été fait pour atteindre une certaine cohérence. Aujourd'hui, c'est une des ontologies les plus utilisées au plan international, pour l'annotation de gènes (http://www.geneontology.org/) et elle compte plus de 35 000 concepts et 50 000 instances de relations.

Actuellement, la nécessité de disposer d'une méthodologie précise de conception des ontologies s'est imposée. Certaines étapes sont reconnues comme étant indispensables durant la conception. Ces étapes ont été définies dans METHONTOLOGY et dans Kactus, par exemple (Hernandez, 2005) :

- **Spécification** : description de l'ontologie attendue, ses utilisateurs, ses utilisations, le degré de formalisation souhaité.
- **Acquisition de connaissance** : identifier les termes de l'ontologie et leur définition. Cette étape consiste à collecter un ensemble d'informations sur le domaine[30] (analyse d'un corpus documentaire, interview de spécialistes, brainstorming, etc.).
- **Conceptualisation** : identifier les concepts, i.e. les descripteurs non ambigus qui sont caractéristiques du domaine. Former un modèle conceptuel. On va également identifier les relations qui les relient : is-a (*hyperonymie*, *hyponymie*), partOf (*méronymie*), CloseTo, etc.
- **Intégration** : recherche d'ontologies existantes qui peuvent être (partiellement) réutilisées. Des ontologies de haut niveau, par exemple, peuvent servir de base à la construction d'une ontologie. De même on peut être amené à réutiliser certaines parties d'une ontologie, dans un projet particulier. Pour cela des méthodes ont été proposées pour effectuer une segmentation de l'ontologie (Seidenberg & Rector, 2006) ou une extraction de sous-ontologie (V. Ranwez, Ranwez, & Janaqi, 2012).

[30] Certaines méthodes permettent d'amorcer la conception d'une ontologie par des analyses automatiques de documents du domaine. C'est l' "*ontology learning*". Cependant les techniques de TALN mises en œuvre sont très pointues et font l'objet d'un champ de recherche à part entière.

- **Implantation** : formaliser le modèle et le vocabulaire correspondant et coder l'ontologie dans un langage déterminé. Suivant le degré de formalisation, on va pouvoir également décrire les règles pour combiner les concepts et les relations (transitivité, inverse, …)
- **Evaluation** : vérifier et valider l'ontologie (cohérence, complétude, absence de répétitions, etc.)
- **Documentation** : rédaction de documents qui aident à comprendre l'ontologie ; pourquoi et comment elle a été conçue.

L'ontologie sert souvent de vecteur de communication dans le cadre d'un projet. Tout au long de la conception, il est nécessaire de respecter au mieux certaines propriétés (Gruber, 1993) :

- **Clarté** : toute ambiguïté doit être levée ;
- **Cohérence** : la cohérence des modèles formalisés doit être vérifiée ;
- **Extensibilité** : Il est indispensable de s'assurer de pouvoir compléter l'ontologie (la mettre à jour) sans (trop) remettre en cause la structure déjà construite ;
- **Indépendance** de tout formalisme : lors de la conception on se doit de rester aussi générique que possible et indépendant de tout langage de formalisation pour favoriser le partage.

Les sections suivantes[31] précisent les différentes étapes présentées dans la Figure 2 au regard de notre expérience personnelle. J'ai participé à la conception de trois ontologies dans des contextes et pour des applications différentes : l'ontologie du football pour le projet Prompter[32] avec la société Nétia, l'ontologie de la musique pour le projet MBox-*composer*[33] toujours avec la société Nétia, et plus récemment l'ontologie de la toxicologie nucléaire (avec comme partenaire principal le CEA). Dans les trois cas, nous avons conçu une ontologie *légère*, c'est-à-dire contenant essentiellement des relations de type *is-a* entre les concepts (c.f. section I.2.4).

I.2.3.1. Spécification et évaluation

Une méthodologie particulière doit être adoptée lors des phases de détection des besoins (en amont de la création de l'ontologie) et de validation (en aval). Il se peut, et c'est même souvent le cas, que les personnes qui ont identifié un besoin d'ontologie (lié à un besoin logiciel, par exemple) ne soient pas les spécialistes du domaine considéré. Il est donc primordial d'apporter un soin particulier au recueil de données et à l'étude de l'ergonomie et des usages auprès des experts. On peut procéder par des entretiens, des questionnaires, des sondages. A ce stade, il est possible de modéliser des scénarios.

De même, une fois l'ontologie créée et mise en œuvre dans une application, une évaluation rigoureuse doit être menée. Les mêmes techniques que pour la spécification peuvent être employées, auxquelles on va ajouter les retours d'utilisation.

En ce qui concerne le projet Prompter, nous voulions utiliser l'ontologie du football (*soccer* en anglais) comme support de notre outil d'indexation, afin de pouvoir qualifier des segments vidéo à l'aide de triplets en contraignant le vocabulaire. Puis, la même ontologie était utilisée lors de la recherche de segments vidéo pertinents pour la composition du résumé. Cette ontologie a été conçue et est disponible en ligne[34]. Cependant dans cette application nous avons été confrontés au problème des instances "éphémères". En effet, un joueur occupe un poste dans un contexte déterminé, sur une période précise, lors d'un match précis, etc. Pour un autre match l'ensemble des instances est à redéfinir. Pour le projet MBox Composer, l'objectif était également d'indexer des entités, des

[31] Largement inspiré de http://interstices.info/jcms/c_17672/ontologies-informatiques?part=3

[32] Composition automatique de résumés personnalisés de rencontres sportives. (Financement Région LR – 1999/2000).

[33] Composition automatique de *playlist* pour des chaines commerciales. (Financement Région LR – 2004/2005).

[34] http://www.daml.org/ontologies/273

enregistrements de titres musicaux, pour composer ensuite une *playlist* qui ait une certaine "couleur" musicale. Enfin pour l'ontologie ToxNuc, l'objectif est de mettre à la disposition de la communauté scientifique une ontologie du domaine de la toxicologie nucléaire environnementale pour servir de support à l'indexation de documents textuels et à la recherche d'information. Mais cette ontologie permettrait également de clarifier le vocabulaire de la communauté en unifiant les termes utilisés dans chaque sous-domaine considéré.

Concernant l'évaluation de ces ontologies, elle a été faite, dans les deux premiers projets, en même temps que l'évaluation des outils qui les utilisaient. Pour ce qui est de l'ontologie de la "Toxicologie Nucléaire environnementale", nous allons l'intégrer prochainement aux outils de la plateforme *ToxNuc*[35].

I.2.3.2. Conception et évolution

L'**acquisition des connaissances** doit être menée en collaboration avec des experts du domaine. Elle peut être assistée par des techniques de traitement de langage naturel qui permettent d'identifier les lexiques associés à certains concepts ou par le partage d'information via des plateformes collaboratives (à partir de *folksonomies*). Cette acquisition sera effective tout au long du projet pour assurer une évolution et une mise à jour du modèle de connaissance.

Les interactions avec les experts peuvent être surprenantes. Pour Prompter, il s'agissait de DJ qui ont chacun leur propre classification empirique. Or une première version de l'ontologie de la musique avait été proposée en catégorisant les instruments de musique, les tonalités, les types d'œuvres, les genres musicaux. Cette collecte des concepts du domaine avait été faite à partir de différents documents : encyclopédies, critiques musicales, etc. Or les DJ ne se reconnaissaient pas dans cette classification et finalement, la partie de l'ontologie qui a été le plus utilisée dans l'outil concernait les *moods*, les humeurs que peuvent traduire la musique. On y retrouvait des concepts assez surprenants pour qualifier un morceau de musique : *Gout* (*Doux*, *Suave*, *Epicé*…), *Onde* (*Sombre*, *Pénombre*, *Scintillant*…), *Attitude* (*Respectueuse*, *Irrévérente*, *Conflictuelle*) ou *Sensation* (*Organique*, *Cérébrale*). On voit donc bien, ici, l'importance du dialogue entre les personnes qui vont formaliser l'ontologie et les experts. De la même façon, pour l'ontologie de la toxicologie nucléaire environnementale, de longs entretiens[36] avec des experts du domaine ont permis d'aboutir à la définition d'une terminologie.

La **conceptualisation** et la modélisation des connaissances font également appel à des méthodologies particulières. En effet les connaissances sont souvent récoltées à partir de 'termes' et il faut parvenir à une définition conceptuelle (entretien avec les experts). On peut, par exemple, utiliser des méta-ontologies (ontologies de haut niveau) qui sont spécialisées progressivement – c'est l'approche que nous avions choisie pour l'ontologie de la musique (c.f. Figure 1) en utilisant la hiérarchie de concepts proposée dans (Sowa, 1984). On peut également s'intéresser aux autres ressources disponibles (ontologies de domaines connexes, etc.) et, en utilisant l'alignement ontologique ou des mécanismes de traduction, établir des passerelles entre l'ontologie que l'on est en train de concevoir et les ressources (pré)existantes. Par exemple dans le cadre de l'ontologie ToxNuc, il sera certainement nécessaire de faire le lien avec CHEBI[37] pour ce qui concerne les descriptions des éléments chimiques. En effet, même s'il existe des spécificités propres au domaine de la toxicologie nucléaire, la description des atomes, par exemple est universelle.

[35] http://www.toxcea.org/

[36] Cette étape préliminaire à la conception de l'ontologie a été réalisée par Emilie Montméjean, que je n'ai pas encadrée à l'époque, puis reprise par Imane Anoir dans le cadre de sa thèse. Ces travaux ont été repris récemment par Véronique Gachet lors de son stage de Master CTN (2011), que j'ai co-encadré. L'ontologie produite a dû être "nettoyée" afin de permettre l'indexation de documents de la plateforme ToxNuc (sujet de Post-Doc de François-Elie Calvier, que j'ai co-encadré – 2012/2013).

[37] http://www.ebi.ac.uk/chebi/init.do

La **formalisation** et l'**implantation** font appel, quant à elles, à des langages et formalismes identifiés : logique de description, analyse de concepts formels (FCA), formalismes du Web sémantique (DAML+OIL, OWL, RDF/S), graphes conceptuels… En ce qui nous concerne, nous avons utilisé DAML+OIL (ancêtre de OWL) pour les ontologies du football et de la musique et OWL pour celle de la toxicologie nucléaire. Ce choix était motivé par le fait qu'OWL étant un standard, d'autres applications pourraient bénéficier de ces ontologies.

Afin d'obtenir et de maintenir un consensus sur les choix de représentation et de conceptualisation faits dans l'ontologie, on pourra faire appel à des "collecticiels" (*groupware*) ainsi qu'à des outils de gestion de points de vue, terminologies, langues ou jargons.

L'**évolution** d'une ontologie est particulièrement délicate. En effet, une ontologie décrit des "faits" du monde, des affirmations, des primitives… Des mécanismes de raisonnement, de calculs divers (pertinence, distances sémantiques, inférence) sont basés sur sa structure. Toute évolution de l'ontologie est donc susceptible d'affecter ces calculs et, par conséquent, d'altérer le fonctionnement des applications qui l'utilisent. La maintenance des ontologies soulève de nombreux problèmes et ouvre de nombreuses perspectives de recherche concernant, par exemple, la gestion des versions, la ré-ingénierie, la propagation des changements, etc.

I.2.3.3. Diffusion : déploiement et mise en place de l'ontologie

La diffusion (et le partage) de l'ontologie au sein d'une communauté suppose que les formalismes choisis aient été aussi compatibles que possible avec les technologies utilisées par la communauté. Pour une application Web, par exemple, on pourra utiliser des architectures pair à pair ou des architectures distribuées pour le partage de fichiers. Pour l'intégration d'application, des architectures de service web peuvent être une solution. Cependant, quelle que soit l'architecture choisie, la distribution des ressources (données, modèles, applications, utilisateurs) et leur hétérogénéité (syntaxes, sémantiques, protocoles, contextes…) posent des problèmes de recherche sur l'interopérabilité (alignement, médiation) et le passage à l'échelle (optimisation d'inférences, propagation des requêtes, composition de services…)

I.2.3.4. Utilisation

L'utilisation d'une ontologie peut être profitable pour différents types d'activités, telles que :

- l'indexation sémantique de ressources (traitement de la langue, annotation…) ;
- recherche d'information et résolution de requête (projection de graphes avec contraintes) ;
- déduction de connaissances (moteurs d'inférence à base de règles) ;
- analyse de gros volumes de documents (*clustering*, veille…) ;
- navigation assistée dans un corpus / visualisation ;
- aide à la décision (pouvant faire intervenir différentes techniques listées ci-dessus).

Outre des techniques calculatoires pointues, ces applications posent souvent le problème de l'interaction avec l'utilisateur. Cette interaction suppose des solutions logicielles adaptées en matière d'ergonomie, et de représentation sémantique pour l'utilisateur. Les ontologies peuvent servir de support à des cartes sémantiques et favoriser les interfaces dynamiques (à partir d'inférences, par exemple, ou par filtrage de données). Cependant, si elle peut susciter des usages nouveaux, l'utilisation d'ontologie pose également de nombreux problèmes, en particulier concernant le passage à l'échelle (la taille des données impose des calculs optimisés) et la visualisation.

I.2.3.5. Gestion et planification

Il est important d'avoir un travail de suivi et une politique globale pour détecter ou déclencher, préparer et évaluer les itérations du cycle de vie de l'ontologie et s'assurer que l'on reste dans le

cercle vertueux des systèmes d'information (où se succèdent contribution, utilisation, création). Cela constitue les activités de gestion et planification qui ont lieu tout au long de ce cycle (c.f. Figure 2).

Enfin il est à noter que concernant la gestion des ontologies, il n'y a pas de spécificité particulière liée à la langue. En effet, l'ontologie s'intéressant à la définition de concepts, les entités du monde réel qui sont représentées sont uniques. C'est leur appellation, leur label, qui change d'une langue à une autre, mais la structure conceptuelle (concepts et relations) reste inchangée.

I.2.3.6. Parmi les langages de représentation des ontologies un se détache : OWL

C'est également en fonction de la tâche à effectuer que l'on va déterminer le formalisme de représentation des ontologies. Un des tout premiers formalismes a été initié par John Sowa en 1984 : les graphes conceptuels (Sowa, 1984). Ce formalisme équivalent d'une logique de premier ordre est facile à lire pour un humain et peut être associé à des méthodes mathématiques puissantes en termes de raisonnement. Cependant, ces traitements automatiques peuvent être lourds à mettre en place.

Avec l'essor du Web et l'évolution des technologies associées d'autres langages ont été proposés : SHOE[38] (Simple HTML Ontology Extension), LOOM[39], PowerLoom[40], KIF[41], OIL puis DAML+OIL[42]. Aujourd'hui un standard a émergé de ces différentes approches et il est conseillé d'utiliser OWL pour modéliser les ontologies.

OWL[43] est une surcouche de RDF/RDF-S qui a été conçue pour favoriser les traitements sémantiques des données. Dérivé du DAML+OIL, OWL est devenu un standard préconisé par le W3C, et est désormais utilisé tant pour les applications indépendantes que pour les applications Web. Il enrichit le modèle RDF-S en formalisant un vocabulaire pour la description d'ontologies complexes.

RDF-S permet la définition d'un vocabulaire simple en s'appuyant sur un nombre minimum de notions et de propriétés, dont :

– les notions de classe, ressource, littéral ;

– les propriétés de sous-classe, de sous-propriété, de champ de valeur, de domaine d'application.

OWL est un langage beaucoup plus riche qui, aux notions définies par RDF(S), ajoute les propriétés de classe équivalente, de propriété équivalente, d'identité de deux ressources, de différence de deux ressources, de contraire, de symétrie, de transitivité, de cardinalité, etc., permettant de définir des rapports complexes entre des ressources (Richy & Després, 2007).

OWL est composé de trois sous-langages d'expressivité croissante :

- **OWL Lite** répond à des besoins de hiérarchie de classification et des contraintes simples. Par exemple concernant les contraintes de cardinalité il ne permet que les valeurs 0 et 1.
- **OWL DL** (Logique de Description) : répond à des besoins d'expressivité maximale couplée à la complétude de calcul. Il permet d'aider au raisonnement sur une base de connaissance. Il a un pouvoir d'expression supérieur, et garantit la *complétude* (toutes les conclusions sont calculables) et la *décidabilité* (tous les calculs se font en un temps fini). OWL DL inclut toutes les constructions de OWL mais avec certaines restrictions.

[38] http://www.cs.umd.edu/projects/plus/SHOE/

[39] http://www.isi.edu/isd/LOOM/

[40] http://www.isi.edu/isd/LOOM/PowerLoom/index.html

[41] http://ksl.stanford.edu/knowledge-sharing/papers/

[42] http://www.daml.org/

[43] http://www.w3.org/TR/owl-features/

Par exemple la séparation des types : une classe peut être sous-classe de plusieurs classes (multi-héritage) mais elle ne peut pas être instance d'une autre classe.

- **OWL Full** : expressivité maximale, mais ne garantit pas la complétude et la décidabilité.

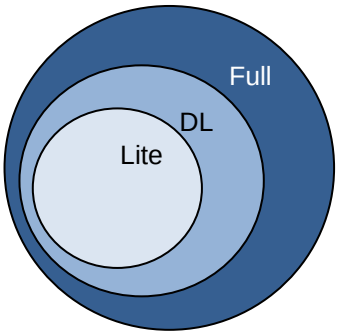


Figure 3 : Imbrication des différents niveaux de langage d'OWL

L'imbrication des trois sous-langages OWL (Figure 3) garantit que :

- toute ontologie OWL Lite valide est une ontologie OWL DL valide ;
- toute ontologie OWL DL valide est une ontologie OWL Full valide ;
- toute conclusion valide de OWL Lite est une conclusion valide de OWL DL ;
- et toute conclusion valide de OWL DL est une conclusion valide de OWL Full.

La structure détaillée du langage OWL, tel que décrit par le W3C est disponible à http://www.w3.org/TR/owl2-primer/ et une synthèse est présentée dans (Mbao, 2007).

I.2.4. Réduction d'ontologies aux liens de spécialisation (ontologies *légères*)

Si les axiomes définis dans les ontologies peuvent permettre de raisonner et d'inférer de nouvelles connaissances, dans certains domaines et pour des cadres applicatifs bien définis, l'ontologie peut être réduite à la structure de graphe que sous-tendent les seules relation transitives de spécialisation (*is_a*) et de composition (*part_of*) ; ce sont des ontologies dites légères. Une telle réduction est fréquemment appliquée pour le calcul de similarité sémantique et en recherche d'information où l'expansion de la requête n'est profitable que lorsqu'elle utilise ces deux types de relation (c.f. chapitre 3).

Une ontologie O peut être considérée comme un graphe sémantique enraciné G = (V, E) où V est un ensemble des nœuds correspondant aux concepts de O et E un ensemble d'arcs étiquetés correspondant à leurs relations sémantiques[44]. Si l'on restreint E à l'ensemble des relations transitives correspondant aux liens de spécialisation (hyponymie) et de composition (méronymie), on obtient un sous-graphe $G_r = (V_r, E_r)$ qui a la particularité d'être sans circuit et orienté. Cette réduction est certes une vision partielle de la connaissance du domaine, mais elle offre une structure sur laquelle des algorithmes efficaces peuvent être développés. Ceci nous a permis de proposer une nouvelle mesure de distance sémantique (S. Ranwez, Ranwez, Villerd, & Crampes, 2006) – c.f. chapitre 3, section 4 ; de définir une méthode et un outil d'extraction de sous-ontologies, *ontoFocus* (V. Ranwez, Ranwez, & Janaqi, 2009; V. Ranwez, Ranwez, et al., 2012) – c.f. chapitre 3, section 5 ; et de concevoir des outils de recherche d'information : OBIRS, OBIRS-feedback et CoLexIR (S. Ranwez et al., 2013; S. Ranwez, Ranwez, Sy, Montmain, & Crampes, 2010; M. F. Sy, Ranwez, Montmain, & Ranwez, 2012; M.-F. Sy et al., 2011) – chapitre 3.

[44] Cette formalisation a été proposée dans (V. Ranwez et al., 2009) et développée par Sébastien Harispe (rapport interne).

I.3. Synthèse

A la fin des années 90, la croissance exponentielle de la quantité de données numériques accessibles a conduit à la définition de nouveaux modèles de représentation de l'information et à de nouvelles stratégies de requêtage. Les ontologies, offrant un cadre formel à la représentation des connaissances d'un domaine, se sont peu à peu imposées comme solution privilégiée à la désambigüisation et comme support efficace en recherche d'information. Elles atteignent aujourd'hui au degré de maturité suffisant pour que les solutions théoriques imaginées dans le cadre académique trouvent des applications dans le monde industriel. Cependant, leur mise en œuvre nécessite une lourde tâche préalable de modélisation.

Dans nos travaux, nous utilisons une réduction des ontologies de domaine, en travaillant essentiellement sur la structure de graphe enraciné, orienté et acyclique formé par les concepts du domaine et leurs relations transitives d'hyponymie et de méronymie. Les applications que nous privilégions concernent la recherche d'information (c.f. chapitre 3) dont le socle est l'indexation des entités manipulées. C'est pourquoi nous détaillons dans le chapitre suivant, nos contributions et nos perspectives en matière d'indexation.

CHAPITRE II

« How can we automatically expand annotations of certain already manually annotated multimedia objects to other objects that have the same or similar attributes without presenting them to the user for manual annotation? »

Michalis Lazaridis et al. 2012.

Les ontologies comme support à l'indexation conceptuelle

II.1. QUALITES ATTENDUES D'UNE INDEXATION 56

II.1.1. EXPRESSIVITE ET FIABILITE 56

II.1.2. TECHNICITE 57

II.1.3. EXPLOITABILITE 57

II.2. ENTITES A INDEXER ET CARACTERISTIQUES A PRENDRE EN COMPTE. 58

II.2.1. CATALOGAGE OU DESCRIPTION D'EDITION 59

II.2.2. L'INDEXATION DU CONTENU POUR SUPPLEER A UNE RECHERCHE CONTEXTUELLE..... 59

II.3. L'INDEXATION CONCEPTUELLE AU CŒUR DE NOTRE DEMARCHE 60

II.3.1. ATOUTS DE L'INDEXATION CONCEPTUELLE..... 60

II.3.2. LIMITES IDENTIFIEES DE L'INDEXATION CONCEPTUELLE..... 61

II.4. L'INDEXATION PAR PROPAGATION : PRINCIPE ET MISE EN ŒUVRE 63

II.4.1. PRINCIPE DE L'INDEXATION PAR PROPAGATION 64

II.4.2. MISE EN ŒUVRE DE L'INDEXATION PAR PROPAGATION 64

II.4.3. RESULTATS ET LIMITE DE L'APPROCHE 66

II.5. L'INDEXATION CONCEPTUELLE PAR PROPAGATION..... 67

II.6. SYNTHESE..... 68

Les index qui peuvent être associés à des éléments d'information correspondent à une information secondaire qui facilite leur accès. L'indexation est ainsi définie comme "une opération qui consiste à décrire et à caractériser un document à l'aide de représentation 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" (United Nations International Scientific Information System (UNISIST), 1975). Les termes caractérisant l'entité indexée sont appelés les *descripteurs* et l'ensemble de ces descripteurs constitue le langage d'indexation (Baziz, 2005).

Si l'on s'accorde à penser que la qualité de la recherche d'information constitue un enjeu majeur dans nos sociétés, tant du point de vue industriel qu'académique et gouvernemental, elle ne peut être dissociée de l'exigence de qualité d'indexation des informations manipulées. Après plusieurs années de réflexion sur cette problématique, il apparaît que les ontologies de domaine peuvent intervenir efficacement dans les systèmes de recherche d'information pour améliorer l'indexation en permettant d'associer des concepts à un document, levant par-là l'ambiguïté liée à une indexation par mots clés. Dans ce cas les descripteurs sont des concepts, le langage d'indexation correspondant à l'ontologie de domaine dont ils sont issus. Cependant, si on la veut de qualité, cette indexation conceptuelle nécessite une intervention humaine qui est souvent fastidieuse. Des méthodes et outils doivent être mis en place pour alléger cette tâche. Ce chapitre retrace notre réflexion sur cette problématique en rappelant les qualités attendues d'une indexation, puis expose nos propositions et ouvre des perspectives en matière d'indexation par propagation et d'indexation collaborative.

II.1. Qualités attendues d'une indexation

Nos premiers travaux sur l'indexation ont été menés dans le contexte de la composition automatique de documents à partir de documents virtuels personnalisables (DVP) avec comme domaine d'application principal les environnements pédagogiques. Un DVP est un ensemble d'éléments numériques, autrement appelés briques d'information, associés à des mécanismes d'identification, de sélection et d'assemblage sous contrainte(s) pour produire un document composite adapté aux besoins d'un utilisateur ou d'un groupe d'utilisateurs dans un contexte particulier (Iksal & Garlatti, 2001; S. Ranwez & Crampes, 1999a). Or pour pouvoir mettre en œuvre des algorithmes de composition, le système informatique doit avoir une représentation interne du contenu des briques d'information manipulées. Nous avons caractérisé les qualités attendues d'une indexation dans (Crampes, Ranwez, Plantié, & Vaudry, 2003), en fonction de trois catégories qui sont synthétisées ici et complétées car elles fixent les objectifs qui sous-tendent nos travaux.

II.1.1. Expressivité et fiabilité

Cette catégorie englobe à la fois la capacité d'expression de l'indexation en regard de l'entité indexée et sa cohérence sémantique. Pour cela elle doit répondre à des critères de :

- **Fidélité** et **objectivité** : représenter strictement le contenu de l'information primaire sans y adjoindre d'interprétation subjective. Pour être fidèle, l'indexation ne doit pas refléter le point de vue de l'indexeur. Or, dans certains cas d'application (comme pour le partage de documents scientifiques dans un collectif, par exemple), l'avis d'expert sur un document revêt un intérêt certain, particulièrement si cet expert fait autorité dans son domaine. Dans ce cas, les métadonnées doivent pouvoir être tracées : qui est l'indexeur ? Comment justifie-t-il son interprétation ? Dans quel contexte a été faite l'indexation ? Quelle est l'utilisation envisagée de cette indexation ? Etc.
- **Complétude** : représenter l'ensemble du contenu et donc se détacher d'une application particulière pour essayer de représenter toutes les facettes d'une entité. Dans les faits, ce critère est difficile à respecter car l'indexation est souvent réalisée pour une application précise et est donc fortement dépendante de cette application.
- **Consistance** : l'indexation ne doit pas contenir de contradictions logiques ni de contradictions sémantiques. D'un point de vue logique, une affirmation et son contraire ne doivent pas coexister. Sur le plan sémantique, l'indexation doit respecter les définitions (règles) du domaine auquel le document se réfère. On perçoit bien l'apport que peuvent représenter les ontologies sur ce point.

- **Précision** : l'indexation doit éviter les ambiguïtés qui peuvent entraîner des mauvaises interprétations sur l'utilisation attendue du document et le sens que l'utilisateur peut être amené à extraire.
- **Evaluabilité** : une indexation doit pouvoir donner ses propres limites afin qu'un utilisateur, ou un outil, puisse juger de ce qu'il peut attendre du document, mais aussi de l'indexation elle-même : est-elle fidèle ? Complète ? etc.

II.1.2. Technicité

Le contexte dans lequel nous nous situons est clairement celui du traitement automatique de l'information. Nous nous intéressons particulièrement aux outils de stockage, d'accès, d'échange, de fusion, de calculs sémantiques, etc. Les contraintes techniques liées à cette automatisation représentent donc des qualités incontournables.

- **Accessibilité** et **calculabilité** : qu'elle soit à usage local (corpus spécialisé) ou dans un milieu plus ouvert (le Web), l'indexation doit pouvoir être facilement accessible à des systèmes informatiques. De plus la formalisation choisie doit favoriser l'exploitation automatique de ces indexations, en particulier lors du calcul de proximité sémantique, d'appariement entre une requête et des documents, etc. A cette fin, l'utilisation de standards pour coder l'indexation est conseillée.
- **Flexibilité** et **interopérabilité** : si l'effort d'indexation est coûteux, on comprendra aisément que l'on veuille mutualiser cet investissement en proposant une indexation réutilisable dans différents contextes et différents systèmes. Là encore, l'utilisation de standard est préconisée.
- **Rigueur sémantique** : ce critère rejoint les attentes décrites dans la catégorie précédente concernant l'expressivité, mais d'un point de vue technique. En effet, il est indispensable que, outre les capacités calculatoires qu'elles présentent, les technologies mises en œuvre pour l'indexation respectent les qualités liées à l'expressivité de l'indexation (section II.1.1).

II.1.3. Exploitabilité

Cette catégorie regroupe les qualités d'indexation qui relèvent essentiellement du principe d'économie. En effet, dans un contexte socio-économique fortement contraint, ce sont souvent les solutions les moins coûteuses qui sont à privilégier. Il s'agit donc d'obtenir le document indexé le plus valorisable possible à moindre effort d'indexation.

- **Concision** ou **automatisation** : On l'a dit, l'efficacité des systèmes de recherche d'information repose en grande partie sur la richesse de l'indexation des documents qu'ils exploitent. Or, l'indexation manuelle est chronophage. En l'absence d'outil (semi) automatique dédiés, l'indexation la plus économique à produire et à maintenir est l'indexation la plus concise.
- **Lisibilité** : Elle reflète la capacité pour un humain à lire et à comprendre. La lisibilité peut être améliorée à l'aide d'outils de représentation textuelle ou graphique.
- **Réutilisabilité** : ce critère rejoint les notions d'interopérabilité et de complétude. Si le document peut être réutilisé dans différents contextes (ce qui est le but recherché), l'indexation doit en témoigner. Le langage d'indexation doit donc être le plus standard possible, et l'expression sémantique la plus universelle possible. Pour ce faire on peut, par exemple, rendre compte de différentes dimensions du document indexé (Rondeau, Ranwez, & Crampes, 2005).
- **Granularité** : dans certains cas, surtout pour de longs documents, seules certaines sous-parties sont pertinentes, l'indexation doit permettre d'identifier ces parties utiles.

- **Valorisation sémantique** : la recherche de documents ou d'extraits par les *robots* exploite l'expressivité de l'indexation. L'expressivité est donc indirectement partie prenante dans l'exploitabilité du document. Cette propriété peut être rapprochée de la *discrimination* et de la *représentation* évoquées dans (Baziz, 2005). En effet, l'indexation devra essayer de rendre compte au maximum de la totalité du contenu informationnel de l'entité indexée (exhaustivité) tout en préservant la discrimination avec les autres entités (spécificité).
- **Evolutivité** : dans la mesure où l'indexation est amenée à évoluer, par exemple, au fur et à mesure de l'évolution d'une représentation de la connaissance du domaine (mise à jour), il convient de disposer de documents indexés qui pourront évoluer avec la connaissance du domaine et les technologies de l'ingénierie ontologique. Dans le cas contraire, l'exploitabilité de ces ressources serait compromise.

Cet inventaire montre que certaines propriétés sont en opposition (*concision* v.s. *expressivité*, *calculabilité* v.s. *lisibilité*, *complétude* v.s. *objectivité*) et le concepteur d'un environnement d'indexation aura donc à cœur de trouver le juste équilibre pour satisfaire au mieux l'ensemble de ces contraintes. C'est en fonction de la finalité de l'indexation que l'on va privilégier l'une ou l'autre des qualités au détriment parfois des autres.

II.2. Entités à indexer et caractéristiques à prendre en compte.

Les entités que nous avons manipulées au cours des différents projets auxquels nous avons contribué, sont de différents types et caractérisées par différents attributs, que nous appelons *descripteurs*. L'indexation qui transcrit ces descripteurs pour les rendre exploitables dans un système automatisé, est fortement dépendante de leur nature. Cette section distingue les types d'entités et leurs spécificités en matière *d'indexation*.

Au cours de mes travaux de thèse, j'ai défini les entités que je manipulais comme étant des *Briques d'Information* (BI). Ces BI désignent des fragments de documents, disponibles sous (au moins) un média, caractérisés par un modèle conceptuel, et qui peuvent être intégrés dans un document *réel* (je rappelle que ces travaux concernaient la composition automatique de documents personnalisés). Une BI peut être composite, i.e. composée de plusieurs sous-briques, sous réserve que chaque sous-brique ait un modèle conceptuel consistant. Dans les applications d'alors, ces BI contenaient des sections de cours et étaient, la plupart du temps, composées de texte et d'images. Leur indexation était divisée en deux parties spécifiques : une première description concernant les caractéristiques d'édition : auteur, date de création, titre, etc., et une deuxième description concernant le contenu, décrit à l'aide de l'ontologie du domaine. On retrouve là une distinction qui est faite dans le domaine de la *documentation*.

Le terme "indexation" est souvent employé de manière générique pour désigner l'association de métadonnées (informations secondaires) à une entité. Or d'un point de vue *documentaliste*, une distinction est à faire entre les métadonnées qui peuvent être utilisées dans le cas d'une recherche *documentaire* (à partir de l'auteur, de la date d'édition, du titre, etc., c.f. chapitre 1), on parle alors de *catalogage*, et celles qui peuvent intervenir dans une recherche contextuelle (en lien direct avec le contenu), on parle alors *d'indexation* (c.f. Figure 4).

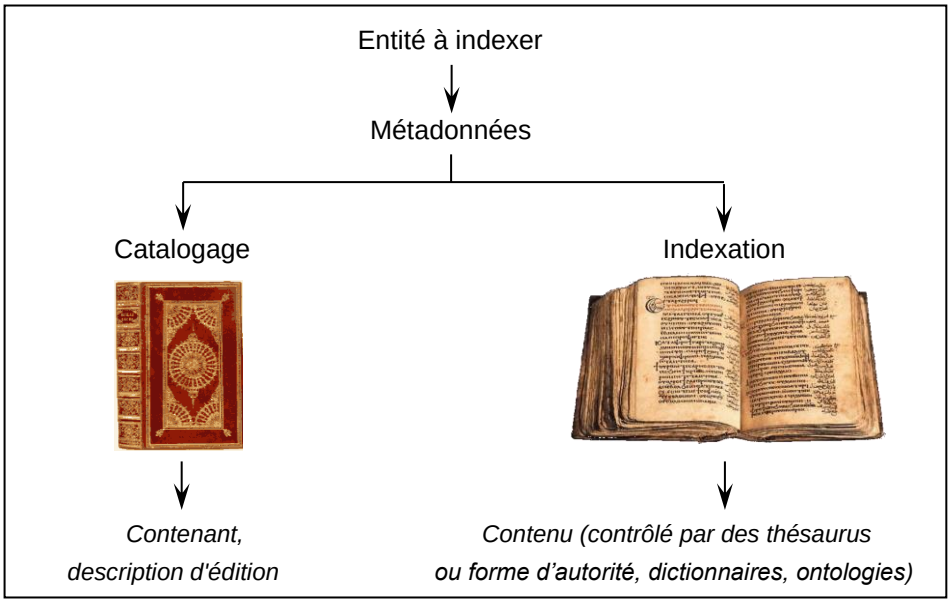


Figure 4 : Catalogage et indexation, deux moyens d'accéder à l'information (Baar, 2004)

II.2.1. Catalogage ou description d'édition

Les informations relatives au catalogage ne sont pas lourdes à renseigner. En effet, elles peuvent souvent être extraites de bases de données existantes, car les informations qu'elles contiennent sont souvent *universelles* (identiques pour toutes les applications). De plus, ces informations sont souvent communes à un groupe d'entités et peuvent être renseignées en une seule fois pour tout le groupe. Par exemple dans les applications pédagogiques, les segments appartenant à un même cours ont le même auteur, etc. Dans le projet *Prompter* qui avait pour but la composition automatique de résumés de rencontres sportives, tous les segments vidéo partageaient une description d'édition qui contenait : l'auteur de la vidéo, la date de la rencontre, le lieu de la rencontre, etc. D'autres informations peuvent être renseignées automatiquement : durée du segment vidéo, date de l'indexation, format… Pour ce qui est des documents textuels, publications scientifiques notamment, les informations d'édition (journal ou conférence, date, auteur, ISBN, nombre de pages, etc.) peuvent être extraites des bases de données que nous interrogeons (PubMed, par exemple). Il en est de même pour les titres musicaux (projet MBox-*composer*) pour lesquels la base de données ABC4 de Nétia contenait déjà toutes ces informations. Enfin, pour l'indexation de photos d'évènements, il suffit à l'utilisateur de renseigner une fois ces données et elles s'appliquent à toutes les photos de la collection. L'ensemble de ces métadonnées de catalogage peut donc être renseigné rapidement et à moindre effort pour l'opérateur humain.

II.2.2. L'indexation du contenu pour suppléer à une recherche contextuelle

Si les métadonnées de catalogage sont grandement utilisées pour la recherche documentaire, les recherches actuelles (les requêtes sur Internet, par exemple) sont souvent des recherches contextuelles où l'opérateur humain, via des moteurs de recherche, va directement *fouiller* le contenu du document. Dans le cas de documents textuels, des analyses automatiques du langage naturel permettent d'obtenir des résultats souvent satisfaisants. Cependant, des ambiguïtés peuvent subsister, qui parasitent les résultats (nous y reviendrons longuement dans la suite). Pour les photos ou les segments vidéo, de même que pour les enregistrements sonores, les techniques d'analyses automatiques du contenu peuvent être plus coûteuses en temps de calcul et il est donc nécessaire de proposer des alternatives. Il convient alors de faire une indexation du contenu de façon à transcrire

dans les métadonnées, de la façon la plus synthétique possible, ce qui est contenu dans les entités manipulées. Pour cela, une indexation manuelle est souvent indispensable.

Par exemple dans le projet Prompter, les segments vidéo devaient être décrits en fonction des actions que l'on pouvait voir sur ce segment (pour un match de foot, par exemple, un *but* ou un *coup franc*, etc.) et les acteurs impliqués dans cette action (c.f. section II.4). Mais parmi tous les projets cités, celui pour lequel l'étape d'indexation a été la plus délicate, car la plus spécifique, est sans doute le projet MBox-*composer*. En effet, la composition automatique de *playlists* suppose une indexation concernant différentes caractéristiques *artistiques* traduisant les ressentis de DJ (*Disc-Jockey*). A l'aide de DJ, nous avons défini des *humeurs – moods* en anglais (23 au total), qui peuvent décrire un titre musical pour traduire ses propriétés *psycho-sensorielles* (quel sentiment, quelle sensation évoque ce titre chez son auditeur), *sociologiques* (traduit-il un positionnement politique, social, etc.), ou son *style* (sentimental, romantique…). Ce travail a été réalisé à partir de lectures de critiques musicales et à partir de réunions d'échanges avec les DJ. Ce sont principalement ces caractéristiques qui vont influencer la composition d'une *playlist*, ciblée en fonction d'objectifs fortement dépendants du contexte et du lieu de diffusion (radios, galeries commerciales, boutique de luxe…). Parmi les 23 critères retenus, on trouve par exemple : "*goût*", "*attitude*", "*intelligence*", "*élévation*", "*sentimental*", "*sensuel*", "*festif*", "*voix*". L'objectif de l'indexation est d'associer à chaque titre un vecteur de 23 réels variant de 0 à 100 et traduisant le degré de représentativité de ces 23 critères dans le titre considéré (échelle sémantique). Par exemple 0 pour "*goût*" traduit "*doux*" alors que 100 dénote le caractère "*épicé*" d'une musique ; 0 pour "*attitude*" traduit "*respectueux*" alors que 100 traduit "*rebelle*".

En ce qui concerne les autres applications, manipulant des publications scientifiques ou des gènes, nous avons utilisé les indexations réalisées par les experts du domaine et disponibles en ligne (PubMed ou Ensembl).

Dans la suite de ce mémoire, nous parlerons principalement d'indexation, car il s'agit bien de la description du contenu des entités à l'aide d'un modèle de connaissance du domaine qui sera utilisée dans nos outils de recherche d'information, et non de l'information de catalogage.

II.3. L'indexation conceptuelle[45] au cœur de notre démarche

S'il est possible de retrouver des textes à partir de simples mots-clés, en recherchant directement dans leur information primaire, on ne peut nier les problèmes liés à la *synonymie*, à la *polysémie*, aux fautes de frappe, aux ambiguïtés, etc. Certaines solutions ont été proposées dans le domaine de l'analyse de texte, en particulier l'approche LSI (*latent semantic indexing*) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) qui a pour but de représenter les termes dans un contexte, pour évaluer leurs affinités sémantiques. Cependant, ces solutions ne règlent pas tous les problèmes, et quoi qu'il en soit, pour les entités non-textuelles (documents multimédia, gènes, images), la recherche d'information nécessite l'utilisation de métadonnées, d'index.

II.3.1. Atouts de l'indexation conceptuelle

L'indexation conceptuelle consiste à associer à une entité, une description de son contenu sous forme d'un ensemble de concepts du domaine (*sac* de concepts), ceux-ci étant définis formellement dans une ontologie. Ces concepts peuvent être organisés en réseaux sémantiques (e.g. approche DocCore de (Baziz, 2005)) ou en fonction de la topologie de l'ontologie dont ils sont issus (e.g. DocTree de (Baziz, 2005)).

Le principe d'utilisation des ontologies pour décrire des fragments de connaissance n'est pas nouveau et c'est même l'une des raisons qui ont conduit à leur avènement et favorisé leur essor (Gruber, 1993). Pendant de nombreuses années, l'apport des ontologies en matière d'indexation était

[45] On parle également *d'indexation sémantique* (Hubert et al., 2009).

principalement associé au maintien de la conformité avec un vocabulaire contrôlé. Les avancées en matière de formalisation ontologique permettent aujourd'hui de les intégrer dans tout le processus d'indexation puis de recherche d'information. En effet, elles fixent un cadre sémantique où les relations d'hyponymie (relation "*est_un*" ou "*sous_classe_de*"), d'hyperonymie ("*super_classe_de*"), de méréonymie (relation "*se_compose_de*"), et de relations spécifiques au domaine (causalité, temporalité, etc.) peuvent être activées pour inférer des connaissances supplémentaires, étendre les requêtes ou effectuer des calculs de distance/proximité sémantique. Déjà (Menzies, 1999) présentait quatre avantages apportés par les ontologies : l'interopérabilité, la réutilisabilité, la structuration et la recherche d'information (calculabilité). D'autres avantages de l'indexation sémantique sont décrits dans (Hubert, Mothe, Ralalason, & Ramanonjisoa, 2009). Ainsi les ontologies semblent favoriser les différents critères évoqués dans la section II.1 de ce chapitre. Cependant pour répondre à ces critères, l'opérateur en charge de l'indexation doit posséder à la fois une bonne connaissance du domaine (afin de caractériser au mieux le contenu de l'entité à indexer) et une bonne maîtrise de l'ontologie (afin de positionner l'indexation à un degré de spécialisation adéquat). Concernant ce dernier point, indexer à un niveau trop générique affecterait toute la chaîne de traitements des éléments de connaissance indexés (ce problème est connu et particulièrement discuté dans la communauté bioinformatique : *shallow annotation problem* (Sevilla et al., 2005; Wang, Du, Payattakool, Yu, & Chen, 2007)). Par ailleurs, il faut également éviter les redondances : ne pas indexer avec plusieurs concepts liés entre eux (e.g. par une relation de spécialisation).

L'utilisation d'ontologies permet également d'évaluer la qualité de l'indexation conceptuelle et aussi la qualité de l'ontologie elle-même. D'une part, le processus d'indexation permet de détecter des *trous* conceptuels (des parties de l'ontologie qui ne sont pas définies avec suffisamment de finesse). D'autre part, l'ontologie permet d'estimer le niveau de redondance de l'indexation (l'index contenant un concept et certains de ses fils, par exemple), son niveau de généralité (un document décrit uniquement avec des concepts très spécifiques est vraisemblablement très spécialisé alors qu'un document décrit par des concepts de haut niveau relèvera plus de la vulgarisation). Cette notion est très présente dans le domaine des sciences de la vie (qualité des annotations de gènes). Ces différents aspects de l'indexation *sémantique* et de l'indexation *conceptuelle* sont largement discutés dans (Baziz, 2005).

II.3.2. Limites identifiées de l'indexation conceptuelle

Cependant ne soyons pas dupes : les ontologies ne sont pas LA solution idéale et leur utilisation se heurte à certaines limites. Les quatre qui sont présentées ici nous semblent être les principales.

La première limite est une évidence : il faut **disposer d'une ontologie du domaine** ! Or, la conception dédiée d'une ontologie est un processus long et coûteux en ressources humaines et calculatoires. L'idéal consiste à pouvoir réutiliser des ontologies existantes, mais, comme nous l'avons dit dans le premier chapitre, une ontologie est toujours conçue dans un contexte particulier et avec une finalité précise. En conséquence, tant dans sa structure que dans le langage de formalisation choisi ou de son contenu, elle va présenter des spécificités qui auront de fortes répercussions sur les applications qui l'utilisent et donc sur l'indexation. Par exemple le degré de couverture de l'ontologie (niveau de détail) et le fait que certaines parties peuvent être beaucoup plus détaillées que d'autres impactera les calculs de similarité sémantique et donc la recherche d'information. Des ajustements sont donc souvent nécessaires. Il se peut qu'on souhaite un degré de détail supplémentaire par rapport au modèle existant ou au contraire une version moins détaillée. Cette adaptation de l'ontologie permet de ne pas noyer l'information utile dans un trop grand nombre de concepts et d'améliorer ainsi les performances de calcul et les interactions avec un opérateur humain. Il se peut aussi que les appellations (*labels*) des concepts ne correspondent pas à ceux en usage dans le domaine : par exemple entre physiciens, biologistes et chimistes, les appellations peuvent varier, comme nous l'avons expérimenté dans le projet ToxNuc-E. Dans ce projet, une ontologie du domaine de la Toxicologie Nucléaire a été conçue (c.f. chapitre I, section I.2.3.2) pour

répondre à un besoin de recherche d'information. Basée sur cette ontologie, une indexation lexicale est en cours basée sur la détection des labels des concepts dans les documents à indexer. Cette indexation automatique à partir d'analyse lexicale, on le sait, présente des limites d'un point de vue conceptuel (c.f. point suivant), mais elle offre l'avantage de pouvoir indexer un grand nombre de documents en peu de temps. Ces documents indexés seront mis en commun via la plateforme collaborative[46] dédiée afin d'être exploités par les membres du projet.

La seconde limite posée par l'indexation conceptuelle concerne son **automatisation**. Si les techniques d'analyse de texte ont considérablement progressé durant les dernières années, elles ne permettent pas (et ne permettront jamais ?) de franchir le "*gap sémantique*". Même si des solutions partielles et ponctuelles ont été proposées dans la littérature, cette problématique reste ouverte. On désigne par *gap sémantique* la difficulté à traduire au niveau sémantique des résultats d'analyse lexicale automatique. Ce problème est généralisable, en effet que ce soit en traitement de langage comme en traitement d'images, les caractéristiques identifiées par des analyses automatisées (dites de bas niveau) doivent être mises en relation avec des entités sémantiques décrites dans les images ou dans les textes. Or, cela nécessite une expertise humaine, une indexation manuelle, qui ne peut pas s'appliquer aux systèmes d'information modernes dont le contenu est mis à jour régulièrement et augmente sans cesse (Lazaridis, Axenopoulos, Rafailidis, & Daras, 2012). Nous tentons néanmoins de réduire au maximum ce fossé sémantique, en couplant les analyses lexicales et ontologiques (S. Ranwez et al., 2013). Pour ce faire, nous envisageons d'utiliser la méthode Synopsis développée par Benjamin Duthil (Duthil, Trousset, Dray, Montmain, & Poncelet, 2011) pour apprendre le vocabulaire rattaché à un concept. Ensuite, par analyse lexicale des documents présents dans le corpus, il serait possible de rattacher certains concepts à ces documents. La qualité d'une telle indexation serait certes de moins bonne qualité qu'une indexation faite par un expert du domaine, mais elle permettrait de traiter un grand nombre de documents. Synopsis a déjà été couplée avec une approche conceptuelle, dans un contexte légèrement différent, lors du développement de *CoLexIR* (S. Ranwez et al., 2013). Le facteur le plus limitant dans ce type d'approche hybride est l'apprentissage et la construction des lexiques associés aux concepts qui se fait via de nombreuses requêtes sur le Web. Ce temps d'apprentissage peut être très long. Cependant, une fois fait, les résultats de cet apprentissage peuvent servir à indexer de nombreux documents ou être utilisés en temps réel, par exemple pour visualiser les segments de texte où apparaissent les différents concepts, comme c'est le cas dans *CoLexIR*.

Parfois il est nécessaire de baser l'**indexation conceptuelle sur plusieurs ontologies**. Se posent alors des problèmes d'alignement entre ces ontologies. C'est une problématique que nous avons rencontrée dans le cadre de la thèse de Fabien Jalabert et pour laquelle il a choisi d'utiliser UMLS comme passerelle entre différents modèles. C'est également un des verrous soulevés dans notre proposition ANR CIRCE, puisque l'étude de la complémentarité d'un ensemble de ressources provenant de domaines différents peut nécessiter l'alignement des ontologies de ces différents domaines. Le sujet de thèse qui a été proposé à Nicolas Fiorini a pour objectif de proposer des avancées sur le thème de la complémentarité en RI. Il est à noter que dans le cas où différentes ontologies sont utilisées pour l'indexation, il sera nécessaire de mettre en place des techniques de requêtage qui exploitent également plusieurs ontologies. Dans ce cas, des mesures de distance entre concepts de différentes ontologies alignées doivent être imaginées. Si des travaux ont été proposés dans la littérature (Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Sánchez, Solé-Ribalta, Batet, & Serratosa, 2012), leurs limites actuelles laissent le champ libre à de nouvelles propositions. La collaboration en cours avec l'université Rovira i Virgili et plus particulièrement avec Montserrat Batet[47] et David Sánchez[48] lors de leurs séjours au LGIP2 nous a permis de proposer une nouvelle approche. Une publication va être soumise prochainement à un journal international.

[46] http://www.toxcea.fr/

[47] Automne 2012.

[48] Printemps 2013.

Enfin, l'indexation conceptuelle ne permet pas toujours **de représenter toutes les informations** caractérisant une entité. Nous avons été confrontés à plusieurs difficultés. En tout premier lieu, les informations concernant la *temporalité* sont difficiles à gérer : des instances, voire le modèle de connaissance lui-même, peuvent évoluer. Par exemple dans certains cas les instances du modèle sont temporaires. C'est une des difficultés que nous avons rencontrées dans le projet *Prompter*, où chaque match de football devait avoir sa propre description et ses propres instances, un joueur pouvant même changer de rôle (poste) en cours de match. Nous avions alors une composante "volatile" des instances de l'ontologie qui était propre à chaque match à indexer et susceptible d'évoluer en temps réel. Par ailleurs il peut également être difficile de formaliser certaines informations relatives au degré de représentativité d'un concept dans une ressource : son aspect *fréquentiste*. Ce degré peut représenter, par exemple dans un document textuel, la fréquence d'apparition des labels relatifs à ce concept par rapport à la même fréquence dans tout le corpus. Cela revient à estimer à quel point ce concept est discriminant pour le document. Dans ce cas, un poids associé à chaque concept de l'indexation permettrait de prendre en compte ce facteur dans la suite du processus de RI, mais affecter ces poids au moment de l'indexation n'est pas chose facile : leur attribution nécessite de connaître l'ensemble des indexations du corpus (une mise à jour du corpus remet en cause l'intégralité des indexations). Enfin une information concernant la *fiabilité* de l'indexation peut être intégrée. Concernant l'indexation de gènes, il existe un facteur de confiance associé aux annotations utilisant la Gene Ontology (*evidence code*, facteur dépendant de la façon dont a été attribué le concept : soit de façon automatique, soit par un expert du domaine). Un tel poids permettrait de prendre en compte la *fiabilité* de l'indexation, mais son affectation n'est pas facile dans le cas général car difficile à estimer de façon automatique et très subjectif lorsqu'il est attribué manuellement (c.f. qualité d'objectivité dans la section II.1.1).

Au vu des limites de l'indexation classique, il est clair qu'une indexation manuelle est trop fastidieuse pour de grands corpus mais qu'une indexation entièrement automatique est souvent trop approximative. La conception et le développement de solutions pour une indexation semi-automatique nous ont donc semblé incontournables.

II.4. L'indexation par propagation : principe et mise en œuvre

La nécessité d'associer des métadonnées à des documents numériques de plus en plus nombreux a conduit à imaginer des mécanismes pour déduire de nouvelles indexations à partir d'indexations existantes. Par exemple dans (Marchiori, 1998), l'auteur propose de propager des métadonnées sur la toile, en utilisant les hyperliens contenus dans les pages comme support de cette propagation. Les auteurs de (Abrouk, Gouaïch, & Raïssi, 2005) se servent également des citations entre les documents pour propager leurs annotations. On peut cependant regretter que cette approche n'analyse pas le contenu des pages et se focalise sur des liens de référencement. Des caractéristiques fausses peuvent être propagées (e.g. si une page ou un document est cité comme contre-exemple). Pour améliorer significativement la recherche d'information conceptuelle, l'indexation utilisée et donc la propagation permettant de l'obtenir doit être précise.

Dans le cadre du projet Prompter, nous avons proposé un environnement d'indexation qui permet d'associer des triplets (*Source – Action – Destination*) à des segments vidéo. Une DTD a été proposée pour contraindre la structure de cette indexation (S. Ranwez & Crampes, 1999b). Cependant, la saisie étant relativement lourde pour les utilisateurs, nous l'avons simplifiée afin de la rendre la plus intuitive possible. Sur la Figure 5, on voit dans le bandeau du haut, des raccourcis qui permettent d'accéder directement à certains concepts de l'ontologie (arbitre, joueur, équipe…), ou à certaines de leurs actions. Sur le cadre de gauche, on visualise l'ontologie et il est possible de sélectionner des concepts et des relations. Le cadre de droite récapitule les concepts saisis. Cependant dans un contexte plus large, confrontés à un plus grand nombre de données, ce type d'interface n'est plus réaliste. C'est ce qui est appelé le problème du passage à l'échelle (*scalability*). Nous avons été confrontés à cette limite lors du projet *MBox-composer* où la base de données de test

de l'entreprise Nétia contenait plus de 4 000 titres musicaux à indexer suivant 23 critères. Il a alors fallu chercher des alternatives, c'est ce qui nous a conduits vers l'indexation par propagation.

Figure 5 : Interface d'indexation des segments vidéo sous forme de triplets dans le projet Prompter.

II.4.1. Principe de l'indexation par propagation

Ce type d'indexation repose sur un pré-requis : on dispose d'un échantillon représentatif d'entités préalablement indexées : le support d'indexation. Ce support peut être représenté graphiquement par une carte sémantique sur laquelle le positionnement des différentes entités préalablement indexées est fonction de leur indexation. Autrement dit, la proximité physique des éléments sur la carte traduit une proximité sémantique de ces éléments. Le paysage obtenu peut alors être vu comme une projection à deux dimensions d'une indexation à *n* dimensions. Pour cette projection, nous avons mis en œuvre une implémentation du MDS (*MultiDimensional Scalling*) par l'algorithme des *ressorts* permettant d'obtenir une projection 2D, respectant au mieux les distances sémantiques calculées dans l'espace à *n* dimensions.

Sur ce paysage, qui peut être explicité pour l'utilisateur grâce à différentes techniques de visualisation (c.f. chapitre 4), les nouveaux éléments à indexer sont positionnés par *glisser-déposer*. En positionnant une nouvelle entité sur la carte, l'utilisateur indique graphiquement les entités indexées qu'il estime être les plus proches sémantiquement de celle qu'il doit indexer. Celle-ci va se voir attribuer automatiquement une indexation déduite à partir de l'indexation de ses *k* plus proches voisins. Pour cela, le système utilise les caractéristiques implicitement présentes dans la carte sémantique, afin de les propager pour le nouvel élément. Une boucle de rétroaction est envisageable pour affiner les résultats, l'utilisateur pouvant retoucher ou préciser cette indexation afin de la rendre plus pertinente.

II.4.2. Mise en œuvre de l'indexation par propagation

Cette solution a été mise en œuvre et validée dans le cas de l'indexation de titres musicaux (Crampes, Ranwez, Velickovski, Mooney, & Mille, 2006). Dans ce cas, comme nous l'avons dit plus haut, l'indexation d'un titre possède deux composantes. La première concerne sa description éditoriale : titre, éditeur, durée, tonalité, interprète, auteur, compositeur, etc. Cette composante était disponible dans la base de données de notre partenaire industriel et des outils de traitement de signal

permettent d'automatiser en grande partie cette description (tonalité, rythme…). La seconde composante concerne une description plus artistique de l'œuvre. C'est cette dernière composante qui a fait l'objet de l'indexation par propagation.

Dans la section II.2.2, nous avons présenté les différents critères permettant de caractériser des titres musicaux. On peut remarquer que ces critères sont fortement subjectifs et dépendants d'un contexte culturel, social, politique. Il est donc impossible de réaliser cette indexation de manière entièrement automatique. Cependant le nombre de titres à indexer rend impossible une indexation entièrement manuelle. C'est pour trouver le juste milieu entre automatisation et expertise humaine, que nous avons proposé l'indexation par propagation dans l'outil MBox-Composer (c.f. Figure 6).

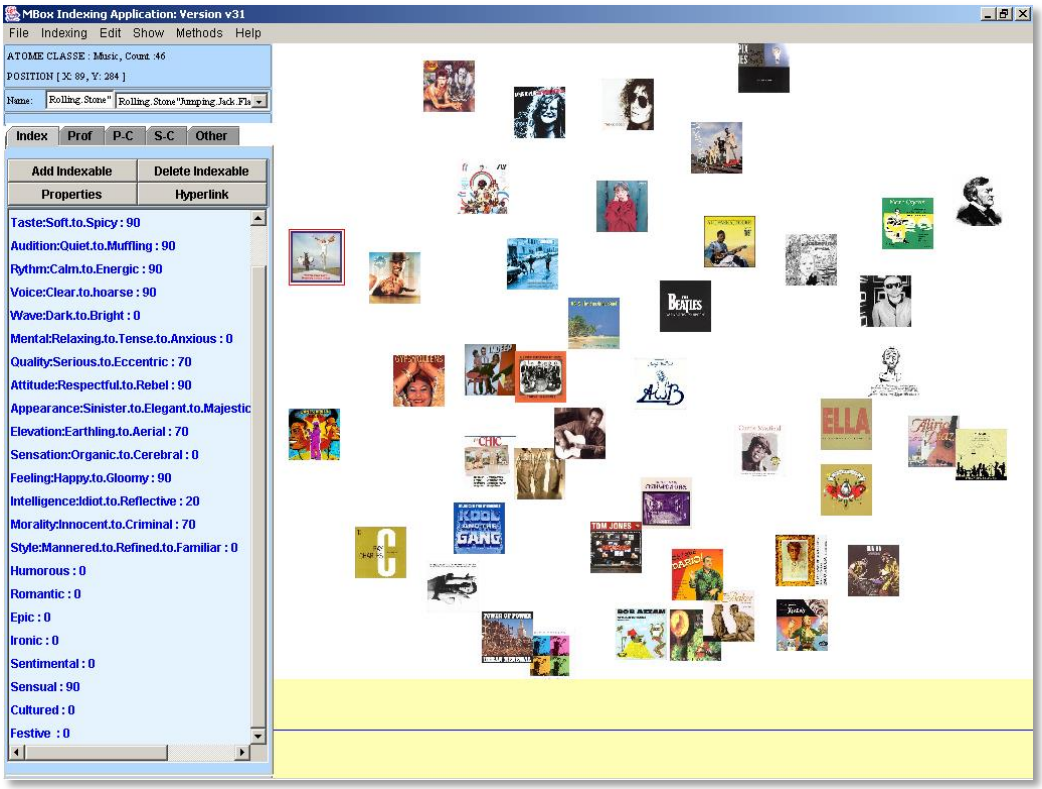


Figure 6 : Interface de l'outil MBox-Composer.
Sur la droite on voit le paysage musical composé des titres indexés représentés par la pochette du disque. L'un de ces titres est sélectionné (les Rolling Stones encadrés en rouge).
Le détail de son indexation (les valeurs des poids des différents critères) est donné en partie gauche.

Le support d'indexation initial est composé de 45 titres, pour lesquels nous avons demandé à un expert d'attribuer pour chacun des 23 critères, un poids correspondant au degré de représentativité de ce critère dans chaque titre. Ainsi un titre peut être considéré comme romantique avec un poids de 70 et rythmé avec un poids de 50… Cette tâche a confirmé l'aspect fastidieux que représenterait une indexation entièrement manuelle de la base de titres musicaux. Une fois cette tâche réalisée, nous en proposons une projection sur un paysage dit "de référence". L'expert a validé cette projection qui faisait sens pour lui, puisqu'elle lui a permis d'identifier des catégories musicales dans certaines zones de la carte (c.f. Figure 7).

Disposant de ce paysage, la phase d'indexation de la base consiste à proposer à plusieurs experts chacun des nouveaux titres musicaux à indexer sous la forme d'une icône représentant, par exemple, la pochette du disque et de déposer cette icône à l'endroit le plus pertinent de la carte,

l'expert humain étant à même d'effectuer des rapprochements, d'estimer des similitudes entre titres musicaux. Cette opération par *glisser-déposer* permet au système d'inférer une description pour chaque nouveau titre, en fonction de sa proximité à d'autres titres déjà indexés. Différents algorithmes de propagation ont été testés (Crampes, Ranwez, Velickovski, et al., 2006) parmi lesquels nous avons retenu celui des *k plus proches voisins* avec *k = 4*, car c'est celui qui donnait le meilleur compromis entre la pertinence des résultats et le temps de calcul. Au cours de ce processus, une assistance visuelle permet de rendre compte de cette propagation grâce à un spectre sémantique représentant les valeurs des poids de chaque descripteur par des barres d'autant plus longues que leur poids est élevé (ce point sera discuté dans la section 3.2.1 du chapitre 4).

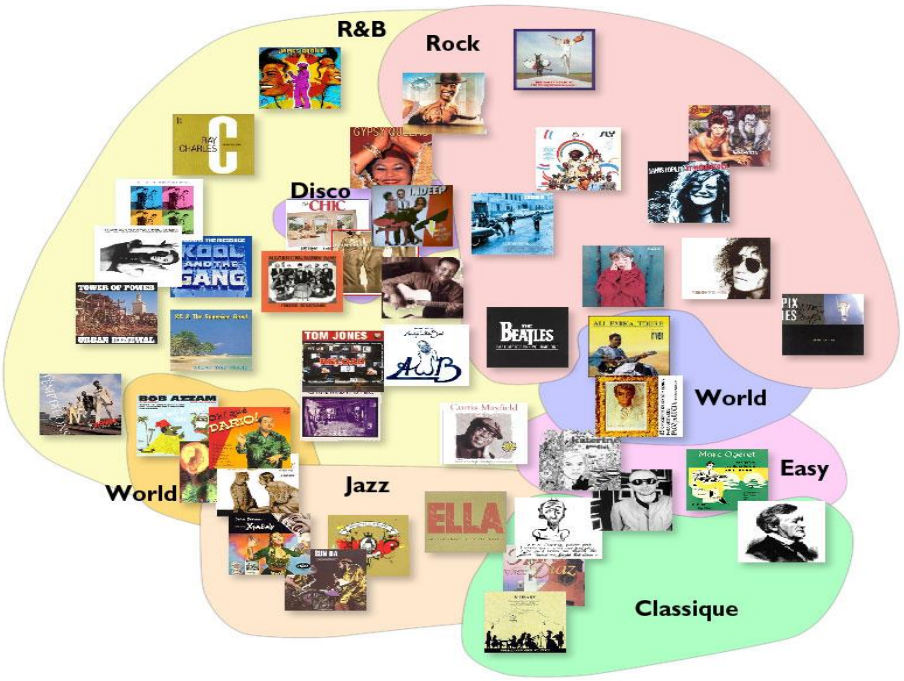


Figure 7 : Paysage de référence obtenu par projection du support d'indexation
Il est possible d'identifier certains courants musicaux en regardant les regroupements des titres à partir de cette projection (identification réalisée manuellement par un DJ).

II.4.3. Résultats et limite de l'approche

L'indexation complètement manuelle avec une liste de critères à renseigner (les poids des 23 descripteurs musicaux) a montré que l'opérateur délaissait rapidement certains descripteurs et se cantonnait à attribuer une valeur à seulement deux ou trois d'entre eux. Il en résultait une indexation très incomplète, qui pénalisait fortement les traitements appliqués par la suite (composition automatique de *playlist*). Au contraire, l'indexation par propagation était vécue comme plus ludique et rapidement, tous les titres étaient indexés sur l'ensemble des descripteurs proposés. Ainsi, même si, en théorie, elle fournissait des résultats plus approximatifs qu'une indexation entièrement manuelle, l'indexation par propagation constituait en pratique une solution très efficace pour le passage à l'échelle (indexation de milliers de titres) et permettait d'obtenir des indexations sur la totalité des critères. Ce résultat s'explique d'une part par la simplicité du processus (facilité d'interaction), mais aussi par l'analogie avec les classements de disques ou de photos que l'on peut faire "à la main". En effet, les DJ étalent souvent leurs disques devant eux sur une table, en les regroupant en fonction de certains critères. Ils utilisent ensuite ces différents tas pour concevoir leurs *playlists*. On a pu faire des tests avec quatre DJ de sensibilités différentes qui se sont reconnus dans

les paysages musicaux proposés et ont indexé rapidement un grand nombre de titres (en moyenne une vingtaine de titres en 5 minutes). Les limites qu'ils ont exprimées concernaient principalement la trop grande flexibilité de l'indexation (trop de critères à renseigner et donc des affinages difficiles quand ils voulaient retoucher l'indexation calculée automatiquement) et les limites liées à la visualisation (surcharge de l'écran après un long moment d'utilisation de la carte). Toutefois, la rapidité et l'expressivité des indexations convenaient à leurs attentes.

Ce procédé d'indexation par propagation fonctionne donc particulièrement bien lorsque chaque élément est indexé par un petit nombre de critères indépendants dont seul le poids varie d'un élément à l'autre. Il suffit, dans ce cas, d'attribuer à chaque descripteur le poids moyen de ses voisins. Cependant cette indépendance entre les critères n'est pas toujours assurée. En particulier dans le cas où ces critères correspondent à des concepts d'une ontologie de domaine, ils sont liés par différentes relations sémantiques qui peuvent traduire une dépendance. Pour généraliser cette approche, on ne peut donc pas se contenter d'indexer les entités par un vecteur de poids dont la taille correspond au nombre de concepts d'une ontologie. Outre le problème lié à la taille de l'ontologie de domaine (et donc à la performance en temps de calcul), une telle indexation supposerait que l'indexeur maîtrise l'ensemble de l'ontologie pour ajuster si nécessaire une indexation. Or bien souvent, les personnes qui indexent maîtrisent une partie bien délimitée de l'ontologie de domaine, qui correspond à leur centre d'intérêt.

Pour toutes ces raisons, une amélioration du modèle doit être imaginée pour pouvoir être appliquée à l'indexation conceptuelle. Nous avons donc débuté une réflexion en ce sens.

II.5. L'indexation conceptuelle par propagation

Avant d'aller plus avant dans la présentation de notre réflexion sur ce sujet, il faut souligner que ces travaux sont actuellement en cours. Un ingénieur de recherche va travailler avec l'équipe pendant l'année 2013 pour finaliser l'approche et développer le prototype. Cette section présente donc des pistes de recherche plus que des résultats aboutis.

L'idée de *propagation d'annotation sémantiques*, telle qu'elle est décrite dans (Pastorello Jr., Daltio, & Medeiros, 2008), repose sur les technologies du Web sémantique (triplets RDF). Les annotations propagées concernent des données multimédia qui étant impliquées dans différents processus, peuvent être soumises à une ou plusieurs transformations. Dans ce cas, il est nécessaire de conserver les annotations présentes sur les données initiales voire de les compléter à partir des axiomes présents dans l'ontologie et qui peuvent également provenir de la caractérisation du processus de transformation. Ce modèle ne correspond pas à notre approche. En effet, il ne concerne pas l'indexation de nouvelles données, mais la propagation d'index sur une même donnée de base. Dans leur cas, il n'y a pas de problème posé par la propagation d'annotations contradictoires, par exemple. Le terme de propagation n'est donc pas employé dans le même sens que nous l'entendons.

De façon analogue à ce que nous avons présenté dans la section précédente, nous disposons d'un ensemble d'entités, déjà indexées par un ensemble (sac) de concepts issus d'une ontologie de domaine. Nous souhaitons attribuer une indexation conceptuelle à de nouvelles entités, par propagation. Cette propagation peut être réalisée lorsque l'on dépose un nouvel élément sur une carte conceptuelle présentant le corpus indexé. Pour des documents textuels, on peut aussi imaginer une première estimation des distances à partir de l'analyse lexicale de leur contenu et la propagation de certains concepts en fonction de ce calcul de distance.

Dans les deux cas, si l'on transpose le principe de la propagation évoqué plus haut à une indexation conceptuelle, plusieurs points sont à (re)définir :

i) Quel contenu d'indexation propage-t-on ? Est-ce que l'indexation résultant de la propagation contient l'union de tous les concepts c_{kx} des k-plus-proches voisins ? Doit-on propager également d'autres concepts liés à ces concepts c_{kx} – notion de sous-

ontologie (V. Ranwez et al., 2009; V. Ranwez, Ranwez, et al., 2012) ? En cas de propagation de concepts liés par une relation de subsomption, garde-t-on le plus générique, le plus spécifique ou les deux ?

ii) Quelle mesure de similarité sémantique est la plus appropriée pour cette propagation en fonction de différents objectifs (lissage/séparation) ? Sur la carte sémantique, quelle est la distance sémantique la plus pertinente pour définir la proximité entre les documents ?

iii) Où s'arrête le voisinage (i.e. quand met-on un terme à la propagation) ?

iv) Comment peut-on résoudre les possibles conflits qui surviennent : si des indexations contradictoires doivent être propagées, quelle stratégie peut-on utiliser pour décider de l'indexation qui doit être retenue ?

v) Quels poids attribuer aux concepts propagés ?

Ces travaux font partie de notre projet de recherche à court terme. En effet nous avons des pistes de réflexions en cours d'évaluation. Concernant le deuxième point, par exemple, nous souhaitons exploiter la propriété de connectivité démontrée dans un article actuellement en soumission (M. F. Sy, Ranwez, Montmain, & Ranwez, n.d.). De plus, nos travaux sur les distances sémantiques, devraient également nous permettre de sélectionner la mesure la plus appropriée en fonction de différentes stratégies de propagation. L'avantage majeur de notre approche porte sur la justification et le contrôle de l'indexation en donnant une vue sémantique des plus proches voisins qui tienne compte du calcul de similarité entre concepts pour proposer à l'utilisateur une explication de l'indexation proposée. Notons que l'opérateur humain reste au centre du processus et que c'est lui qui "dirige" l'indexation. Pour cela, une connaissance au moins partielle de l'ontologie du domaine est nécessaire et un environnement adapté doit être pensé. Si cette intervention humaine reste nécessaire, on envisage également une approche communautaire dans le cadre des projets de plateformes collaboratives qui permettrait de répartir l'effort d'indexation entre les différents acteurs. Nous comptons également tirer parti de ces travaux sur l'indexation conceptuelle par propagation pour compléter l'indexation du corpus lié à la cancérologie (stage de Pierre Fontana, 2012). Ces travaux pourraient éventuellement être enrichis pour prendre en compte certains aspects de contexte et suggérer une convergence de *tags* à partir, par exemple des documents provenant des mêmes auteurs ou lus par un même lecteur. Cette mise en pratique des usages pour suggérer un enrichissement des indexations conceptuelles rejoint en partie le courant *folksonomique* qui tend à conseiller des objets d'intérêt entre utilisateurs. Mais, dans notre cas différentes contraintes impliquent des exigences bien supérieures à de simples conseils entre "amis" : le type d'utilisateurs visé (experts provenant de communautés connexes, par exemple), le caractère sémantique de la fouille (importance de la *fiabilité* de l'indexation), le besoin de croiser des documents sémantiquement complémentaires (possibilité de croiser plusieurs ontologies), le souci de visualisation et la nécessité de justifier les résultats.

II.6. Synthèse

L'indexation est au centre de tout processus de recherche d'information. Si l'approche conceptuelle a apporté des éléments de réponse aux problèmes de la synonymie, de la polysémie, et levé par là même de nombreuses ambiguïtés, l'indexation préalable qu'elle nécessite reste difficile à automatiser. L'opérateur humain reste garant de la qualité de l'indexation (en termes de fiabilité) et il est nécessaire de l'assister dans cette tâche, en lui proposant des outils intuitifs et adaptés à son contexte.

Pour répondre à ce besoin, nous avons proposé une méthode d'indexation par propagation qui a été éprouvée dans le domaine de la *musique* et mise en œuvre pour indexer des titres musicaux. Une déclinaison de cet environnement a été proposée pour l'indexation conjointe de photos et de

musiques, avec des résultats plus mitigés (Crampes, Gout, Mille, Villerd, & Ranwez, 2007) ou à l'indexation de photos seules (Crampes, De Oliveira-Kumar, Ranwez, & Villerd, 2009). Dans ce dernier cas, notre recherche se focalisait sur la représentation du paysage initial qui sert de support à l'indexation. Ce paysage doit être aussi explicite que possible. Dans notre cas, l'indexation contenant la liste des personnes présentes sur la photo, un diagramme de Hasse était proposé. La visualisation d'un tel diagramme dans le cas d'une indexation conceptuelle ne serait peut-être pas la plus appropriée car peu explicite pour l'indexeur.

De par notre expérience en matière d'indexation par propagation et nos résultats concernant l'exploitation des ontologies, nous envisageons de poursuivre nos travaux sur l'indexation par propagation dans un contexte conceptuel. Cette indexation est un préalable à la recherche d'information conceptuelle et constitue donc une étape indispensable à notre démarche d'exploitation sémantique des données.

CHAPITRE III

«He possesses two out of the three qualities necessary for the ideal detective. He has the power of observation and that of deduction.

He is only wanting in knowledge; and that may come in time. »

Arthur Conan Doyle, "The Sign of the Four".

La recherche d'information basée sur les ontologies

III.1. LES PRINCIPES DE BASES DE LA RECHERCHE D'INFORMATION 72

III.1.1. EXPRESSION DE LA REQUETE 74

III.1.2. ORDONNANCEMENT DES RESULTATS 75

III.1.3. EVALUATION DES SYSTEMES DE RECHERCHE D'INFORMATION 76

III.2. LA RECHERCHE D'INFORMATION CONCEPTUELLE 77

III.2.1. UTILISER LA STRUCTURE SEMANTIQUE DES ONTOLOGIES POUR ETENDRE LA REQUETE 77

III.2.2. UTILISER L'ONTOLOGIE POUR LA REFORMULATION 80

III.3. OU APPROCHES CONCEPTUELLES ET LEXICALES PEUVENT SE COMPLETER 81

III.3.1. ANALYSE LEXICALE POUR COMPLETER L'INDEXATION 81

III.3.2. ANALYSE LEXICALE POUR UNE MEILLEURE COMPREHENSION DES RESULTATS 82

III.4. LES MESURES DE SIMILARITE SEMANTIQUE AU CŒUR DE LA RI 84

III.4.1. MESURES DE SIMILARITE SEMANTIQUE PAR INTENTION 84

III.4.2. MESURES DE SIMILARITE SEMANTIQUE PAR EXTENSION 85

III.4.3. PROPRIETES ET CHOIX DES MESURES DE SIMILARITE SEMANTIQUE 85

III.5. FOCUS SUR DES ALGORITHMES UTILISES DANS DIFFERENTES PHASES DE LA RI 86

III.5.1. CALCUL DES LCA D'UN ENSEMBLE DE CONCEPTS 86

III.5.2. EXTRACTION DE SOUS-ONTOLOGIES 87

III.6. SYNTHESE 88

Les différents enjeux de la Recherche d'Information (RI) ont été discutés dans le premier chapitre de ce mémoire. La quantité des informations disponibles est en forte croissance, ces informations ne sont plus seulement émises par quelques professionnels (journalistes, écrivains) sur des thèmes fédérateurs, mais sont produites quotidiennement par des millions d'anonymes, ce qui augmente leur diversité et leur hétérogénéité. De plus, il faut répondre à une évolution des besoins des utilisateurs de plus en plus habitués à des outils rapides et efficaces, aux interfaces mobiles et interactives. Le problème n'est plus de rassembler l'ensemble de l'information, mais de trier, dans les informations disponibles, celles qui sont vraiment utiles. Ce besoin est particulièrement marqué dans le domaine biomédical, où les innovations technologiques récentes (puces à ADN, séquençage haut-

débit, etc.) renforcent la production massive de données. Exploiter ces données nécessite des outils de traitement dédiés et performants (d'où l'essor de la bio-informatique), mais également des outils qui permettent de croiser l'analyse de ces données biologiques avec les publications scientifiques y afférent et favorisent ainsi les découvertes scientifiques. Dans ce contexte, la RI constitue donc un enjeu majeur.

L'essor des technologies liées au Web Sémantique a suscité une mutation des usages, et une adaptation des systèmes de recherche d'information (SRI) est nécessaire. L'utilisation des ontologies comme support à la fois de l'indexation et de la recherche d'information, semble désormais incontournable. En définissant formellement les *concepts* d'un domaine, une ontologie favorise l'interaction entre les opérateurs humains et les systèmes informatiques. Dans le même temps, les *relations* définies entre ses concepts enrichissent l'expressivité sémantique de l'indexation autant que de la requête. Cependant, si les ontologies sont au cœur de nombreux SRI actuels, ces derniers tombent souvent dans un des deux travers suivants : soit ils permettent une grande expressivité sémantique et exigent donc une forte expertise de l'utilisateur et un langage souvent complexe pour l'expression de sa requête (connaître à la fois l'ontologie de domaine, ses subtilités et les stratégies de recherche mises en œuvre), soit ils utilisent un langage de requêtage très simple, qui réduit souvent l'ontologie à un simple vocabulaire contrôlé.

Ainsi, pour faire face à une demande variée, souvent dans l'urgence, de nombreux travaux de recherche ont été initiés par la communauté RI pour proposer de nouveaux modèles et des stratégies innovantes, intégrés à des outils efficaces. Ils ont donné lieu à de multiples publications dont une synthèse très complète est disponible dans (Baeza-Yates & Ribeiro-Neto, 2011). L'objectif de ce chapitre n'est donc pas de passer en revue les différentes méthodes existantes, mais plutôt de souligner nos contributions dans ce domaine. Dans une démarche d'automatisation cognitive, nous privilégions l'automatisation de certaines tâches de l'utilisateur en essayant de respecter au mieux ses capacités cognitives et son contexte d'application. Ainsi nous apportons un grand soin à la traduction de son besoin en information en amont de la recherche d'information, puis, lors de la restitution des résultats, à la justification de ceux-ci.

Une première section dresse le cadre de nos travaux, en présentant la recherche d'information de manière générale avec ses diverses composantes. Il est à noter que nous ne nous plaçons pas dans le cadre d'une recherche ouverte sur Internet, mais dans le cadre d'une recherche dans un corpus préalablement indexé. La section 2 propose un recentrage sur l'utilisation des ontologies dans ce contexte et souligne les avantages d'une recherche conceptuelle par rapport à une recherche traditionnelle (booléenne ou lexicale). Cependant, convaincus qu'on ne peut renvoyer dos à dos les approches lexicales et conceptuelles et qu'au contraire, elles peuvent s'enrichir mutuellement dans un continuum de solutions, nous présentons dans la section 3 des solutions pour les faire cohabiter dans un système de RI. Un SRI nécessite forcément un module permettant d'évaluer la proximité sémantique des concepts afin d'estimer la pertinence d'un document vis-à-vis d'une requête ou la *ressemblance* de deux documents. La section 4 fournit une vue très synthétique des différentes mesures présentées dans la littérature et discute de leur influence sur les résultats de la recherche d'information. La section 5, enfin, présente certains algorithmes utilisés au cours d'un processus de recherche d'information. En effet, les ontologies étant de plus en plus volumineuses, il est nécessaire de disposer de solutions algorithmiques performantes pour pouvoir effectuer ces traitements en temps réel.

III.1. Les principes de bases de la recherche d'information

L'objectif principal d'un SRI peut être défini de la façon suivante : "trouver des entités (généralement des documents) extraites de larges collections (généralement stockées informatiquement) qui répondent à un besoin d'information" (Manning, Raghavan, & Schütze, 2008). Le principe de base de la recherche d'information est donc de retrouver et de proposer à

l'utilisateur un maximum d'éléments d'information pertinents, tout en lui proposant un minimum d'éléments non pertinents.

La recherche d'information (RI) concerne tout ce qui a trait à la représentation, au stockage, à l'organisation et à l'accès à des éléments d'information contenus dans des documents, des pages Web, des catalogues en ligne, des enregistrements structurés ou non, des objets multimédias. La représentation et l'organisation de ces éléments doit permettre à un utilisateur d'accéder facilement à l'information dont il a besoin (Baeza-Yates & Ribeiro-Neto, 2011).

On englobe donc sous le terme RI l'ensemble des modèles et des techniques qui peuvent être mis en œuvre dans ses trois principales composantes :

i) **L'indexation** qui concerne l'organisation des données et leur description. On y trouvera en particulier des travaux s'intéressant à la modélisation de la connaissance, à la modélisation des index associés aux documents (et donc à la représentation de ces documents par rapport à cette connaissance) et à la représentation des requêtes (toujours en fonction de la représentation de connaissance retenue).

ii) La **recherche d'information** elle-même, c'est-à-dire les différentes stratégies de RI mises en œuvre lors de l'appariement entre un élément d'information du corpus et une requête d'un utilisateur (ou d'un autre système d'information). Cela concerne à la fois les architectures logicielles qui peuvent être mises en œuvre, les calculs de pertinence permettant de classer les éléments en fonction de leur adéquation avec la requête et les préférences de l'utilisateur, les mesures de distance/similarité sémantique qui peuvent être utilisées au cours de cet appariement, les méthodes de classification, etc.

iii) Enfin, **l'interface** SRI-utilisateur qui regroupe les techniques de filtrage, de présentation, de visualisation, d'interaction, d'expression de la requête, etc. qui sont mises en œuvre pour favoriser l'interaction entre le SRI et l'utilisateur. Cela concerne en particulier toutes les fonctionnalités liées au paramétrage de la recherche d'information, et donc à la personnalisation de celle-ci.

Le lecteur attentif aura remarqué que ces trois composantes font l'objet des trois chapitres qui constituent le cœur de ce mémoire. Après avoir présenté notre vision de l'indexation dans le chapitre précédent, il convient de détailler la RI telle que nous l'envisageons. Dans notre démarche, nous nous attachons à conserver un juste équilibre entre l'optimisation des résultats de la recherche d'information et le respect des contraintes et attentes de l'utilisateur.

On peut, de façon très pragmatique résumer le principe d'un SRI de la façon suivante : dans un contexte applicatif donné, un utilisateur soumet une requête au système, de façon à obtenir un ensemble d'informations pertinentes. Si l'on met momentanément de côté les aspects liés à l'interface, le SRI est généralement constitué de trois processus de base (Belkin, Ingwersen, & Pejtersen, 1992):

- Le processus d'indexation qui consiste à représenter les ressources (souvent les documents) et les requêtes avec un ensemble de termes ou concepts, pondérés ou non, qui synthétisent au maximum leur contenu informationnel (ce point a été détaillé dans le chapitre précédent).
- Le processus de recherche lui-même. Il contient la stratégie d'appariement des ressources avec la requête de l'utilisateur. Le SRI procède à un appariement entre ces ressources, les éléments d'information indexés suivant un certain formalisme, et une requête exprimée (ou traduite) dans ce même formalisme. Les ressources sont ordonnées (*ranking*) en fonction d'un calcul de scores (appelés RSV pour *Retrieval Status Value*) et celles qui ont obtenu les meilleurs scores sont présentées comme résultats à l'utilisateur. Ce résultat est donc très fortement dépendant de l'indexation (des documents et de la requête).
- L'expansion de requête est un processus intermédiaire qui permet de reformuler la requête de l'utilisateur, pour améliorer la qualité des résultats (c.f. section III.5).

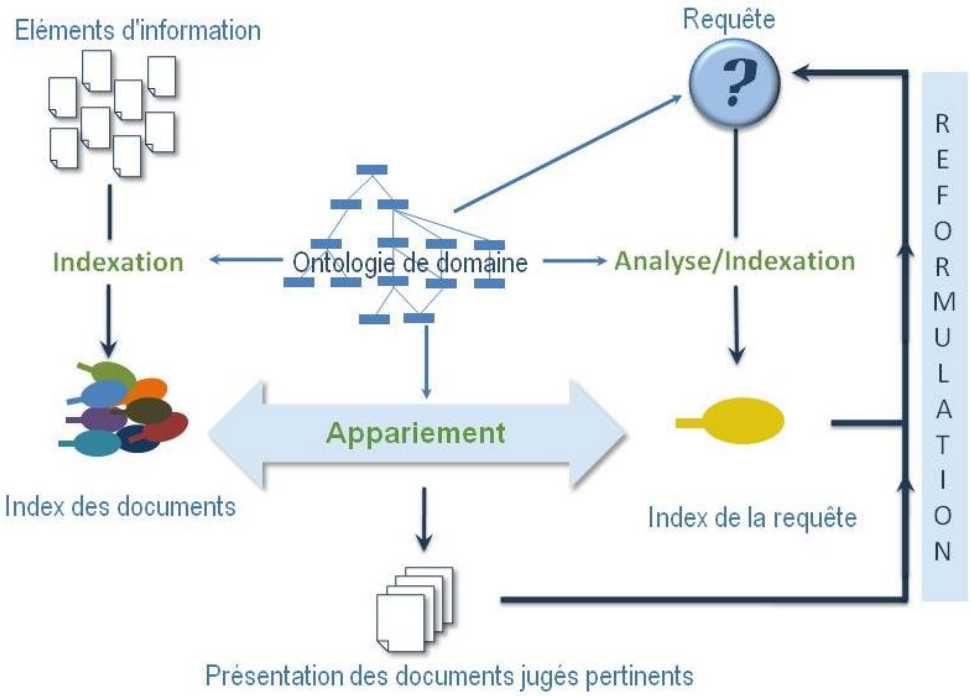


Figure 8 : Schéma global d'un SRI conceptuel utilisant une ontologie de domaine

Le schéma global du fonctionnement d'un système de recherche d'information (SRI) est représenté dans la Figure 8. Si l'indexation et/ou le processus d'appariement sont basés sur une ontologie de domaine, le SRI sera qualifié de *conceptuel*.

L'indexation choisie et l'expression de la requête jouant un rôle clé dans le processus de recherche d'information, ils sont développés dans les sections suivantes.

III.1.1. Expression de la requête

La requête formalise le besoin en information de l'utilisateur et sert de base au calcul d'appariement entre ce besoin et les entités présentes dans le corpus. Nécessairement synthétique, cette requête ne traduit pas forcément toutes les attentes de l'utilisateur. Si elle est exprimée en langage naturel, elle peut être analysée, les termes qui la constituent sont alors transformés : lemmatisés, par exemple pour une recherche booléenne ou appariés avec des labels de concepts dans le cas d'une recherche conceptuelle (*concept mapping*). Cette analyse peut être vue comme une indexation de la requête. Dans le cas d'une recherche conceptuelle, il est également possible de contraindre l'expression de la requête aux seuls concepts de l'ontologie de domaine, comme nous l'avons fait dans OBIRS[49] avec une assistance à la saisie des concepts par complétion automatique. Les concepts de la requête peuvent être pondérés ou non pour traduire les préférences de l'utilisateur, cette pondération peut avoir une grande influence sur les résultats (Farah & Vanderpooten, 2007). Parmi les facteurs qui dégradent les performances d'un SRI, on peut citer le degré de généralisation des termes de la requête. En effet, des termes trop génériques (positionnés trop haut dans la hiérarchie de concepts de l'ontologie) conduisent à des résultats peu précis.

Au terme de cette formulation, le besoin en information de l'utilisateur est contenu dans une requête système afin que celui-ci effectue l'appariement entre cette requête et les indexations des documents du corpus. Certains systèmes mettent en œuvre une étape intermédiaire appelée l'*expansion de requête*, qui intervient entre l'indexation et l'appariement. Elle a pour but de compléter automatiquement la requête de l'utilisateur pour y ajouter de nouveaux termes, enlever

[49] Ontological Based Information Retrieval System, http://www.ontotoolkit.mines-ales.fr/ObirsClient/

des termes jugés trop "pauvres", i.e. trop génériques, ou modifier les poids associés à certains termes. Cette expansion peut être faite en fonction d'un historique lié à l'utilisateur (les documents qu'il a consultés, etc.) ou se baser sur des ressources externes (ontologies, thésaurus) ; on parle alors de méthodes *globales*. La plupart du temps, on utilise pour cette expansion les liens de subsomption ou de synonymie de l'ontologie. C'est notamment le cas du SRI de la base de données PubMed[50], qui inclut dans la requête de l'utilisateur tous les termes du MeSH[51] subsumés par un des termes de la requête initiale. Lorsque cette requête étendue est traitée par le système, aucune distinction n'est faite entre les termes rajoutés automatiquement et les termes initiaux qui la composaient, aussi, les utilisateurs peuvent parfois être déroutés par l'ensemble des documents retrouvés, leur indexation en regard de la requête qu'ils avaient formulée. Cela est particulièrement vrai lorsque les documents retrouvés ne sont indexés par aucun des termes initiaux de la requête. C'est pourquoi, nous privilégions, dans nos travaux, l'explicitation des résultats à l'utilisateur, en détaillant, à sa demande, les concepts qui matchent exactement avec des concepts de la requête, ceux qui sont obtenus par des liens de subsomption ou encore par d'autres liens, en précisant, pour chacun d'eux, la distance sémantique qui les sépare des termes de la requête. Nous reviendrons sur ce point plus loin.

L'expansion de requête peut également avoir des effets négatifs : en incluant plus de termes à la requête, elle accroit le nombre de résultats et peut noyer les documents les plus pertinents dans une trop grande quantité de réponses. Là encore, nous privilégions la justification des résultats pour l'utilisateur : le score global qui est associé à chaque document retourné par le système, est fonction de son indexation et des termes de la requête. Ce score, dépendant des distances sémantiques introduites plus haut et détaillées dans la section 4, permet d'ordonner les documents et de les disposer graphiquement sur une carte conceptuelle où leur position par rapport à la représentation de la requête, traduit leur degré de pertinence par rapport à cette requête.

Si, comme on vient de le voir, l'expansion de la requête présente des avantages, par exemple en évitant les *silences* pour certaines requêtes, certains auteurs mentionnent le fait que l'ensemble des termes de la requête exprimés par un utilisateur peut être considéré comme un seul concept et ils préconisent donc une expansion *prudente* (approche DocTree p. 18 dans (Baziz, 2005) où la requête est envisagée comme le sous-graphe de l'ontologie obtenu en ne conservant que les concepts de la requête et leurs relations). Nous verrons dans la section III.2, que nous avons adopté, au moment de l'appariement entre les documents et les requêtes, une stratégie d'expansion qui tient compte de la structure de l'ontologie de domaine via les mesures de proximités sémantiques qui peuvent y être appliquées.

III.1.2. Ordonnancement des résultats

Le processus de *ranking* ou d'ordonnancement consiste à attribuer à chaque élément du corpus une valeur numérique (*Retrieval Status Value* – RSV) qui reflète son degré de pertinence par rapport à la requête et aux préférences de l'utilisateur. En fonction du modèle d'indexation utilisé, cette fonction d'ordonnancement peut utiliser une approche et des calculs ensemblistes, vectoriels ou probabilistes, pour ordonner les résultats de façon à ce que les documents les plus pertinents pour l'utilisateur soient en tête de liste – chapitre 3 dans (Baeza-Yates & Ribeiro-Neto, 2011). Cette pertinence est difficile à évaluer correctement car elle est fortement dépendante de l'utilisateur, de son contexte de recherche et même du moment où est effectuée cette requête. Or c'est ce calcul de pertinence qui détermine les documents qui seront retournés à l'utilisateur, c'est pourquoi cette étape est critique dans le processus global de RI.

Une fois les RSV attribués à chaque entité du corpus, celles qui ont obtenu un score supérieur à un certain seuil sont présentées à l'utilisateur. Ces entités pertinentes peuvent alors être organisées dans des listes ("à la Google") ou bien positionnées sur des cartes conceptuelles afin de faciliter la

[50] Base de données de publications scientifiques biomédicales, indexées par des termes du MeSH – http://www.ncbi.nlm.nih.gov/pubmed/

[51] Médical Subject Heading – http://www.ncbi.nlm.nih.gov/mesh/

lecture des résultats (ce point sera largement développé dans le chapitre 4). Il est également possible d'appliquer des fonctions de *clustering* afin de regrouper les résultats similaires dans une même *classe*. L'utilisateur a alors une perception des résultats à plusieurs niveaux (*classes* et ressource). Si l'on utilise des techniques de *clustering*, il se peut qu'il y ait certains recouvrements entre les classes. La présentation des résultats sous forme de liste peut alors être plus difficile à déchiffrer. Là encore, la visualisation sur des cartes de connaissance peut fournir des éléments de solution.

Les ressources pertinentes pour l'utilisateur étant identifiées, et les résultats restitués à celui-ci, il reste une phase importante : l'évaluation du système. Elle fait l'objet de la section suivante.

III.1.3. Evaluation des systèmes de recherche d'information

Deux types d'évaluation sont à distinguer : les évaluations basées sur des corpus de tests (*benchmarks*) et utilisant des indicateurs de performance, et les évaluations faites par les utilisateurs eux-mêmes qui estiment leur degré de satisfaction en fonction de différents critères.

Concernant le premier type d'évaluation, les deux principaux indicateurs utilisés pour évaluer les SRI sont la *précision* (proportion de documents pertinents identifiés par le système, par rapport au nombre total de documents sélectionnés) et le *rappel* (proportion de documents pertinents retrouvés par rapport au nombre total de documents pertinents présents dans le corpus) – chapitre 4 de (Baeza-Yates & Ribeiro-Neto, 2011). Pour ce type d'évaluation il est donc nécessaire de disposer i) d'un corpus indexé, ii) d'un ensemble de requêtes écrites par des utilisateurs potentiels, et iii) pour chacune des requêtes de l'ensemble de documents jugés pertinents par des experts du domaine. Dans nos travaux nous avons utilisé le corpus test MuchMore[52]. Des détails sont présentés dans (M.-F. Sy et al., 2011) concernant l'évaluation des performances d'OBIRS et dans (M. F. Sy et al., 2012) concernant l'évaluation d'une extension d'OBIRS intégrant la reformulation : OBIRS-*feedback*. MuchMore est un corpus de test (*benchmark*) regroupant des résumés de documents scientifiques dans le domaine biomédical. Chaque résumé est indexé par un ensemble de concepts de l'ontologie MeSH (*Medical Subject Headings*). Des requêtes, exprimées par des listes de concepts de la même ontologie, sont disponibles ainsi que les documents pertinents leur correspondant. La méthodologie d'évaluation choisie suivait le protocole TREC (Voorhees, 1994). A l'aide de ce benchmark, nous avons étudié l'influence de différents paramètres de notre approche sur la qualité des résultats.

Ce genre d'évaluation permet de comparer les différents SRI entre eux, mais ne peut pas être mise en place dans le cas d'une recherche ouverte, sur le Web, par exemple. De plus ce genre d'évaluation ne permet pas d'apprécier l'ergonomie et l'interaction IRS-utilisateur qui constitue des éléments incontournables de l'automatisation cognitive.

Le second type d'évaluation (par des utilisateurs) est donc également nécessaire. Au-delà des seuls critères de rappel et de précision, la satisfaction de l'utilisateur doit également être évaluée (c.f. point iii, section III.1). Plusieurs facteurs peuvent être pris en compte : réactivité du système (*latence*), satisfaction de l'utilisateur (convivialité de l'interface, facilité de prise en main, …), qualité de la présentation des résultats, etc. Nos travaux de recherche s'inscrivent dans le cadre de partenariats et les outils développés sont destinés à une communauté d'utilisateurs.

[52] http://muchmore.dfki.de/

Nous souhaitons respecter au mieux les limites cognitives d'un opérateur humain et proposer des outils qui satisfassent le plus grand nombre. C'est pourquoi des études ont été menées en particulier avec des chercheurs de l'Inserm dans le cadre des projets liés aux plateformes collaboratives de l'AvieSan (S. Ranwez et al., 2013; M.-F. Sy et al., 2011). Dans ces deux publications, le lecteur pourra trouver des cas d'études où les outils ont été testés par des experts. Actuellement, un ingénieur de recherche effectue le transfert de technologie visant à intégrer les fonctionnalités du prototype OBIRS sur la plateforme AvieSan *Cancer*. Par ailleurs, un corpus de plus de 30 000 publications scientifiques liées au cancer est en cours d'indexation et d'intégration sur cette même plateforme. Nous envisageons donc des tests de notre SRI à plus grande échelle dans les mois à venir, ce qui permettra de raffiner le modèle et d'ajuster l'interface aux besoins des utilisateurs.

III.2. La recherche d'information conceptuelle

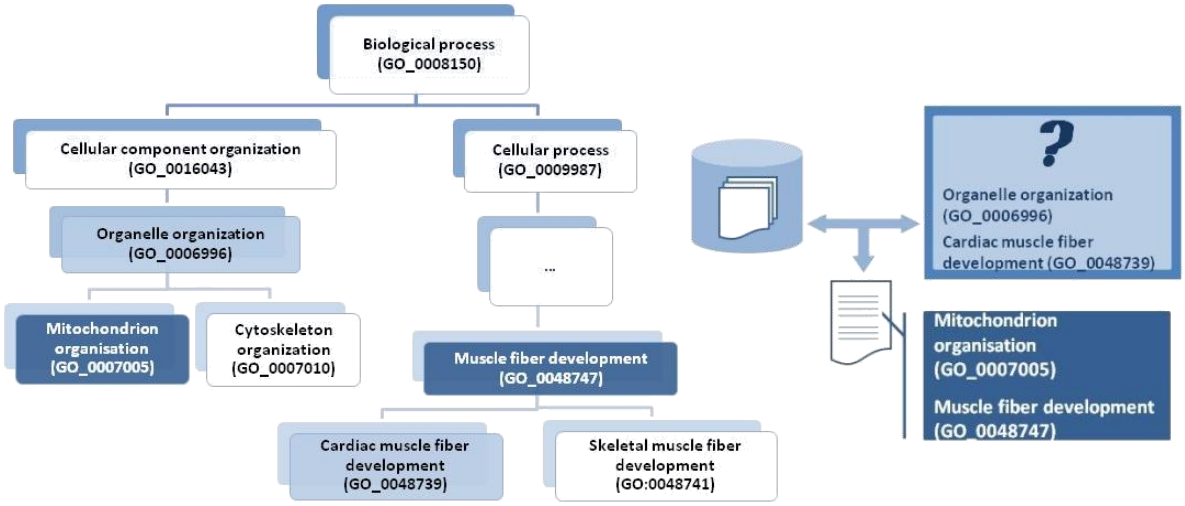
Dans le chapitre précédent, nous avons longuement discuté des avantages d'une indexation conceptuelle. Tout ce qui a été dit alors trouve son corollaire dans les techniques de recherche d'information qui peuvent être appliquées sur une telle indexation. La plupart des SRI considèrent une indexation, tant des documents que des requêtes, par sac de termes (mots-clés), éventuellement pondérés. Nous l'avons dit, de tels systèmes peuvent être pénalisés par l'ambiguïté portée par les termes (e.g. *homonymie*) et par l'absence de prise en compte des relations qu'ils peuvent avoir avec d'autres termes (e.g. *synonymie* ou *hyperonymie*) (Baziz, Boughanem, Pasi, & Prade, 2005). Le traitement des requêtes au niveau conceptuel s'impose donc pour pallier ces limites (Haav & Lubi, 2001). Les structures conceptuelles qui sont utilisées vont des dictionnaires aux ontologies (e.g. Gene Ontology), en passant par les thésaurus (WordNet, UMLS). S'il est désormais admis que leur utilisation augmente significativement la performance des SRI (Andreasen, Bulskov, & Knappe, 2003), il reste une marge d'amélioration puisque la plupart des ontologies ne sont pas conçues et optimisées dans ce but (Jimeno-Yepes, Berlanga-Llavori, & Rebholz-Schuhmann, 2010). Un état de l'art plus complet des SRI est fourni dans (Haav & Lubi, 2001).

Un des écueils de la recherche d'information concerne la (re)formulation de la requête. Les utilisateurs sont, en effet, souvent étrangers aux technologies mises en œuvre dans le processus de recherche et ont du mal à synthétiser et à exprimer leur besoin en terme de requête formulée de manière optimale pour le SRI. Une autre limite est qu'ils ne connaissent pas le corpus, dans l'exemple de la Figure 9, le problème n'est pas la formulation de la requête, mais qu'en l'absence de documents satisfaisant pleinement cette requête, on veut l'élargir. La richesse du corpus sur une question mène à la reformulation : le fait de n'avoir pas assez de documents nécessite un élargissement de la requête, le fait d'en avoir trop nécessite de la préciser. A force de pratique, les utilisateurs adaptent cependant leur formulation, à la lumière d'expériences passées, pour exprimer au mieux leurs attentes. Cela impose donc une première formulation de requête, puis, au vu des résultats obtenus, une reformulation afin de mieux exprimer leurs besoins. Parfois le système effectue lui-même cette reformulation pour peu que l'utilisateur lui signale les résultats qui lui semblent pertinents. On parle alors de *relevance feedback*. Anticipant ce besoin de reformulation, certains systèmes n'attendent pas le retour des utilisateurs pour étendre la requête en y incluant des termes jugés pertinents par rapport à ceux qui sont présents dans la requête initiale. On parle dans ce cas d'*expansion de requête*.

III.2.1. Utiliser la structure sémantique des ontologies pour étendre la requête

La recherche d'information *conceptuelle* est basée sur un modèle conceptuel, i.e. une ou plusieurs ontologies de domaine. On se place ici dans un contexte où les entités manipulées ont été au préalable indexées par un ensemble de concepts de ce modèle (c.f. chapitre 2). La requête est également constituée d'un ensemble de concepts. L'objectif de la RI est d'apparier la requête avec les indexations des différentes entités afin de les classer par ordre de pertinence.

L'utilisation d'ontologies permet, nous l'avons vu, de désambiguïser les requêtes de l'utilisateur. Mais au moment de l'appariement entre requêtes et documents, on peut également utiliser la structure sémantique de l'ontologie pour compléter la requête de l'utilisateur et affiner les résultats. Supposons que nous nous trouvions dans le cas illustré par la Figure 9 : alors que la requête de l'utilisateur comporte les deux concepts "*Organelle organization (GO_0006996)*" et "*Cardiac muscle fiber development (GO_0048739)*", aucune entité du corpus ne contient un de ces deux concepts dans son annotation. Or, il existe une entité (un gène en l'occurrence), qui a été annotée par les concepts "*Mitochondrion organisation*" et "*Muscle fiber development*" qui sont respectivement un *fils* et un *père* des concepts de la requête. Il paraît donc pertinent, au lieu de ne rien renvoyer dans la liste des résultats, de lui présenter celui-ci, en précisant bien pourquoi il a été sélectionné (justification de la sélection).



A) B)

Figure 9 : Utiliser une ontologie pour éviter les *silences* dans un système basé sur une ontologie de domaine (extrait de la Gene Ontology – A) et un corpus indexé, interrogé par un ensemble de concepts (exemple de requête et de ressource indexée, ici, un gène – B).

Comme nous l'avons dit plus haut, l'expansion de la requête doit se faire *prudemment*. Aussi, nous utilisons une mesure de distance entre concepts, afin de n'inclure dans l'ensemble de concepts de la requête, que des concepts *proches* des concepts initiaux. Cela permet de pénaliser le RSV des documents en fonction de leur éloignement à la requête, et ainsi de garder en tête de liste les entités indexées exactement par les concepts de la requête tout en en proposant d'autres. Pour estimer cette proximité, l'utilisation de mesures de similarité sémantique appropriées est nécessaire. Ces mesures seront discutées plus loin.

En ajoutant des termes à ceux de la requête exprimée par l'utilisateur, il devient nécessaire de justifier la sélection effectuée. En effet, nous l'avons dit, l'utilisateur est au centre de notre démarche d'automatisation cognitive et nous souhaitons lui rendre un résultat aussi détaillé et compréhensible que possible. Ces éléments sont évidemment pris en compte dans la restitution des résultats (visualisation détaillée dans le chapitre 4), mais ils sont également à l'origine du modèle d'agrégation que nous avons mis en place pour le calcul du RSV.

Nous avons proposé un modèle de pertinence utilisant les proximités sémantiques pour attribuer les RSV aux documents du corpus (S. Ranwez et al., 2010; M.-F. Sy et al., 2011). Dans le cas de recherche conceptuelle, le RSV permet de classer les documents de telle façon que ceux qui sont indexés par des termes exacts de la requête ont un meilleur classement que ceux qui ont été indexés par des hyperonymes, des hyponymes ou encore des concepts "cousins" de ceux de la

requête. Considérant une requête Q représentée par son modèle conceptuel (ensemble des concepts $C(Q) = \{C_1, C_2, \dots, C_n\}$) et un document d également représenté par un ensemble de concepts issus de la même ontologie $C(d) = \{C'_1, C'_2, \dots\ C'_m\}$, ce calcul se décompose en trois étapes.

Dans un premier temps on évalue la proximité π d'un concept de la requête avec un concept annotant le document (différentes mesures peuvent être employées que nous détaillons dans la section III.4). Ensuite, on évalue la proximité de ce concept avec l'indexation du document (ensemble des concepts qui l'annotent). Pour cela une agrégation des proximités π est utilisée. Dans notre cas, nous avons choisi la stratégie de *best match*, en utilisant l'opérateur *max*, considérant que dès lors qu'un concept est représenté dans l'indexation d'un document il traduit l'intérêt du document par rapport à ce concept. Si d'autres concepts apparaissent dans cette indexation (montrant que le document traite également d'autres thèmes), cela ne doit pas pénaliser le document par rapport à ce concept de la requête.

$$\pi(C_x, d) = \underset{C'_y \in C(d)}{agreg}\left(\pi(C_x, C'_y)\right) = \underset{C'_y \in C(d)}{max}\ \pi(C_x, C'_y)$$

Enfin, pour évaluer l'adéquation entre un document et une requête, nous nous sommes inspirés de la famille des opérateurs de Hölder, et plus particulièrement de l'opérateur d'agrégation proposé par Yager (Yager, 1979) :

$$RSV(Q, d) = \underset{C_x \in C(Q)}{agreg}\left(\pi(C_x, d)\right) = \left(\frac{\left(\sum_{x=1}^{|Q|} \pi(C_x, d)^q\right)}{|Q|}\right)^{1/q}, q \in \mathbb{R}$$

Chaque concept C_x d'une requête Q étant considéré comme un "critère", il s'agit de considérer les documents du corpus comme des alternatives pour lesquelles l'évaluation par rapport aux critères est donnée par $\pi(C_x, d)$.

Dans le cas de requêtes effectuées sur le Web, il a été montré que les opérateurs appliqués aux termes de la requête (AND, OR, NOT) sont rarement utilisés (B J Jansen, Spink, & Saracevic, 2000) et, quand ils le sont, c'est souvent mal à propos (Bernard J Jansen, 2000; Lucas & Topi, 2004). Cette constatation peut également s'appliquer à de nombreux outils de RI pour lesquels il est difficile d'exprimer ces opérateurs. Or le mode de calcul du RSV décrit ci-dessus permet de faire intervenir la préférence de l'utilisateur concernant le type d'agrégation choisi, en faisant simplement varier le paramètre q. Ce paramètre permet de spécifier l'exigence plus ou moins prononcée de la satisfaction simultanée de tous les critères, i.e. la contrainte plus ou moins nécessaire que D_i soit sémantiquement proche de tous les concepts de la requête. En effet, si q tend vers -∞, la requête tend à être conjonctive (l'agrégation tend vers le *min*, équivalent à un AND entre les concepts) ; lorsque q tend vers +∞ elle tend à être disjonctive (l'agrégation tend vers le *max* équivalent à un OR entre les concepts) ; enfin lorsque q tend vers 0, le calcul se rapproche de la moyenne géométrique. Ce contrôle est particulièrement intéressant car il permet, en ne jouant que sur un seul paramètre (qui dans notre cas est représenté par un curseur), de modifier le type d'agrégation.

Suivant ce modèle, un système de recherche d'information basée sur une ontologie a été développé, OBIRS, dont nous avons déjà parlé. L'application disponible en ligne concerne la recherche de gènes, indexés avec des concepts de la *Gene Ontology* (GO). Nous avons évalué cet environnement en collaboration avec Armelle Regnault, chercheur à l'Inserm et responsable de la coordination de l'ITMO IHP (Immunologie, Hématologie et Pneumologie). Cette collaboration a permis, entre autres, d'améliorer sensiblement l'interface graphique. Une utilisation à plus grande échelle nous permettrait de poursuivre ces améliorations. C'est ce que nous espérons au travers de la mise à disposition de l'outil sur d'autres plateformes collaboratives dont nous avons la charge (e.g. ITMO Cancer). Dans ce but, nous avons également adapté cet outil pour la recherche de publications scientifiques indexées avec le MeSH (*Medical Subject Headings*) (S. Ranwez et al., 2013).

III.2.2. Utiliser l'ontologie pour la reformulation

Contrairement à l'expansion de requête qui est réalisée en amont du processus d'appariement, la reformulation consiste à produire une nouvelle requête en fonction des résultats fournis par le système. Durant cette reformulation, la requête peut être modifiée en ajoutant de nouveaux concepts, en supprimant des concepts *pauvres* (non discriminants) ou encore en modifiant les poids qui leur sont associés. De nombreuses techniques de reformulation ont été proposées dans la littérature, dont la plus répandue est sans doute le retour de pertinence (*relevance feedback*) (Crouch J. & Yang, 1992). Cette technique se base sur les documents jugés pertinents par l'utilisateur parmi ceux proposés en réponse à une requête initiale et propose une nouvelle requête, utilisant les techniques d'expansion et de modification des poids (Abdelali, Cowie, & Soliman, 2007).

Le premier à avoir proposé un modèle pour la reformulation en recherche d'information est Rocchio (Rocchio, 1971). Dans son approche il utilise les documents jugés pertinents par l'utilisateur, en réponse à une requête donnée, pour formuler une nouvelle requête qui tienne compte de la requête initiale, et qui favorise des documents proches de ceux jugés pertinents parmi les résultats obtenus et défavorise ceux qui sont proches des documents jugés non pertinents. Si elle a démontré son efficacité, cette approche, dite *explicite* car faisant intervenir directement le jugement de l'utilisateur, peut être fastidieuse pour ce dernier. Même si on peut essayer de mutualiser les efforts en prenant en compte différents utilisateurs de même profil (ce qui a plu à certaines personnes devrait plaire à d'autres qui sont dans un contexte similaire), l'interaction de l'utilisateur reste souvent un facteur limitant. D'autres approches, dites *implicites* ont donc été proposées, dans lesquelles le système déduit de la liste des résultats, ceux qui peuvent être pris en compte dans la reformulation. Pour ce faire, on peut se référer aux documents qui ont été consultés par l'utilisateur, ou bien se baser uniquement sur les premiers documents de la liste (cette dernière solution nécessite une certaine confiance dans les résultats initiaux) – on parle alors d'analyse *locale*. Alternativement, la reformulation peut utiliser une ressource externe (thésaurus, ontologie, etc.) couplée à une analyse lexicale du corpus et des relations entre termes qu'il contient – on parle d'analyse *globale*. Ces dernières techniques transparentes pour l'utilisateur (n'augmentent pas sa charge cognitive) et sont donc d'un grand intérêt. Cependant elles sont moins précises que les approches explicites.

Dans nos travaux, nous utilisons une méthode *locale* de retour de pertinence explicite (*relevance feedback*) pour compléter et reformuler la requête de l'utilisateur (M. F. Sy et al., 2012). Celui-ci doit sélectionner les documents qu'il juge pertinents parmi ceux qui lui ont été fournis en résultat de sa requête et notre SRI, OBIRS, utilise l'ontologie de domaine pour compléter la requête initiale (liste de concepts) afin de trouver d'autres documents proches de ceux sélectionnés. La reformulation est, ici, envisagée comme la recherche d'une nouvelle requête maximisant un indicateur de performance. Si les approches vectorielles utilisent de tels indicateurs depuis longtemps (Rocchio, 1971), leur transposition à la reformulation conceptuelle n'avait, à notre connaissance, pas encore été explorée. Les solutions utilisant une ressource sémantique se contentent souvent de rajouter des concepts en relation (synonymie par exemple) avec ceux de la requête initiale (Baziz, Aussenac-Gilles, & Boughanem, 2003; Voorhees, 1994).

Nous envisageons la reformulation comme la recherche d'une nouvelle requête maximisant un indicateur de performance. Elle est donc vue comme un problème d'optimisation. Différents indicateurs sont discutés dans (M. F. Sy et al., 2012) principalement basés sur les notions de *précision* et de *rappel*. Si D_s est la liste des documents présentés à l'utilisateur par le SRI suite à la requête Q_{init}, les nouveaux résultats attendus doivent se rapprocher des documents D_u jugés pertinents par l'utilisateur ($D_u \subseteq D_s$). Nous y intégrons une reformulation négative, de manière à ce que les résultats proposés s'éloignent des documents implicitement jugés non pertinents ($D_s \backslash D_u$). L'idée est de trouver une requête Q qui maximise l'indicateur ind ci-dessous :

$$ind(Q, D_u, D_s) = \alpha\, RSV(Q, Q_{init}) + \beta \underset{d_i \in D_u}{agreg} \left(RSV(Q, d_i)\right) - \gamma \underset{d_j \in D_s \backslash D_u}{agreg'} \left(RSV(Q, d_j)\right)$$

$agreg$ et $agreg'$ étant des fonctions d'agrégation, potentiellement de nature différente, en fonction de la stratégie de reformulation choisie. α, β et γ sont trois paramètres compris dans l'intervalle $[0,1]$ qui permettent d'attribuer plus ou moins d'importance à la requête initiale, aux documents jugés pertinents ou au contraire aux résultats jugés non pertinents. $RSV(Q, d_i)$ est le score de pertinence du document d_i par rapport à la requête Q. A l'aide du benchmark MuchMore, nous avons étudié l'influence de différents paramètres, en particulier le choix de la mesure de similarité utilisée dans la recherche d'information ainsi que les opérateurs d'agrégation choisis (valeurs du paramètre q) et l'influence du paramètre γ dans la reformulation.

Notre contribution concerne non seulement les indicateurs de performance que nous utilisons pour ce calcul, mais également des résultats théoriques sur lesquels s'appuient des solutions heuristiques qui permettent des traitements rapides, même dans le cas d'ontologies contenant un très grand nombre de concepts. Une solution exacte nécessiterait de tester toutes les requêtes possibles (i.e. tous les sous-ensembles de concepts de l'ontologie). Ceci n'est évidemment pas possible. Le principe de base de l'heuristique que nous utilisons pour cette reformulation de requête consiste en une approche gloutonne qui ajoute à la requête courante les concepts de l'ontologie qui améliorent le plus la fonction *objectif*. Même ainsi, cela resterait trop coûteux en temps de calcul car il faudrait tester chaque concept de l'ontologie pour vérifier s'il améliore ou non la requête courante. Or, comme nous l'expliquons dans (M. F. Sy et al., n.d.), il n'est pas nécessaire de tester tous les concepts de l'ontologie, mais seulement ceux proches des concepts présents dans l'indexation des documents sélectionnés (ceux de D_u). Dans ce même article, nous prouvons que plusieurs mesures sémantiques permettent d'identifier cet ensemble de concepts à tester en un temps proportionnel à sa taille. On restreint donc drastiquement l'espace des possibles et les temps de calcul en sont d'autant plus rapides.

III.3. Où approches conceptuelles et lexicales peuvent se compléter

Comme en témoigne l'ajout de chapitres dédiés à la recherche d'information à partir de texte dans la seconde édition de (Baeza-Yates & Ribeiro-Neto, 2011), l'analyse lexicale et les nombreux travaux qui y affèrent sont d'un grand intérêt en RI. En effet, si les approches sémantiques sont reconnues pour améliorer les systèmes de recherche d'information, les indexations qu'elles utilisent ne couvrent qu'une partie du contenu des documents qu'elles décrivent, ce qui ne fait qu'accentuer le *gap sémantique*. Une partie de la communauté d'ingénierie des connaissances s'intéresse à la complémentarité des approches lexicales et conceptuelles pour enrichir l'indexation (à rapprocher de la notion de *peuplement* ontologique). Dans notre cas nous l'avons également éprouvé pour améliorer la justification des résultats. Ce sont ces deux points qui sont discutés dans cette section.

III.3.1. Analyse lexicale pour compléter l'indexation

Appelés à des postes de gouvernance de grand consortiums reconnus au niveau national et international, plusieurs chargés d'affaires, directeurs ou managers se trouvent confrontés à une surcharge cognitive due, en grande partie, à la gestion des connaissances propres à ces regroupements. Ces connaissances sont multiples : il peut s'agir de documents scientifiques, de données de gestion, de fiches personnelles décrivant les activités de chaque membre ou bien de ressources plus confuses représentant des équipes de recherche, des laboratoires, des services… Difficile (impossible ?) de maîtriser l'ensemble de ces informations et d'en avoir une vision globale. C'est le constat qu'ont fait certains de nos partenaires ayant à gérer les instituts thématiques multi-organismes (ITMO) de l'AvieSan. Dans ce cas, des outils capables de rechercher l'information, de la synthétiser et d'en présenter une vue globale sont indispensables. En ce qui concerne les ITMO, les ressources sont produites par des communautés différentes car provenant de plusieurs institutions, de personnes ayant des profils et des domaines d'expertise spécifiques. Ces ressources peuvent avoir été indexées au préalable, comme c'est le cas pour les publications scientifiques liées au cancer qui sont indexées dans PubMed en utilisant des concepts du MeSH, mais un besoin a été identifié d'utiliser

également des ressources *ad hoc*, comme par exemple la classification CSO (Common Scientific Outline) pour catégoriser des projets liés au cancer. Ce processus de classification peut être très personnel et ne prend pas toujours en compte l'indexation préexistante des ressources.

Pourtant parfois, il peut s'avérer utile d'établir des relations entre les deux indexations, par exemple pour identifier la bibliographie associée à un projet de recherche et donc les experts qui peuvent être sollicités dans le cadre de ce projet et éventuellement identifier certaines complémentarités entre des équipes, des personnes.

Nous avons travaillé dans ce sens au cours des derniers mois (Post-doc de François-Elie Calvier, 2012/2013) et proposé une méthode de classification de documents textuels, à partir d'un apprentissage basé sur une ontologie de domaine. Ces travaux ont fait l'objet d'une publication dans une conférence (Calvier, Plantié, Dray, & Ranwez, 2013), et nous en exposons ici les grandes lignes.

L'objectif principal de cette approche est de minimiser l'intervention humaine dans la construction du modèle de classification tout en respectant certains processus logiques propre à chaque utilisateur.

Le processus de classification débute par une phase d'apprentissage. Un utilisateur-expert ayant défini des catégories (par exemple des catégories du CSO), associe une liste de documents à chacune de ces catégories. Le système analyse les index conceptuels (sacs de concepts issus d'une ontologie de domaine, par exemple le MeSH) associés à chaque document. Cette indexation conceptuelle est supposée fiable. A partir de l'analyse des index, il est possible d'identifier des concepts qui caractérisent les différentes catégories et ainsi modéliser le système de classification de l'expert. Quand un nouveau document doit être classé, on peut trouver la ou les catégories auxquelles il peut être rattaché, en fonction de son indexation conceptuelle. Pour cela, une méthode d'agrégation particulière a été imaginée pour caractériser les concepts qui décrivent une catégorie, en fonction des concepts qui indexent les documents appartenant à cette catégorie. Bien sûr, lors de la phase de classification, des mesures de similarité sémantique doivent être utilisées pour estimer la proximité entre les concepts indexant le nouveau document et les concepts caractérisant chaque catégorie. L'originalité de l'approche que nous avons proposée réside dans le classement dans plusieurs catégories (appartenance *floue*).

Le système de recherche d'information peut dès lors utiliser l'une ou l'autre ou les deux indexations : l'indexation conceptuelle basée sur l'ontologie ou la classification en différentes catégories.

III.3.2. Analyse lexicale pour une meilleure compréhension des résultats

Les approches lexicales sont particulièrement adaptées à un contexte de recherche d'information *ouvert*, contenant des documents hétérogènes et non-structurés (e.g. Web) tandis que les approches sémantiques, reposant sur des ontologies de domaine, nécessitent une formalisation stricte de la connaissance du domaine et une description conceptuelle des éléments du corpus. Ces indexations se situent principalement au niveau du document (elles synthétisent l'ensemble du contenu) alors que les approches lexicales permettent une analyse plus fine, en descendant au niveau du paragraphe, voire de la phrase. Il semble donc nécessaire d'envisager la recherche d'information comme un continuum des solutions terminologiques jusqu'aux approches ontologiques. Si auparavant ces approches étaient considérées comme concurrentes et exclusives, il paraît indispensable aujourd'hui de les rapprocher pour tirer parti de leurs avantages respectifs (Prévot, Borgo, & Oltramari, 2010). Des solutions hybrides sont à imaginer. Nous avons proposé dans (S. Ranwez et al., 2013) une architecture, CoLexIR[53], qui utilise l'analyse lexicale en complément de la RI conceptuelle. Le schéma global est présenté dans la Figure 10.

[53] Conceptual and Lexical Information Retrieval

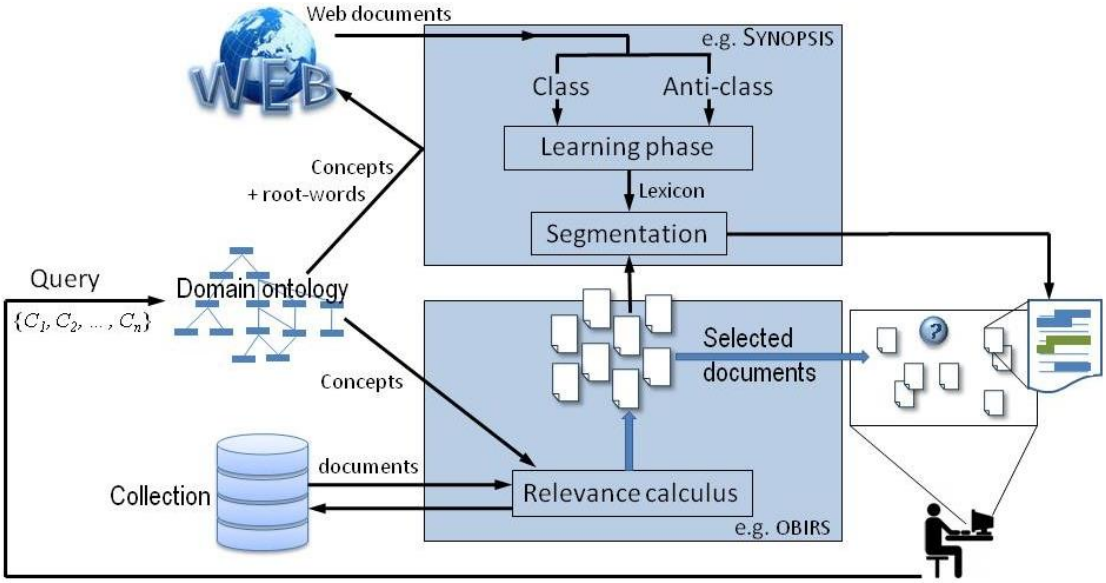


Figure 10 : Schéma global de l'approche CoLexIR

Cette application, bien sûr, n'est mise en œuvre que pour la recherche de documents textuels. L'idée sous-jacente est de rechercher ces textes par une méthode conceptuelle classique (par exemple en utilisant OBIRS), puis, pour chaque document retrouvé, d'utiliser une analyse lexicale – dans notre cas, il s'agit de *Synopsis* (Duthil, Roche, Montmain, & Poncelet, 2011), pour segmenter le texte et identifier les passages qui correspondent à la demande de l'utilisateur (cadre D dans la Figure 11). Ainsi, le système justifie les résultats à l'utilisateur, et l'aide à les appréhender plus efficacement. Une étape d'apprentissage est nécessaire pour construire un lexique associé à chaque concept de l'ontologie. Cette étape peut être automatisée en exploitant la structure de l'ontologie et les labels des concepts comme *root-words* pour construire les requêtes Web nécessaires à cet apprentissage (c.f. Figure 10). Elle reste néanmoins coûteuse en temps de calcul. Cependant, elle est effectuée une fois pour toute et l'utilisation de ces lexiques par la suite, peut se faire en temps réel. Les segments identifiés comme étant relatifs à certains concepts, ne contiennent pas nécessairement des labels associés à ces concepts mais un sous-ensemble des termes qui ont pu leur être associés pendant la phase d'apprentissage.

Pour ce travail, nous avons collaboré avec Benjamin Duthil, étudiant en thèse au LGI2P, pour la conception et le développement de la solution, et avec Patrick Augereau de l'unité Inserm IRCM[54] de Val d'Aurelle, pour la partie évaluation. Nous avons demandé à cet expert du domaine d'utiliser le SRI pour fouiller des publications scientifiques liées au Cancer (publications de BMC Cancer de 2007 à 2011) et indexées par le MeSH. L'utilisation visée était l'assistance à la recherche bibliographique. Si les résultats étaient positifs, ils ont également soulevé certaines limites. En particulier, il semblerait également intéressant de tenir compte de la structure des documents lors de la sélection des extraits de l'article, un passage de la partie "Discussion" étant souvent jugé plus informatif qu'un passage de la section "Matériel et Méthode".

[54] Institut de Recherche en Cancérologie de Montpellier

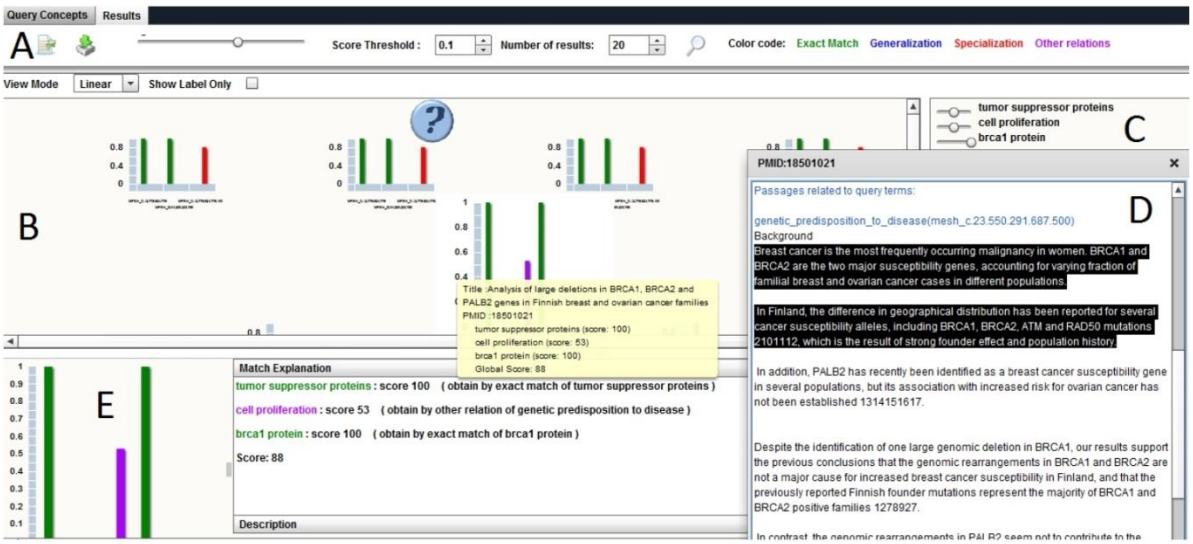


Figure 11 : Interface de CoLexIR

Ce travail s'inscrit pleinement dans notre démarche d'automatisation cognitive. En effet notre objectif est de respecter les limites cognitives de l'opérateur humain et de lui permettre, le plus intuitivement possible, d'appréhender les résultats dans leur globalité et de se focaliser, si besoin, sur des parties spécifiques. L'interactivité avec l'environnement est donc primordiale. Ce point sera développé au chapitre suivant.

III.4. Les mesures de similarité sémantique au cœur de la RI

Que ce soit au moment de l'appariement entre requête et documents (e.g. calcul de π dans la section III.1.2) ou au moment de la reformulation (c.f. section III.2.2), l'estimation du RSV est fortement dépendante de la mesure de similarité sémantique employée. Ces mesures peuvent également intervenir soit pour évaluer la ressemblance (sémantiquement parlant) entre deux documents (utilisés pour la recherche de documents complémentaires, par exemple). Nous y avons donc consacré une partie de nos recherches.

On notera que certaines de ces mesures satisfont les axiomes de *distance*, tandis que d'autres évaluent la *proximité* (voisinage de champ lexical : par exemple une *tasse* est proche de *café*) ou la *similarité* (une *tasse* est un concept proche de *bol*). C'est pourquoi nous englobons toutes ces notions dans le terme *mesures de similarité*.

Il est possible de distinguer deux types de mesures suivant qu'elles sont définies par intention, i.e. basées sur la structure de la connaissance du domaine, ou par extension, i.e. basées sur une analyse statistique du corpus sur lequel elles sont employées.

III.4.1. Mesures de similarité sémantique par intention

Ces mesures utilisent le graphe sémantique (réseau sémantique) associé à la représentation de la connaissance du domaine (e.g. ontologie) comme espace métrique (Resnik, 1999). Si cette représentation peut inclure différents types de relations entre les concepts, la plupart de ces mesures se focalisent sur la relation de subsomption (*is_a*) (Rada, Mili, Bicknell, & Blettner, 1989). Cela s'explique d'une part parce que cette relation est la seule qui est toujours présente dans une ontologie (c.f. la définition formelle des ontologies présentée dans (Maedche & Staab, 2001)), et d'autre part par la sémantique particulière qu'elle exprime et les propriétés qui en découlent (e.g. orientation, transitivité, …)

La réduction d'une ontologie $\mathcal{O}$ à la seule relation *is-a* constitue un graphe orienté sans circuit dans lequel les nœuds représentent les concepts de $\mathcal{O}$ et les arcs les relations *is-a* entre ces concepts. Ce graphe ayant des propriétés particulières, de nombreuses mesures de similarité sémantique utilisent cette réduction. La mesure la plus naïve consiste à définir la proximité de deux concepts comme la longueur du plus court chemin entre les nœuds correspondants (Rada et al., 1989). Cette mesure a ensuite été complétée par (Hirst & St Onge, 1998) pour tenir compte du nombre de changements de direction dans le graphe. Avec cette variante, dans l'exemple de la Figure 1 du chapitre 1(ontologie de la musique), *Cor* est plus proche d'*Instrument* (chemin de longueur 2 sans changement de direction) que de *Trompette* (chemin également de longueur 2 mais avec un changement de direction). De nombreuses mesures ont été proposées à partir de celles-ci, dont beaucoup utilisent la recherche du plus petit ancêtre commun entre deux concepts. Cependant, ces mesures supposent pour la plupart une homogénéité des arcs entre les concepts. Or en fonction du degré de profondeur dans le graphe induit par la relation *is_a*, une relation entre deux concepts ne traduira pas le même degré de spécialisation. Toujours dans l'exemple de la Figure 1 du chapitre 1, l'arc entre *Cor* et *Cuivre* représente une variation de spécialisation plus faible que celui entre *Instrument* et *Objet*. De plus ces mesures ne respectent pas les propriétés des distances, c'est ce qui nous a poussés à proposer une nouvelle mesure qui respectent les propriétés de *symétrie*, de *triangularité* et de *positivité* (S. Ranwez et al., 2006). Cette mesure a la particularité de considérer non seulement les ancêtres, mais également les descendants communs à deux concepts. Cela répond à l'intuition selon laquelle au plus deux concepts partagent des descendants, au plus ils sont proches.

III.4.2. Mesures de similarité sémantique par extension

Ces mesures sont basées sur une analyse statistique de l'occurrence des termes dans les différents documents présents dans le corpus. La plupart de ces mesures utilisent l'IC (Information Content) défini dans (Resnik, 1999) : $\mathrm{IC(c)} = -\log(\mathrm{P(c)})$, avec $\mathrm{P(c)}$ la probabilité d'apparition du concept c ou d'un de ses descendants dans l'ensemble des indexations des documents du corpus (on note $hypo(c)$ l'ensemble des hyponymes de c, y compris lui-même). Cette probabilité est d'autant plus grande que le nombre de descendants du concept considéré est grand. Le degré d'expressivité d'un concept est donc inversement proportionnel à son nombre de descendants. Sur la base de cette remarque, et pour faire abstraction du corpus au moment des calculs, (Lin, 1998) propose une définition de l'IC exploitant la structure de l'ontologie utilisée pour définir une distance sémantique :

$$IC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(|C|)}$$

$$\pi_{lin}(c_1, c_2) = \frac{2 * IC\,(\,MICA(c_1, c_2)\,)}{IC(c_1) + IC(c_2)}$$

avec $|C|$ étant le nombre de concepts présents dans l'ontologie et MICA l'ancêtre commun à c_1 et à c_2 le plus informatif (*most informative common ancestor*).

Il existe dans la littérature de nombreuses mesures de similarités basées sur ces estimations du contenu informationnel. Les répertorier toutes sortirait du cadre de ce mémoire. Un état de l'art complet est présenté dans (Harispe, Ranwez, Janaqi, & Montmain, n.d.).

III.4.3. Propriétés et choix des mesures de similarité sémantique

Qu'elles soient définies par extension ou par intension, on peut distinguer deux niveaux de calcul : les mesures dites *par paires* (*pairwise*) qui évaluent le degré de similarité entre deux concepts, et les mesures *par groupes* (*groupwise*) qui évaluent des similarités entre des ensembles de

concepts. Ces dernières sont utilisées dans le domaine biomédical, où, par exemple, des mesures de similarité entre l'indexation de différents gènes permettent de déterminer ceux impliqués dans un même processus biologique facilitant ainsi la prédiction d'interactions entre protéines (Pesquita & al., 2009).

De par leur mode de calcul et les spécificités liées à leur mise en œuvre, les mesures de similarité sémantique ont des propriétés différentes. Par exemple, comme elle respecte les propriétés d'une distance (*positivité, symétrie* et *inégalité triangulaire*), nous avons utilisé notre mesure (S. Ranwez et al., 2006) dans le cas de visualisation de données par projection MDS (c.f. chapitre suivant). D'autres mesures cherchent à se rapprocher au maximum de l'avis d'experts (Lee, Shah, Sundlass, & Musen, 2008; Pakhomov et al., 2010). Les mesures peuvent ne pas vérifier la symétrie, mais elles sont pertinentes en recherche d'information. C'est donc l'expressivité qui est évaluée ici. Enfin, si on cherche dans un corpus un document qui *ressemble* à un autre, on ne requiert pas le même type de mesure que lorsqu'on recherche un document par rapport à une requête. En effet, dans un cas la mesure utilisée doit être symétrique (entre documents), alors que dans l'autre cas elle doit être asymétrique : s'il semble raisonnable de pénaliser un document parce qu'un concept d'une requête est absent de son index, l'ignorer parce qu'il est indexé par un concept absent de la même requête ne le serait pas.

Ainsi, on le voit, le choix d'une mesure de similarité est fortement dépendant d'une application et d'un contexte donné. Or il n'existe pas de guide précis pour aiguiller un utilisateur (concepteur d'une application dédiée) vers une mesure ou une autre en fonction de son contexte. C'est pour répondre à ce besoin que nous avons orienté la thèse de Sébastien Harispe sur l'étude de ces mesures et leur classification. Ce travail a conduit à la définition d'un cadre abstrait qui généralise ces mesures, permet de les caractériser par rapport à leurs propriétés et éventuellement d'en définir de nouvelles (Harispe et al., n.d.). Ce cadre abstrait, se base sur les travaux de (Tversky, 1977), et utilise l'estimation de l'information partagée par des concepts (*commonality*) et de leur spécificité (*distinctiveness*). Basée sur ce cadre abstrait, une librairie a été développée, qui permet de créer un grand nombre de benchmarks, de tester les différentes mesures de la littérature et ainsi de sélectionner la meilleure dans un contexte donné. Cette librairie, *Semantic Measures Library*, est accessible en ligne[55].

III.5. Focus sur des algorithmes utilisés dans différentes phases de la RI

Au cours de nos recherches, nous avons été à plusieurs reprises confrontés à des problèmes de temps de calcul lors de traitements sur des ontologies contenant un trop grand nombre de concepts. Nous avons donc proposé des algorithmes efficaces pour effectuer certains de ces traitements.

III.5.1. Calcul des LCA d'un ensemble de concepts

Nous avons vu que de nombreux traitements, en particulier le calcul de mesures de similarité sémantique, sont basés sur le calcul du MICA, ce qui nécessite le calcul du plus petit ancêtre commun (lca pour *least common ancestor*). Cette opération est bien définie et couramment utilisée dans le domaine des graphes orientés sans circuit (*dag* pour *directed acyclic graph*) où plusieurs solutions ont été proposées dans la littérature pour calculer le lca d'un couple de nœuds. Cependant la notion de lca d'un groupe de nœuds peut également s'avérer utile notamment dans le cas de l'extraction de sous-ontologies détaillée dans la section suivante.

Nous avons étendu la définition du lca d'un couple de nœuds donnée dans (Bender, Pemmasani, Skiena, & Sumazin, 2001) et proposé la définition du lca pour un groupe de nœuds dans un dag et nous l'avons appliquée aux ontologies :

[55] http://www.semantic-measures-library.org/sml/

Soit le dag $\mathcal{O}_{isa} = (C, E)$ la réduction d'une ontologie $\mathcal{O}$ à ses seules relations de subsomption, i.e. où C correspond à l'ensemble des concepts de $\mathcal{O}$ et E à l'ensemble des relations de type *is_a* qui les relient. Soient S un sous-ensemble de concepts $S \subseteq C$ et $A(S)$ l'ensemble des ancêtres de S, l'ensemble lca(S) de ce sous-ensemble S est formé par les nœuds u ∈ A(S) dont le degré entrant[56] est nul dans le graphe $H = G[A(S)]$ induit par A(S).

Sur la base de cette définition, nous avons introduit l'opérateur $U_{lca}(S)$, qui calcule l'union des lca de tous les couples contenus dans S. Si l'on considère un ensemble de départ S_0, on peut définir la suite croissante $S_{k+1} = U_{lca}(S_k)$ qui converge au bout d'un certain nombre d'itérations. On définit ainsi la fermeture de S pour l'opérateur *lca*, notée *lca*-fermeture(S). Dans (V. Ranwez, Janaqi, & Ranwez, 2012) nous avons donné deux définitions équivalentes de cette fermeture et proposé un algorithme efficace pour la calculer. On peut définir de manière analogue la *gcd*-fermeture de l'opérateur U_{gcd} (gcd pour *greatest common descendant*). Ces opérateurs sont le fondement de l'outil OntoFocus, que nous avons développé pour l'extraction de sous-ontologies.

III.5.2. Extraction de sous-ontologies

De par le succès du Web sémantique et des techniques associées, de plus en plus d'applications basées sur les ontologies sont disponibles. De nombreuses communautés ont donc choisi de représenter leurs connaissances à l'aide d'une ontologie de domaine, pour servir de guide à l'indexation de documents, assister l'opérateur dans sa navigation ou améliorer la recherche d'information. On peut citer WordNet dans le domaine général, FOAF (*Friend of a Friend*) pour les réseaux sociaux ou, dans le domaine biomédical, les initiatives du groupement OBO (Open Biological and Biomedical Ontologies[57], à l'origine par exemple de la Gene Ontology). Or, dès lors qu'elles sont partagées par une large communauté, les ontologies ont tendance à croître et à rapidement atteindre plusieurs milliers de concepts et tout autant de relations, dépassant ainsi la capacité d'un opérateur humain à les appréhender dans leur globalité. Par ailleurs, certains traitements deviennent trop longs et ne peuvent plus être effectués en temps réel. Nous avons identifié plusieurs cas où il peut être nécessaire de restreindre la taille des ontologies considérées :

- Conception/extension de l'ontologie. Lors de la mise à jour de l'ontologie, il est souvent nécessaire de considérer un sous-ensemble de concepts et d'analyser leurs relations avant d'ajouter un nouveau concept, de modifier une relation ou de réorganiser les concepts.
- Recherche d'information et filtrage basés sur une ontologie. Etant donné l'ensemble des concepts utilisés pour annoter un (ou plusieurs) document(s), l'identification des relations entre ces concepts permet de mieux comprendre le sens de l'annotation.
- Navigation et visualisation supportées par une ontologie. Les limites cognitives et perceptives des opérateurs humains engendrent une forte demande en outils de représentation intuitifs et conviviaux.

En réponse à ce problème, nous avons proposé une méthode pour extraire une sous-ontologie focalisée sur l'ensemble des concepts qui, à un moment donné et pour une application donnée, sont au centre de l'intérêt de l'utilisateur. Cette méthode, appelée OntoFocus, permet d'identifier un ensemble de concepts pertinents pour l'utilisateur et les relations associées, pour lui fournir une sous-ontologie focalisée sur ses centres d'intérêt. Pour réaliser cette extraction, on ne peut pas se contenter de garder seulement les concepts d'intérêt de l'utilisateur. En effet, pour construire un extrait "porteur de sens", c'est-à-dire suffisamment riche pour qu'un utilisateur comprenne bien les liens entre les différents concepts qu'il a identifiés, il est souvent nécessaire d'y inclure les ancêtres et/ou les descendants de ces concepts. Les inclure tous (comme le fait la base de données Ensembl, par exemple) conduit souvent à une sous-ontologie trop volumineuse et on retombe dans la

[56] Le degré entrant d'un nœud est le nombre de relations ayant ce nœud pour destination.

[57] http://www.obofoundry.org/

problématique liée au nombre de concepts considérés. Cependant seuls un petit nombre d'entre eux jouent un rôle central dans la compréhension des liens entre les concepts d'intérêt et ils peuvent être identifiés grâce aux opérateurs U_{lca} et U_{gcd} définis dans la section précédente (V. Ranwez et al., 2009; V. Ranwez, Ranwez, et al., 2012).

A partir d'une ontologie $\mathcal{O}$ de référence et d'un ensemble de concepts d'intérêt S_i fournis par l'utilisateur, *OntoFocus* calcule sa *lca*-fermeture et sa *gcd*-fermeture dans le graphe orienté acyclique $\mathcal{O}_{isa}$. L'union de ces deux ensembles définit les concepts qui sont conservés pour former la sous-ontologie. Les relations entre ces concepts, sont ensuite ajoutées, à partir des relations présentes dans l'ontologie d'origine en respectant la transitivité des relations.

Nous avons proposé un algorithme pour calculer ces fermetures avec une complexité en temps proportionnelle au nombre de concepts conservés multiplié par le nombre d'arcs de l'ontologie d'origine (V. Ranwez, Janaqi, et al., 2012). Cela permet d'envisager son utilisation dans des applications en temps réel ce qui n'était souvent pas possible avec les solutions existantes, plus gourmandes en ressources. On peut donc envisager d'utiliser *OntoFocus* pour créer de nouvelles sous-ontologies à la demande chaque fois que l'utilisateur a besoin d'une vue spécifique sur l'ontologie de référence, cette dernière demeurant inchangée (c.f. chapitre suivant). Nous l'avons testé sur la Gene Ontology en donnant comme concepts d'intérêt les 50 concepts liés à l'annotation de BRCA1 (gène impliqué dans le cancer du sein). La sous-ontologie résultante contient 92 concepts, ce qui permet à un opérateur humain de les appréhender plus facilement. Nous avons utilisé *OntoFocus* à plus large échelle, pour générer les sous-ontologies des quelques 13 000 gènes de la base de données OrthoMamm-v7[58] (V. Ranwez et al., 2007).

Nous pensons utiliser ce module en particulier pour le filtrage lors de la recherche d'information ou pour la visualisation d'ontologies (c.f. chapitre 4).

L'application *OntoFocus* a été mise en ligne (http://www.ontotoolkit.mines-ales.fr/), et une version téléchargeable sera prochainement mise en ligne.

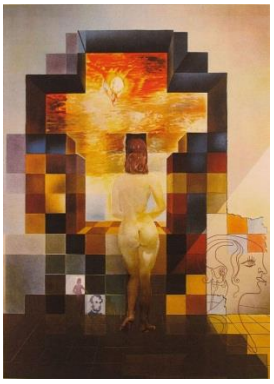
III.6. Synthèse

Dans ce chapitre, nous avons positionné nos travaux par rapport à l'état de l'art en recherche d'information. Nous avons particulièrement insisté sur les apports de l'expansion de requête et de la reformulation. Au travers *d'OBIRS*, *OBIRS-feedback* et *CoLexIR*, nous avons, jusqu'à présent, envisagé une recherche d'information *classique*, où l'indexation de chaque document du corpus est comparée avec la requête de l'utilisateur, afin d'ordonner les documents pour ne présenter à l'utilisateur que les plus pertinents.

Nos perspectives de recherche, à la suite de ces travaux, concernent la recherche d'informations complémentaires. En effet, si à l'heure actuelle, la plupart des moteurs de recherche d'information tentent de trouver des documents qui ressemblent aux documents jugés pertinents par l'utilisateur, il est des situations où l'on peut vouloir trouver un ensemble de documents qui, de par leur complémentarité, couvrent l'ensemble des termes de la requête. C'est le sujet de la thèse qui est proposée à Nicolas Fiorini (commencement de la thèse en octobre 2012). Cette notion se rapproche de la notion de diversification des résultats d'une requête (*query results diversification*) (Minack, Demartini, & Nejdl, 2009), mais s'en distingue car si la diversification tend à sélectionner des ressources qui répondent à une requête mais sont le plus éloignées possible les unes des autres, la notion de complémentarité, quant à elle, suppose que les ressources aient une certaine proportion de recouvrement. Un travail de définition et de formalisation de la complémentarité est en cours.

[58] http://www.orthomam.univ-montp2.fr/

CHAPITRE IV



« Les images du monde extérieur se feront de plus en plus l'illustration de ma pensée. »

Salvador Dali, "La femme visible".

Visualisation et interactions exploitant un modèle sémantique

IV.1. CARTES CONCEPTUELLES ET CARTES SEMANTIQUES 92

IV.1.1. OU IL EST QUESTION DE PROJECTION MDS 93

IV.1.2. DEFINITION DE CARTE SEMANTIQUE, EXEMPLES ET UTILISATION 93

IV.1.3. COMMENT ASSISTER LA CREATION DE CARTES SEMANTIQUES 95

IV.2. ASSISTER VISUELLEMENT LA NAVIGATION DANS DE LARGES BASES DE DOCUMENTS 95

IV.2.1. GUIDES ET REPERES SEMANTIQUES POUR DECRYPTER LA CARTE 96

IV.2.1.1. Repères sémantiques sur la carte 96

IV.2.1.2. Spectres sémantiques pour représenter les ressources lors de l'indexation par propagation 98

IV.2.2. LE PRINCIPE "FOCUS AND CONTEXT" 99

IV.2.3. UTILISATION DE L'ANALYSE DE CONCEPTS FORMELS POUR LA VISUALISATION 100

IV.2.4. VISUALISATION DE PHOTOS SOCIALES SUR UN DIAGRAMME DE HASSE 102

IV.2.5. PASSAGE A L'ECHELLE : PERSPECTIVES DE RECHERCHE CONCERNANT LE NIVEAU D'ABSTRACTION 103

IV.3. EXPLICITER LES RESULTATS DE LA RECHERCHE D'INFORMATION 104

IV.3.1. PRESENTATION GLOBALE DE L'INTERFACE ET DES OUTILS DE PARAMETRISATION 105

IV.3.2. CONSTRUCTION DE LA CARTE SEMANTIQUE SYNTHETISANT LES RESULTATS 106

IV.3.3. REPRESENTATION DES ELEMENTS DU CORPUS DANS LA CARTE 107

IV.3.4. IDENTIFICATION DES PASSAGES PERTINENTS 109

IV.4. SYNTHESE ET PERSPECTIVES 109

Tandis que les SRI de première génération consistaient à automatiser une recherche à base de catalogues à l'aide de champs déterminés (auteurs, titre, etc.), ceux de deuxième génération ont mis l'accent sur les fonctionnalités de recherche à base de métadonnées, d'opérateurs de requêtage, etc., pour en arriver aujourd'hui aux SRI de 3e génération qui se focalisent sur des interfaces utilisateur améliorées graphiquement, dotées de capacité d'interaction, dans des architectures ouvertes (Baeza-Yates & Ribeiro-Neto, 2011) – chapitre 2.

"Le grand atout de la carte réside dans sa faculté à abstraire et révéler de façon immédiate une information contenue de manière non explicite dans les données, résumé ainsi par William Playfair[59] : « faire en sorte que les données parlent aux yeux »" (Villerd, 2008). Dans notre démarche visant à favoriser l'automatisation cognitive nous nous intéressons aux fonctionnalités liées à l'interaction et à la visualisation de données, qui favorisent une compréhension rapide des contenus des éléments d'un corpus et de leur adéquation avec les besoins de l'utilisateur. En effet, l'interface et l'interaction qu'elle permet jouent un rôle décisif et constituent donc des éléments centraux lors de plusieurs étapes du processus de RI : la (re)formulation de la requête, la compréhension des résultats et la sélection des plus pertinents et éventuellement la navigation au travers d'un corpus en fonction de plusieurs critères. Après avoir formulé sa requête, l'utilisateur doit comprendre les résultats qui lui sont proposés et, ayant compris les raisons de la sélection, affiner sa requête pour obtenir de meilleurs résultats.

Dans ce chapitre, nous passons en revue les différentes solutions que nous avons conçues et développées pour répondre à ces besoins. La première section présente notre perception des *cartes conceptuelles (*ou *sémantiques*) et des cartes de *connaissances*. Nous y présenterons les techniques qui peuvent être employées pour créer des cartes *qui font sens*. Plusieurs exemples sont ensuite proposés, qui mettent en exergue certaines caractéristiques des éléments des cartes de connaissance. Enfin, nous verrons comment il est possible d'automatiser certaines étapes de construction de ces cartes.

Ces travaux sur les cartes sémantiques ont été mis en application dans des contextes différents qui répondent à deux façons d'accéder à l'information pour les utilisateurs : soit un accès par "tâtonnements", en utilisant un système de navigation (éventuellement assistée) sur un corpus ; soit un accès direct à certains documents grâce à un moteur de recherche dont les résultats à une requête soumise sont présentés. Ces deux types d'utilisation répondent à des besoins spécifiques et feront l'objet respectivement des deuxièmes et troisième sections dans lesquelles nous présenterons les spécificités liées à chaque contexte applicatif. Les éléments visuels (repères sémantiques, outils de visualisation, etc.) associés à chaque contexte y seront détaillés.

IV.1. Cartes conceptuelles et cartes sémantiques

Nos capacités cognitives favorisent notre perception et notre mémorisation visuelles. Ainsi, il est plus facile d'appréhender un ensemble d'informations graphiques qu'une liste d'éléments textuels ou d'enregistrements – chapitre 14 de (Baeza-Yates & Ribeiro-Neto, 2011). Il suffit souvent de quelques secondes pour se faire une première *idée* des données présentes à l'écran et de leur position les unes par rapport aux autres. C'est ce qui nous a poussés, tout au long de nos travaux de recherche, à privilégier un rendu visuel des résultats. Pour que ceux-ci soient facilement interprétables par l'utilisateur, ils sont présentés sur des *cartes de connaissance* dans lesquelles les entités du corpus sont positionnées en fonction de la sémantique de leurs contenus. Pour cela, il est nécessaire que les distances sémantiques entre ces entités soient prises en compte dans la représentation graphique qui est fournie à l'utilisateur. C'est pourquoi nous avons utilisé des projections qui respectent au mieux ces distances et qui sont présentées dans ce qui suit.

[59] Un des premiers à synthétiser des informations statistiques par une représentation graphique (1759-1823).

IV.1.1. Où il est question de projection MDS

Comme nous l'avons détaillé dans le chapitre 2, les différentes entités que nous avons manipulées sont indexées en fonction de leur type, et du contexte d'application, en utilisant des attributs sémantiques (*descripteurs*) qui décrivent leur contenu. Cette description peut s'appuyer sur une représentation de la connaissance du domaine (e.g. ontologies) qui fournit un espace conceptuel sur lequel des mesures de distance sémantique peuvent être appliquées.

La méthode de projection MDS – *MultiDimensional Scaling* (Borg & Groenen, 2005; J. B. Kruskal, 1979; J. Kruskal, 1964; Platt, 2005), consiste à minimiser une fonction de *stress* entre *n* points (décrits au départ dans un espace à *m* dimensions), lors de la projection de ces points dans un espace à *m'* dimensions, *m'* étant plus petit que *m*. Par exemple, l'espace de départ peut être l'espace conceptuel fourni par l'ontologie de domaine ou encore un espace vectoriel (descripteurs musicaux dans MBox-*composer*). Une matrice de distances sémantiques peut être calculée entre toutes les entités indexées avec les concepts de cette ontologie. Leur positionnement à l'écran respectera au mieux ces distances et donc une proximité physique à l'écran, sur la carte, correspond à une proximité sémantique dans l'espace de départ. Il existe plusieurs approches pour calculer une projection MDS, dont une consiste à calculer la minimisation de la fonction de stress en étendant ou rapprochant les distances euclidiennes entre les points qui doivent être représentés (Buja & Swayne, 2002). Cette approche est connue sous le terme de *Force Directed Placement* (FDP) ou encore algorithme des ressorts – *Spring Embedding Algorithm* (Eades, 1984; Morrison, Ross, & Chalmers, 2003). En effet, lors du placement des éléments, plusieurs forces d'attraction et de répulsion peuvent intervenir : e.g. pour une entité, la proximité sémantique avec d'autres ayant tendance à les attirer, tandis qu'une *limite* a tendance à les repousser pour éviter les chevauchements.

Nous avons testé et validé nos solutions dans l'environnement *MolAge* (pour *Molecular Agents*) qui est une plateforme multi-agents, développée dans notre équipe à l'initiative de Michel Crampes. Dans cet environnement, chaque élément visuel (représentant un document, une personne, un titre de musique,…) constitue un agent qui a certaines propriétés en fonction desquelles il calcule sa position par rapport aux autres. Ce calcul utilise l'algorithme des ressorts, avec comme *force* entre les éléments, les distances sémantiques qui les séparent. Ainsi, MolAge permet d'afficher des cartes sémantiques avec lesquelles il est possible d'interagir en manipulant visuellement des entités.

IV.1.2. Définition de carte sémantique, exemples et utilisation

Cartes de connaissances, *cartes sémantiques*, *cartes conceptuelles* : il est difficile de se faire une idée précise de ce que traduisent ces différentes appellations. L'engouement qu'elles suscitent dans différents domaines (psychologie, sciences sociales, ingénierie de la connaissance, informatique, analyse lexicale…) conduit souvent à les confondre. Il nous semble donc important de préciser la définition que nous leur associons dans ce mémoire.

Nous appelons carte *conceptuelle*, la représentation (le plus souvent sur un écran) de la connaissance d'un domaine, faisant apparaître les différents concepts de ce domaine et les relations entre ces concepts (e.g. *mind map*, *réseau sémantique, ontologie…*).

La carte *sémantique* ou carte de *connaissances*, représente, quant à elle, un ensemble de données (ressources, informations) organisées en fonction de différents critères sémantiques définis par rapport à la modélisation de la connaissance du domaine. Ces cartes sémantiques, en présentant les informations à l'utilisateur en regard de cette connaissance, favorisent leur structuration, leur *assimilation* et donc leur *mémorisation*. Elles peuvent être assorties de techniques d'interaction afin de permettre à l'utilisateur de sélectionner certains éléments, de modifier certains paramètres et d'en observer dynamiquement les effets sur la carte ou encore de changer de vue sur les éléments du corpus qui lui sont présentés.

Si la visualisation de données varie fortement en fonction du contexte applicatif, la carte est généralement un outil qui se partage au sein d'une communauté, que ce soit une communauté

d'usage (gestionnaires, chercheurs, étudiants) ou une communauté liée à la thématique, au domaine (biologie, sport, gastronomie,…). Si la taille des données est trop grande, des outils de filtrage sont nécessaires pour ne pas noyer l'utilisateur. L'utilisation d'un même type de représentation au sein d'une communauté favorise les habitudes, les usages. Pour toutes ces raisons, la carte sémantique constitue un facteur majeur de l'automatisation cognitive. Elle permet, en effet, de réduire les temps d'apprentissage pour un opérateur humain. En favorisant une perception rapide et organisée des informations, l'utilisateur gagne en réactivité et acquiert certains *réflexes* pour décider rapidement des actions à mener, par exemple en situation de crise. Leur évaluation ne peut se faire que par l'utilisation, sur des critères tels que : rapidité, facilité d'utilisation, esthétique, compréhension…

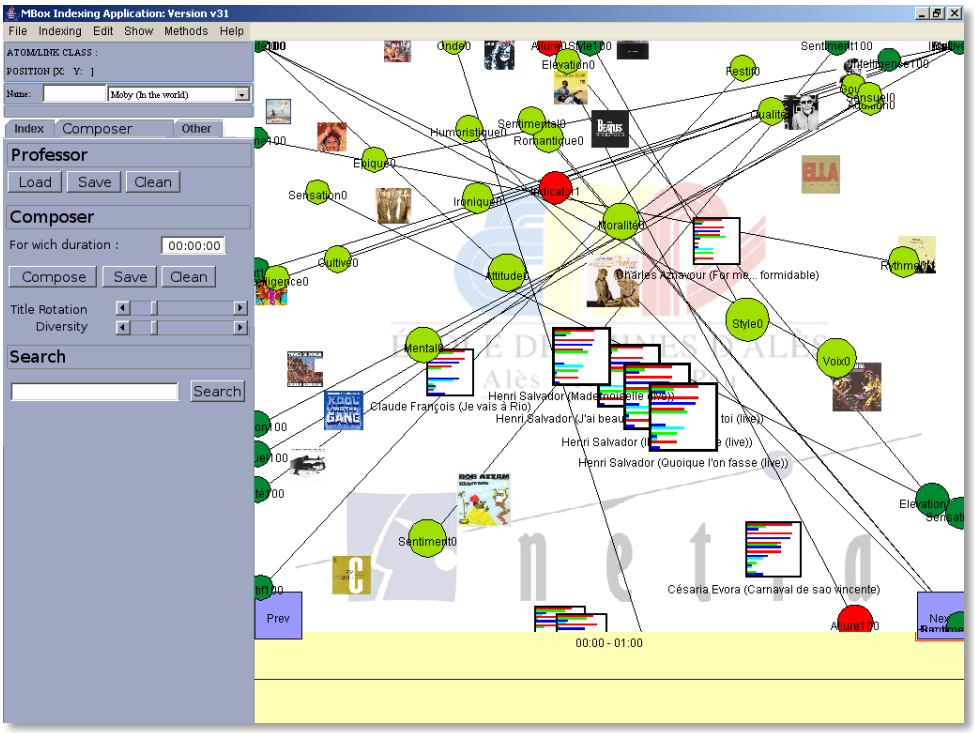


Figure 12 : Carte sémantique représentant les titres musicaux avec des repères sémantiques

La notion de carte et ses différents usages sont présentés dans le chapitre 3 de (Jalabert, 2007). A l'image de la carte géographique, la carte de *connaissance* est spécifique à un besoin et à un domaine. Elle se veut intuitive, utile, partageable, permettant de se *repérer* dans un espace complexe que l'on souhaite percevoir sous un certain angle, à différentes échelles. Elle doit permettre de se déplacer au travers d'un *paysage* afin d'explorer un *espace* pour justifier, mémoriser, partager et capitaliser des connaissances. Pour cela, des repères peuvent être proposés. Le fait qu'elle soit qualifiée de sémantique impose que ces *repères* dépendent de la sémantique contenue dans les entités représentées. Au final, la carte doit proposer à l'utilisateur une vue sur un ensemble de données, de façon à ce qu'il s'en fasse une représentation mentale, appelée *carte mentale*, où il met en relation les données avec leur contexte et la connaissance du domaine. S'il souhaite changer de point de vue sur les données, il est important de respecter cette carte mentale pour laquelle les repères visuels serviront de cadre de référence. Ces repères indiquent, par exemple, la forte représentativité d'un concept à un endroit de la carte, un chemin au travers des données représentées ou encore des zones de regroupement dans cette carte. La Figure 12 montre un exemple de carte sémantique qui a été créée dans le cadre du projet MBox-*composer*. Dans cette carte, les titres musicaux ont été indexés en fonction des *humeurs* (*moods* en anglais c.f. chapitre II, section II.2.2). On obtient ainsi pour chacun d'eux, un vecteur de 23 valeurs, les dimensions de ce vecteur

correspondant aux humeurs définies. Un calcul de distance vectoriel peut donc être appliqué entre les différentes entités, dans cet espace vectoriel. La projection MDS permet de visualiser la proximité des titres.

En résumé, les cartes sémantiques, comme les cartes géographiques, permettent de localiser des données (informations) les unes par rapport aux autres, en fonction de certaines caractéristiques sémantiques. Même si elles peuvent représenter des données dans un espace multidimensionnel, purement abstrait, on y retrouve, par analogie, les notions de distance, de chemin, de relief ou encore de frontières, qui sont connues de tous dans le domaine des cartes géographiques. Leur conception, très dépendante d'un contexte applicatif, peut être longue et fastidieuse. C'est pourquoi, avec Michel Crampes, nous avons imaginé, toujours en utilisant l'environnement MolAge, l'automatisation de certaines phases de cette conception.

IV.1.3. Comment assister la création de cartes sémantiques

Si les cartes sémantiques sont un moyen efficace de représenter visuellement des données et d'interagir avec la connaissance d'un domaine, leur conception n'est pas immédiate. Cela est en partie dû au fait de la complexité des données à représenter. Inspirés par le modèle MVC (*model – view – control*), nous avons proposé la méthode DVC pour *domain – view – controller* (Crampes, Ranwez, Villerd, et al., 2006). Cette méthode a été intégrée à l'environnement MolAge afin de proposer aux concepteurs de cartes sémantiques, un moyen de spécifier la carte qu'ils veulent créer et d'organiser en conséquence les informations à représenter. Dans l'application choisie, toujours en lien avec le projet MBox-Composer, le *domaine* est représenté par une carte conceptuelle de la musique. Les concepts du domaine et leurs relations y sont représentés. Toute l'ontologie ne peut cependant pas être représentée et seuls les concepts d'une *sous-ontologie* utilisée dans un contexte donné (par exemple pour l'indexation) sont représentés. La *vue*, quant à elle, fournit les informations relatives à l'affichage souhaité. Elle est composée de *caractéristiques*, de *propriétés*, qui peuvent être appliquées à des entités. On y trouve, par exemple, la notion de *temps*, l'*ordre alphabétique* ou des contraintes *spatiales*. En mettant en relation la *vue* et le *domaine*, il est possible de spécifier que les *interprètes* apparaissent en haut de la carte et suivant l'ordre alphabétique, que les *compositeurs* soient affichés en bas de l'écran et suivant l'ordre chronologique de leur naissance, etc. Le module de *contrôle*, gère toutes les interactions avec la carte. Cette approche permet une grande flexibilité au niveau de la conception des cartes et permet de les personnaliser en fonction de différents contextes. Il est possible de créer, pour cela, des scénarios prédéfinis.

Les cartes ainsi créées peuvent être utilisées dans différents contextes. La section suivante présente leur utilisation pour naviguer dans de larges corpus.

IV.2. Assister visuellement la navigation dans de larges bases de documents

En fonction du contexte de l'utilisateur, la manière d'accéder à l'information va varier. Dans certaines situations, il préfèrera parcourir l'ensemble des entités du corpus, en naviguant de proche en proche entre ces entités. C'est le cas lors de la phase d'indexation, par exemple, ou, comme nous le verrons, pour la composition de *playlists*. Cependant, la taille du corpus est parfois tellement grande qu'il est nécessaire d'utiliser des techniques particulières (filtrage, guides, etc.) pour assister cette navigation. Au cours de nos travaux, nous avons utilisé différentes fonctionnalités permettant de mettre en évidence certains éléments : *zoom*, *panning zoom*[60], ou de modifier l'apparence de certains éléments lorsque l'on approche le curseur de la souris : *fisheye*, *lentilles sémantiques*,

[60] Lorsque certaines entités se chevauchent, il est possible de les faire apparaître momentanément sans chevauchement, dans une zone autour du curseur de la souris.

filtrage. Ces techniques ne sont pas toutes détaillées ici même si elles peuvent être utilisées dans les approches de navigation que nous proposons.

IV.2.1. Guides et repères sémantiques pour décrypter la carte

Assister l'utilisateur lors de sa navigation peut être fait soit en positionnant des repères sémantiques sur la carte qui mettent en exergue certaines caractéristiques des ressources représentées, soit en modifiant l'affichage de certaines ressources en fonction d'un contexte d'utilisation.

IV.2.1.1. Repères sémantiques sur la carte

Guider lors de la phase d'indexation. Dans le projet MBox-composer, la seule visualisation des titres musicaux, dispersés sur l'écran comme des pochettes de disques pourraient l'être sur la table de mixage, n'est pas suffisamment explicite (c.f. Figure 7 dans la section II.4.2). Certains DJ nous ont demandé d'éclairer la carte en explicitant mieux le positionnement des éléments en fonction de leur description (vecteur de caractérisation). Ce point est crucial lors de l'indexation par propagation car pour positionner l'entité à indexer les experts désirent parfois suivre certaines *directions* sémantiques avant de la déposer sur la carte. C'est pourquoi nous avons proposé d'indiquer certains champs sémantiques (qu'on peut assimiler à des champs d'attraction) en rajoutant des *sondes* (cercles verts sur la Figure 12). Ces sondes indiquent les descripteurs qui ont le plus d'influence dans les différentes zones de la carte. Elles représentent des entités virtuelles qui ne sont indexées que par un seul descripteur ; celui-ci a pour valeur soit 0 (cercles vert clair, indiquant le minimum pour ce descripteur), soit 1 (cercles vert foncé, indiquant le maximum pour ce descripteur). Leur positionnement dans la carte est calculé en utilisant l'algorithme FDP, et donc en tenant compte de tous les autres éléments pour cet unique descripteur. L'indicateur du minimum (respectivement du maximum) va donc se placer dans un secteur de la carte où les valeurs sont faibles (respectivement hautes) pour ce descripteur pour les autres entités. La ligne qui relie le maximum et le minimum indique un axe fictif (le gradient entre le minimum et le maximum n'est pas linéaire, mais en pratique cette représentation est souvent suffisante).

Assister la navigation. Il y a différents critères à partir desquels l'utilisateur peut vouloir entrer dans la base de titres musicaux : les Compositeurs, les interprètes… On peut alors imaginer des guides comme montré dans la Figure 13 : l'utilisateur peut sélectionner un interprète (ils sont classés par ordre alphabétique) et voir l'ensemble des titres qui lui est associé, ou bien choisir un compositeur (par ordre chronologique de leurs dates de naissance) et voir l'ensemble de son œuvre dans la base. On conserve également le même type de navigation que dans la Figure 12, puisque dans le paysage central, les titres sont représentés en fonction de leur indexation sémantique.

Parcourir le paysage musical pendant la phase de composition de *playlists*. Une fois la carte créée et explicitée pour l'utilisateur, différents *chemins* peuvent être exploités afin de parcourir l'espace des données. Ces chemins peuvent être *balisés*, comme nous l'avons dit plus haut, par des gradients sémantiques (utilisés lors de l'indexation), ou par des guides sémantiques pour assister l'utilisateur dans sa navigation (ce point sera développé dans la section IV.2.3). Dans le cadre du projet MBox-*composer*, nous avons exploité cette notion de chemin pour l'étape de composition. L'objectif du projet était la composition automatique de *playlists* musicales pour des galeries commerciales, des concessions automobiles ou des radios. Chacun de nous a pu faire l'expérience, dans de tel lieux (ou parfois il faut attendre longtemps, revenir,…) d'entendre les mêmes fonds musicaux en boucle. Or ces enregistrements ne sont pas le fruit du hasard, ils ont été composés par des DJ, en fonction de critères particuliers (type de clientèle, environnement social, produits en vente,…) pour favoriser le bien-être du client et donc ses achats. Certaines (mauvaises !) radios diffusent, elles-aussi des enchaînements prédéfinis de titres musicaux. La société Nétia voulait fournir à ses clients, une *Box* permettant de composer automatiquement des *playlists*, et éviter ainsi de lasser les auditeurs en répétant toujours la même chose. L'idée était d'apprendre, à partir de

playlists composées par les experts, un certain cheminement dans la carte sémantique représentant la base de données musicale. Cet apprentissage était ensuite mis à profit en utilisant les N-*Grammes* pour construire un cheminement proche de celui de départ, mais impliquant des titres différents. Cette technique consiste à partir d'une succession d'éléments, à prédire (en fonction de calculs de probabilité) le prochain élément de la liste.

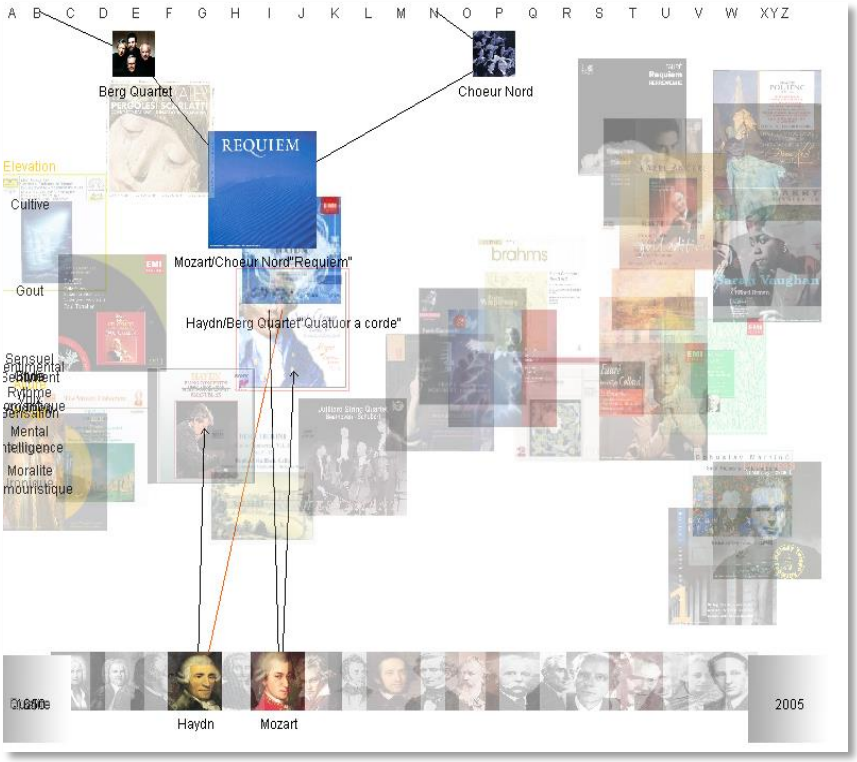


Figure 13 : Interface de navigation dans une collection de titres musicaux.

Les interprètes sont présentés par ordre alphabétique en haut de l'écran ; les compositeurs sont organisés en fonction de l'ordre chronologique de leur naissance. Des lentilles sémantiques permettent de n'afficher que les entités sélectionnées, les autres étant présentées en transparence.

Les chemins appris l'étaient en fonction du vecteur de descripteurs de chaque titre musical. Un chemin correspond à une évolution voulue par le DJ pour susciter certaines émotions chez l'auditeur ; c'est un style de composition. On peut, par exemple, commencer par un titre relativement *calme* mais *lumineux* pour aller vers des titres plus *mélancoliques* et *cultivés* en passant par des titres plus *rythmés* et *rebelles*. En résultat le système MBox-*composer* génère des *playlists* qui empruntent un chemin voisin de ceux appris. Quand plusieurs chemins existent, l'utilisateur peut les combiner de façons différentes en étant plus ou moins permissif, i.e. autoriser plus ou moins de latitude par rapport aux *playlists* d'apprentissage. L'illustration présentée dans la Figure 14 montre ces chemins. Les repères sémantiques sont représentés ici par des bulles bleues correspondant à l'endroit de la carte où elles sont le plus représentées. Le contenu de la *playlist* est détaillé dans le *bandeau* du bas.

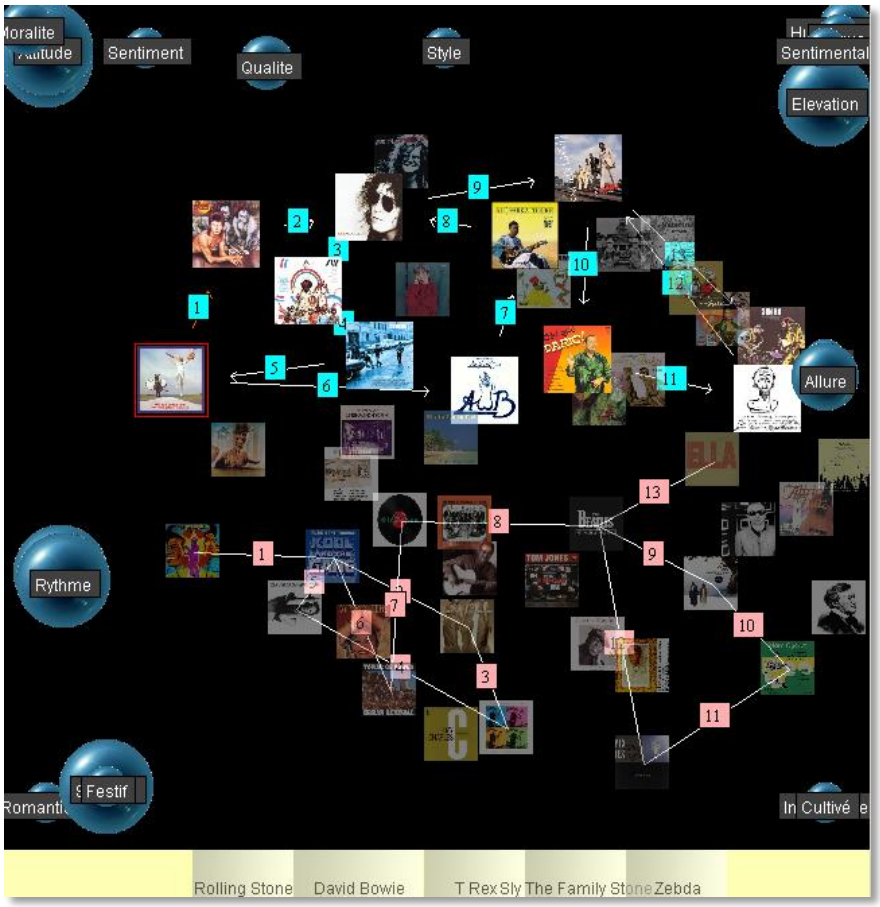


Figure 14 : Deux chemins représentant des *playlists* dans MBox-*Composer*.

IV.2.1.2. Spectres sémantiques pour représenter les ressources lors de l'indexation par propagation

Le principe de l'indexation par propagation, a pour objectif d'indexer rapidement un grand nombre de ressources. Il est donc primordial de fournir à l'opérateur humain toutes les informations nécessaires pour cette indexation. Dans le paysage musical présenté dans la Figure 6 du chapitre II (section 35II.4.2), les titres musicaux sont affichés sous la forme d'une image représentant l'album dont ils sont issus, ou la pochette de leur *single* si elle est disponible. Cette visualisation permet rapidement de savoir de quel titre il s'agit, mais ne donne pas d'indication sur les valeurs d'indexation. Lors de l'utilisation de la carte pour l'indexation par propagation, les titres à proximité du curseur de la souris ne s'affichent plus comme cité précédemment, mais sous la forme d'un pictogramme représentant les valeurs numériques (poids) associées aux 23 descripteurs qui les caractérisent (c.f. Figure 15). L'indexeur peut rapidement voir les valeurs associées au titre qu'il est en train d'indexer, et vérifier que ces valeurs sont bien en accord avec les valeurs de ses K-plus proches voisins. Il peut ainsi suivre les variations du vecteur d'indexation et "déposer" le titre à l'endroit de la carte qui lui semble le plus pertinent. Une fois l'élément déposé, les 23 valeurs calculées sont associées aux descripteurs. L'opérateur humain peut les retoucher pour qu'elles répondent mieux à sa perception du titre : e.g. renforcer le poids d'un descripteur qui lui semble particulièrement important pour un titre mais qui aurait obtenu une faible valeur lors de la propagation.

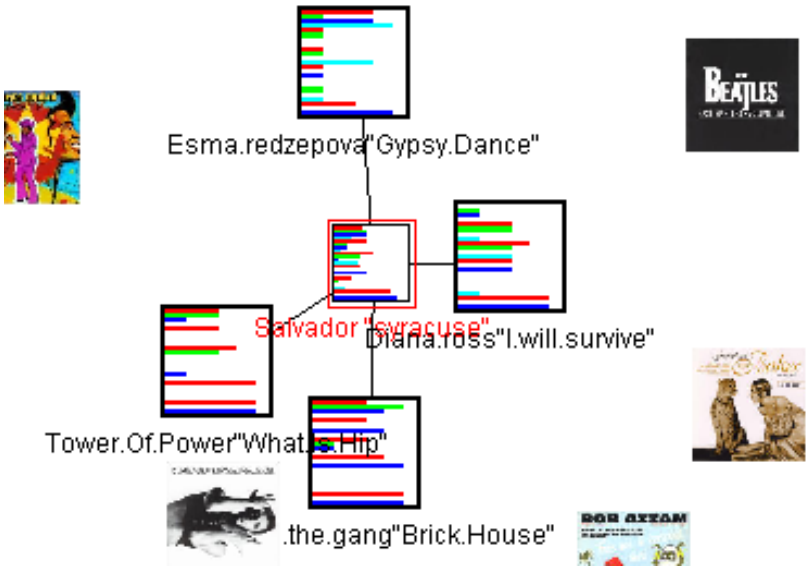


Figure 15 : Visualisation de la propagation des poids des descripteurs grâce à un filtre sémantique. *Ici, un zoom sur le paysage de référence est présenté et le titre en cours d'indexation est Syracuse d'Henry Salvador.*

IV.2.2. Le principe "Focus and context"

A la suite des travaux de (Shneiderman, 1996), l'approche "*focus + context*" a suscité de nombreux travaux dans la communauté InfoVis (visualisation d'information). Cette approche recommande de proposer d'abord une vue globale des données, puis de permettre à l'utilisateur d'identifier une zone d'intérêt (*focus*) et enfin d'afficher le contexte détaillé de cette zone (*context*). La vue globale permet à l'utilisateur de conserver une même *carte mentale* des données (i.e. la représentation qu'il s'en fait), tout au long de sa navigation. Nous avons utilisé une telle approche dans la thèse de Fabien Jalabert (Jalabert, 2007). L'objectif de cette thèse, rappelons-le, était de créer une interface de médiation entre différentes sources de données utilisées par les biologistes, e.g. des bases de gènes, leurs annotations avec des concepts de la Gene Ontology ou des corpus de publications scientifiques. L'environnement développé s'appelait I^{2}DEE pour I*nteractive and Integrated Data Exploration Environment*. L'intégration des données dans un modèle unifié, était réalisée grâce à UMLS. Concernant la visualisation, les différentes ressources qui devaient coexister sur la carte devaient être différenciées. Par ailleurs, leur nombre très élevé imposait de mettre en application les recommandations de Shneiderman. Une vue du prototype développé par Fabien est présentée dans la Figure 16. Ces travaux étaient très intéressants et restent d'actualité si l'on en croit les biologistes qui expriment souvent leur besoin de croiser ce genre de données.

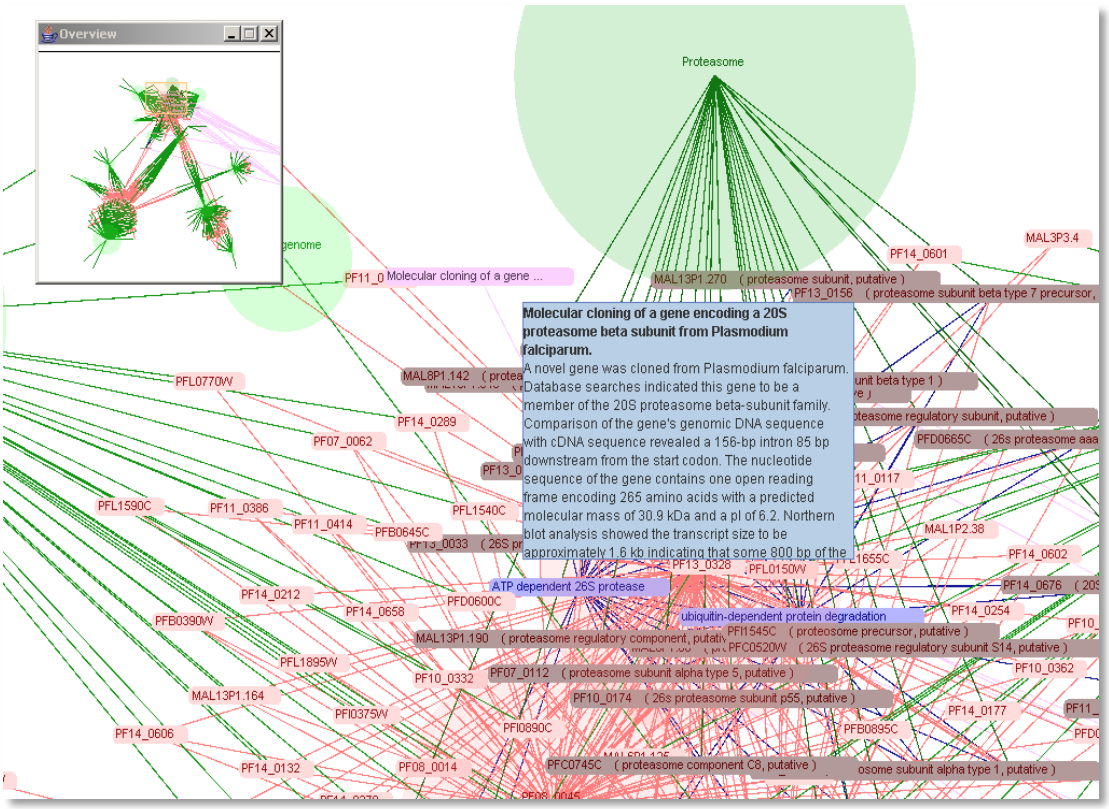


Figure 16 : Interface de l'outil I²DEE.
En haut à gauche, on a un aperçu global des données (context) et une vue rapprochée d'une partie de la carte est visible dans la partie droite (focus). Ici les entités représentées correspondent à des regroupements des gènes (cercles verts), des gènes (en saumon, ou en plus foncé pour ceux qui appartiennent au groupement sélectionné, ici Proteasome), leur annotation avec des concepts de la Gene Ontology (en violet) et des publications (au passage de la souris, le résumé est proposé à l'utilisateur, ici en bleu).

Cependant la solution visuelle présentait plusieurs limites. En premier lieu, l'encombrement de la carte. Malgré la vue globale, il est difficile de se faire une idée précise des données représentées et de maîtriser le contenu de la carte. Il serait préférable dans la vue générale (*context*) de présenter des éléments englobant un ensemble de ressources (*clustering*), que l'utilisateur pourrait découvrir à sa demande. D'autre part, la sémantique de la carte n'est pas explicitée. La notion de distance sémantique n'était pas exploitée lors de la conception de la carte. Les éléments étaient présentés en utilisant l'algorithme des ressorts, mais sans contrainte autre que celle d'éviter les chevauchements. Si bien qu'on peut avoir des éléments qui sont représentés loin les uns des autres, alors qu'ils ont une forte proximité sémantique.

Toutes ces raisons nous ont amenés à proposer de nouvelles solutions pour la navigation.

IV.2.3. Utilisation de l'Analyse de Concepts Formels pour la visualisation

Dans le cadre d'une collaboration avec la société INova[61] il nous était demandé de réfléchir à la navigation dans une base de données de brevets. Nous avons choisi de coupler une analyse de ces

[61] Devenue depuis Inova-software : http://www.inova-software.com/
Cette société développe des logiciels de partage d'information au sein de grand groupes industriels, afin de favoriser

brevets qui utilise la théorie des FCA (*Formal Concept Analysis*) avec des techniques de visualisation, pour suggérer des chemins et aider l'utilisateur à se focaliser progressivement sur certains brevets. En effet, la FCA permet de décrire la structure conceptuelle des données dans un treillis de concepts (appelé treillis de Galois), et de visualiser, au travers de ce treillis, certaines propriétés de ces données. Ainsi, à partir d'une vue globale des données, l'utilisateur peut, en sélectionnant certaines de ces propriétés, filtrer certains documents pour ne conserver que les plus pertinents.

Les fondements de la théorie des FCA sont présentés dans (Ganter & Wille, 1999). Dans nos travaux, le treillis de concepts correspondait à la matrice document/termes d'indexation (mots-clés). Ce treillis servait de support à la navigation et de vue globale qui soulignait la structure conceptuelle de la base. Notre contribution concernait principalement les techniques de visualisation associées à ce modèle (Villerd, Ranwez, & Crampes, 2008; Villerd, Ranwez, Crampes, & Carteret, 2009). En effet, plutôt que de représenter le treillis sous forme de diagramme de Hasse, comme c'est d'usage dans la communauté FCA, les nœuds du treillis sont disposés en utilisant l'heuristique des *ressorts* (*Force Directed Placement*), les forces appliquées à chaque ressort provenant d'un calcul de distance sémantique (S. Ranwez et al., 2006), c.f. Figure 13. Certains regroupements peuvent être identifiés, en fonction de la proximité sémantique des ressources. Dans la partie droite de la figure, on peut voir l'effet de l'application d'une lentille sémantique particulière : un concept sélectionné et les concepts qui ont un lien direct avec lui sont mis en évidence et leur nom apparaît tandis que les autres apparaissent en transparence. Pour éviter les encombrements, toutes les arêtes ne sont pas représentées, seules celles correspondant au concept sélectionné apparaissent. Les nombres qui sont affichés dans chaque concept correspondent au nombre d'instances (de documents) qui leur sont associées. Cette visualisation permet de naviguer sur les nœuds du treillis, tout en estimant l'étendue de la distance sémantique à chaque pas.

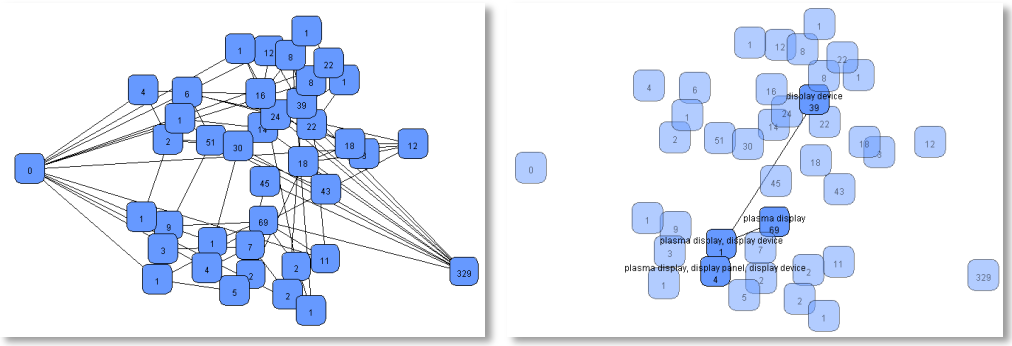


Figure 17 : Représentation d'un treillis de concepts à l'aide d'une projection MDS.
La vue globale est une projection MDS d'un treillis de concepts en utilisant une distance sémantique. Toutes les arêtes sont montrées sur la vue de gauche. Un concept particulier {plasma display, display device} et ses voisins sont mis en évidence dans la vue de droite

Cette visualisation des treillis constitue la vue globale de la base de données des brevets. Une vue locale est associée, pour représenter les données *brutes*, i.e. les éléments du corpus. Là encore, la distance sémantique peut être utilisée, pour observer des regroupements locaux. Cette vue locale (partie gauche de la Figure 18) est constituée des éléments du corpus qui sont en relation avec les concepts sélectionnés dans la vue globale (partie droite).

l'émergence d'innovation technologique et était partenaire des deux propositions de projet soumises à l'appel de l'ANR-ContInt.

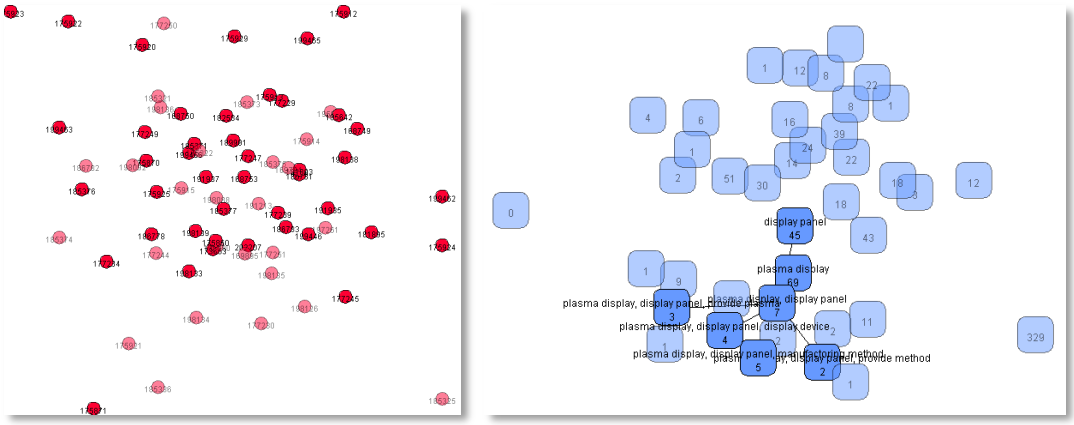


Figure 18 : Interaction entre une vue locale et une vue globale d'une base de brevets.
La sélection d'un concept dans la vue globale (à droite) affiche les brevets contenus dans l'extension de ce concept dans la vue de gauche et suggère des chemins de navigation en mettant certains concepts en évidence dans la vue générale.

L'interaction avec la carte est assurée dans les deux vues. Dans la carte de droite, l'utilisateur peut sélectionner certains concepts et de proche en proche se déplacer dans le graphe de concepts pour cibler ses recherches. Dans la vue locale (partie gauche), les ressources correspondant au concept sélectionné ainsi qu'à ses plus proches voisins apparaissent. Il est possible de sélectionner une ressource pour avoir plus de détail (vue synthétique du document, son indexation, etc.). Le cas d'utilisation typique consiste à partir du nœud le plus élevé dans le treillis pour afficher l'ensemble de la collection. Ensuite, deux possibilités se présentent : l'utilisateur continue sa navigation sur le treillis en sélectionnant un des nœuds adjacents au nœud de départ (mis en évidence). Quand l'utilisateur bouge la souris sur un nœud particulier, les documents correspondants sont mis en évidence pour que l'utilisateur ait une pré-visualisation de la vue qui va suivre et pour observer la répartition de ces documents sur la vue actuelle. Quand l'utilisateur sélectionne un nouveau nœud, la vue locale est mise à jour pour enlever les entités qui n'appartiennent pas à l'extension du concept choisi. Dans ce cas, la distance n'est pas recalculée et la carte ne bouge pas (respect de la *carte mentale* de l'utilisateur) ; seuls les niveaux de visibilité ont changé.

Le positionnement et l'organisation de ces ressources à l'écran sont fonction de la distance sémantique qui les sépare. La distance utilisée peut être indépendante des attributs contenus dans le treillis de Galois. En effet, il peut être intéressant de comparer deux niveaux d'indexation. Par exemple dans le cas de documents multimédia, des descripteurs concernant les *'moods'* peuvent être utilisés pour structurer les données et construire la base conceptuelle (le treillis de concepts) tandis que des descripteurs physiques, obtenus par des techniques d'analyse du signal peuvent être utilisés pour calculer la distance sémantique entre les entités, en supposant que l'utilisateur utilise un niveau d'abstraction très haut pour réaliser le focus de Schneiderman et un niveau d'abstraction plus bas, plus proche des données brutes, pour observer l'organisation locale des documents. Utiliser la même distance pour représenter toutes les vues locales préserve la représentation mentale de l'utilisateur.

IV.2.4. Visualisation de photos sociales sur un diagramme de Hasse

Nous appelons photos sociales, les clichés pris lors de rencontres telles que les réunions de familles ou des évènements sociaux (mariage, anniversaires, remise de diplôme, etc.) et qui représentent généralement des individus ou des groupes d'individus. Ce genre de manifestation est souvent propice à un grand nombre de photos qu'il est fastidieux de classer, d'organiser. Pourtant, une indexation est nécessaire, si l'on veut réaliser, par exemple des albums-souvenirs. La première composante de l'indexation est facile à réaliser : la date, l'heure, le lieu et l'évènement correspondant.

La deuxième composante est plus délicate à automatiser : les personnes présentes. C'est pourquoi nous avons imaginé une interface pour représenter ces photos (Crampes, De Oliveira-Kumar, Ranwez, & Villerd, 2009). Toujours en suivant l'idée de l'utilisation de l'analyse de concepts formels pour organiser la visualisation, nous avons proposé de classer ces photos en fonction des personnes présentes. Les personnes constituent donc les "attributs" de chaque photo et sont organisées dans un treillis.

Figure 19 : Interface d'indexation de photos sociales.
Sur la vue de gauche, on voit les photos organisées suivant un treillis en fonction des personnes présentes sur chaque photo. Sur la vue de droite, une photo est en cours d'indexation.

La Figure 19 présente deux vues de l'environnement développé dans ce contexte. La structure de treillis est représentée dans la partie gauche. Les personnes seules sont situées dans le niveau le plus haut du treillis (le plus à gauche sur la vue) puis les groupes de deux sont formés en lien avec ces photos individuelles, puis des groupes de trois, etc. La vue de droite présente l'environnement d'indexation. Les personnes sont identifiées dans les cercles à gauche. Elles sont classées par ordre alphabétique. Au centre, une carte présente le treillis (les liens ne sont pas visibles pour ne pas surcharger la vue). Les photos à indexer sont à droite. Une d'entre elles est sélectionnée et doit être indexée. L'utilisateur peut la positionner dans le treillis et le système l'indexe automatiquement en fonction de ses plus proches voisins (principe de l'indexation par propagation). Si une personne non encore mentionnée dans le système est présente sur la photo à indexer, il est possible de la rajouter. Ce treillis, représenté comme un diagramme de Hasse, est utilisé à la fois pour naviguer dans le paysage, mais aussi comme guide lors de l'indexation. Une fois l'indexation réalisée, la recherche de photos et leur partage est grandement facilité, nous l'avons testé pour la création d'albums, par exemple. Cet environnement a été soumis à plusieurs étudiants auxquels il était demandé d'indexer un ensemble de leurs photos personnelles. Il en est ressorti une grande facilité d'utilisation et un côté ludique qui plaisait à tous.

IV.2.5. Passage à l'échelle : perspectives de recherche concernant le niveau d'abstraction

Le problème du passage à l'échelle et de la représentation d'un trop grand nombre de données a souvent constitué un verrou dans nos travaux. Nous y avons répondu la plupart du temps par des procédés visuels (lentilles sémantiques), ou par des techniques de filtrage. Parmi ces techniques de filtrage, certaines sont éprouvées, d'autres font partie de nos perspectives à court terme.

Lorsque nous avons imaginé l'extraction de sous-ontologie – *ontoFocus* (V. Ranwez et al., 2009; V. Ranwez, Ranwez, et al., 2012), une des finalités était de pouvoir représenter des vues différentes en fonction du degré d'abstraction souhaité. Par exemple, concernant l'exploitation de la

Gene Ontology et des annotations de gènes utilisant ses concepts, il peut être souhaitable de représenter les sous-ontologies rattachées aux annotations d'un gène ou d'une famille de gènes. Les biologistes utilisent souvent ce genre de vue réductrice de la GO, après les avoir créées manuellement. En utilisant *ontoFocus*, il est possible de construire de telles vues en temps réel. C'est ce qui est fait par exemple dans le site d'interrogation de la base de données OrthoMam[62] et dans le site de valorisation des résultats de recherche du projet ANR Phyl-Ariane[63], qui utilisent *ontoFocus* en ce sens. Cependant dans les deux cas, les vues sont figées. Nous envisageons d'intégrer ces vues dans un environnement interactif afin de permettre à tout moment de revenir dans un contexte plus large (incluant plus de concepts de la GO), en utilisant un procédé qui s'apparenterait à un zoom sémantique. Cette fonctionnalité serait du plus grand intérêt lors de la phase d'annotation de gènes, par exemple, ou pour comparer différentes annotations.

La visualisation de sous-ontologie serait également particulièrement utile dans les environnements de conception d'ontologies, où elle permettrait de ne visualiser que des parties restreintes de l'ontologie (celle correspondant, par exemple, au domaine d'expertise d'un opérateur humain) mais en gardant l'intégralité de l'ontologie comme référence, et faciliter ainsi la conception ou la mise à jour de l'ontologie.

Nous n'avons pas encore testé ce mode d'affichage mais il est envisagé de le développer dans le cadre de la mission confiée à Nicolas Clairon (début de contrat en mars 2013) car elle doit être intégrée à un module d'indexation par propagation de publications scientifiques. Ce module de représentation de sous-ontologies sera également utile pour le rendu visuel de la recherche d'informations par complémentarité (thèse de Nicolas Fiorini), par exemple pour visualiser les compétences de chercheurs et être à même d'évaluer la couverture du domaine assurée par un collectif de chercheurs (dans le cas d'une réponse à un appel à projet, par exemple). On peut même imaginer qu'un chercheur, en voyant la couverture sémantique calculée par le système pour le caractériser (après une analyse de ses publications scientifiques), modifie son 'profil' en ajoutant ou supprimant certains concepts. Cette action plus ludique que la lecture d'une liste de concepts, aurait plus de chances de susciter l'attention de l'utilisateur et donc favoriserait la saisie d'information.

Un autre mode de différentiation des niveaux d'abstraction consisterait à représenter les clusters d'entités comme des ressources à part entière, en développant éventuellement à la demande certains clusters. C'est une notion qui était déjà présente dans les travaux de thèse de Fabien Jalabert (c.f. les regroupements sur *l'overview* de la Figure 16), mais que nous n'avons pas approfondie. Il serait ainsi possible de visualiser progressivement les données en fonction de regroupements sémantiques entre elles. Ce point fait également partie de nos perspectives à court terme car dans la restitution des résultats d'un moteur de recherche d'information, ce *clustering* allègerait l'affichage.

IV.3. Expliciter les résultats de la recherche d'information

Au lieu de représenter l'intégralité des ressources disponibles en laissant à l'utilisateur le soin de naviguer entre ces ressources, la carte sémantique peut être utilisée pour représenter les résultats d'une recherche d'information. Or un inconvénient de beaucoup de systèmes d'IR est le manque d'expressivité de leurs résultats. Dans la plupart des cas, on propose simplement une liste de ressources sans explications concernant le processus d'appariement entre la requête de l'utilisateur et les ressources sélectionnées question : n'avons-nous pas tous été quelque peu déroutés par la liste de résultats proposée par Google ? Se contenter de surligner les mots de la requête qui apparaissent dans une ressource n'est pas suffisant ; il faut organiser les résultats pour qu'on comprenne leur distance à la requête. Mais même dans le cas où un classement est effectué (à partir du calcul du RSV de chaque ressource), celui-ci est rarement explicite. En l'absence de justification concernant

[62] http://www.orthomam.univ-montp2.fr/orthomam/html/

[63] http://www.lirmm.fr/phylariane/

les résultats de systèmes, les utilisateurs sont démunis et peuvent être embarrassés pour modifier leur requête de manière satisfaisante dans un processus de recherche itératif.

Dans le développement d'OBIRS, nous nous sommes attachés à représenter le plus synthétiquement et le plus fidèlement possible les informations utiles à l'utilisateur. L'environnement se veut le plus interactif possible durant tout le processus : de la formulation de la requête à l'exploitation des résultats.

IV.3.1. Présentation globale de l'interface et des outils de paramétrisation

L'interface d'accueil d'OBIRS est présentée dans la Figure 20. Ici le mode de recherche avancé (*advanced search*) est activé. Nous rappelons que cette application est dédiée à la recherche de gènes indexés par des concepts de la GO. Il est possible de sélectionner une espèce parmi les 6 proposées (*Homo sapiens*, *Mus musculus*, *Plasmodium falciparum*, *Danio rerio*, *Oryza sativa*, *Arabidopsis thaliana*), puis de saisir une requête composée de noms de concepts. Pour faciliter cette saisie, une complétion automatique est activée. Il est également possible de naviguer dans l'ontologie (représentée sous une forme hiérarchique, comme dans l'explorateur de fichiers de Windows) et de sélectionner dans la requête des concepts plus spécifiques dans l'arborescence. Un espace de paramétrisation est également accessible, dans lequel il est possible de sélectionner le type de calcul de distance sémantique utilisé et préciser si l'on veut un opérateur d'agrégation plutôt conjonctif (proche du ET logique) ou disjonctif (proche du OU) (correspond au paramètre q dans l'équation du calcul du RSV de la section III.2.1 du chapitre III).

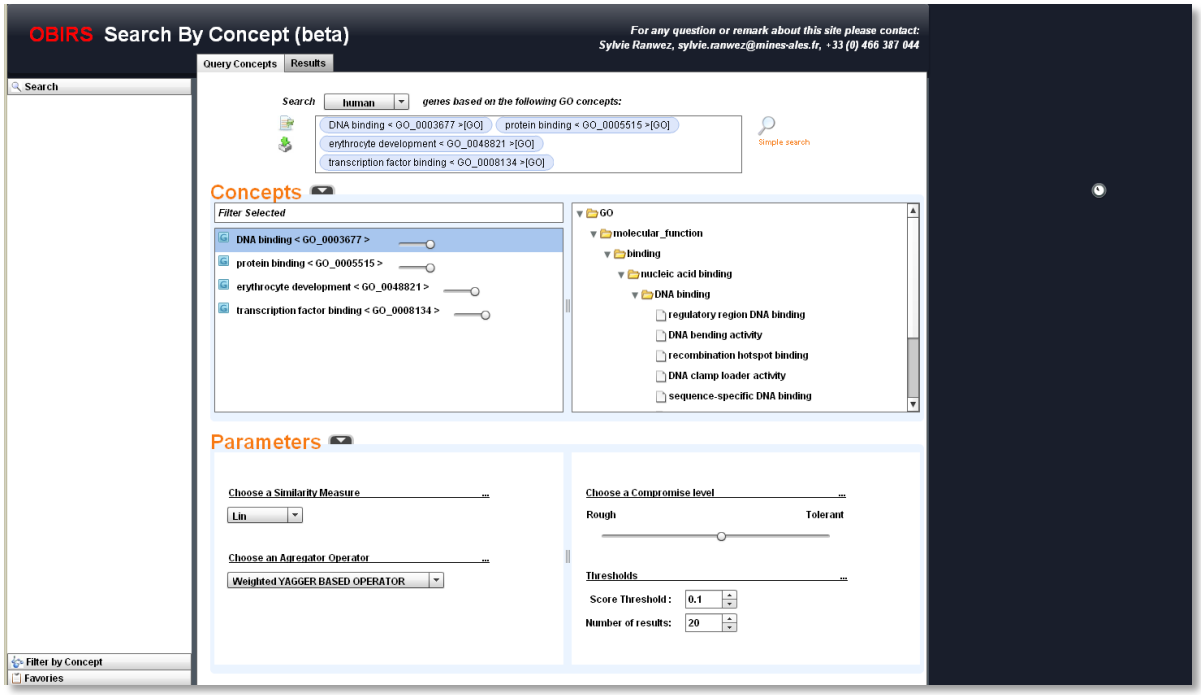


Figure 20 : Interface d'OBIRS permettant d'exprimer une requête
Ici les termes de la requête sont des concepts de la Gene Ontology, dont la saisie est facilitée par une complétion automatique. Une navigation dans l'ontologie est également proposée pour sélectionner des concepts plus ou moins spécifiques. Il est également possible de modifier le mode de calcul de distance sémantique et de préciser si l'on veut un opérateur d'agrégation plutôt conjonctif (proche du ET logique) ou disjonctif (proche du OU).

On notera que ces paramètres sont plutôt destinés à une utilisation experte lors de la phase de tests de l'outil, mais n'est pas à destination des utilisateurs finaux biologistes. Enfin, il est possible de fixer un seuil de

pertinence (RSV) en deçà duquel on ne souhaite pas que les résultats soient affichés ainsi que le nombre maximum de résultats souhaité.

IV.3.2. Construction de la carte sémantique synthétisant les résultats

Une fois la requête exprimée, elle est soumise au moteur qui calcule un score de pertinence pour chaque ressource disponible (ici, chaque gène de la base de données), et les ordonne en conséquence. Celles qui obtiennent les meilleurs scores (leur nombre étant limité au nombre de résultats souhaités par l'utilisateur) sont présentées sur une carte où la requête est positionnée (sous la forme d'un point d'interrogation) et les résultats disposés à une distance de cette requête reflétant leur score de pertinence (c.f. Figure 21). Plusieurs vues sont possibles : soit une présentation linéaire où chaque ligne correspond à une valeur de RSV (c'est cet affichage qui est présenté dans la Figure 21), soit en radial (comme le montre la version d'OBIRS déployée pour l'ITMO Cancer[64] et qui est présentée dans la Figure 22). Ainsi en un coup d'œil, l'utilisateur possède une vision globale des résultats et de leur dispersion par rapport à la requête. Ces résultats peuvent être exportés au format CSV ou en XML. La requête, elle aussi, peut être sauvegardée pour être réutilisée dans des sessions futures.

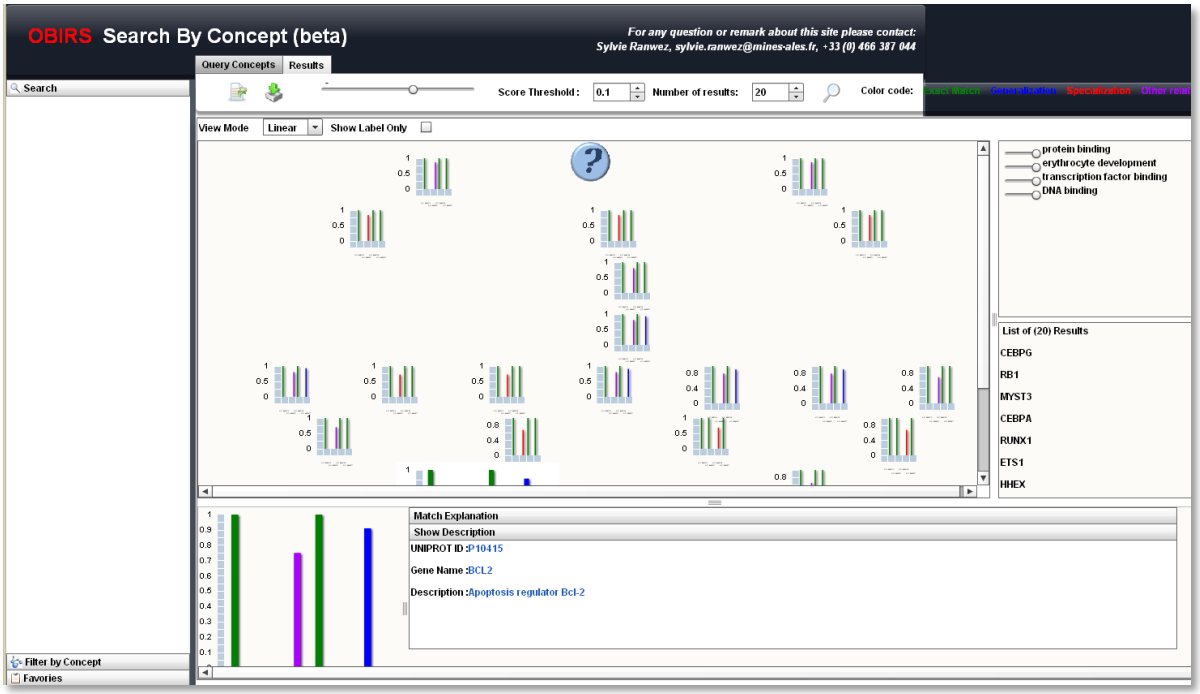


Figure 21 : Interface d'OBIRS présentant les résultats d'une requête

L'utilisateur possède une vue globale des résultats et de leur dispersion par rapport à la requête. Mais pour lui fournir le maximum de justification sur les raisons de la sélection de ces résultats, nous avons proposé une représentation de chaque élément du corpus par un pictogramme.

[64] Portage réalisé par Ansata Dada Balde Sy, et disponible à http://obirs-cancer.mines-ales.fr/ObirsClient/public/Obirs/

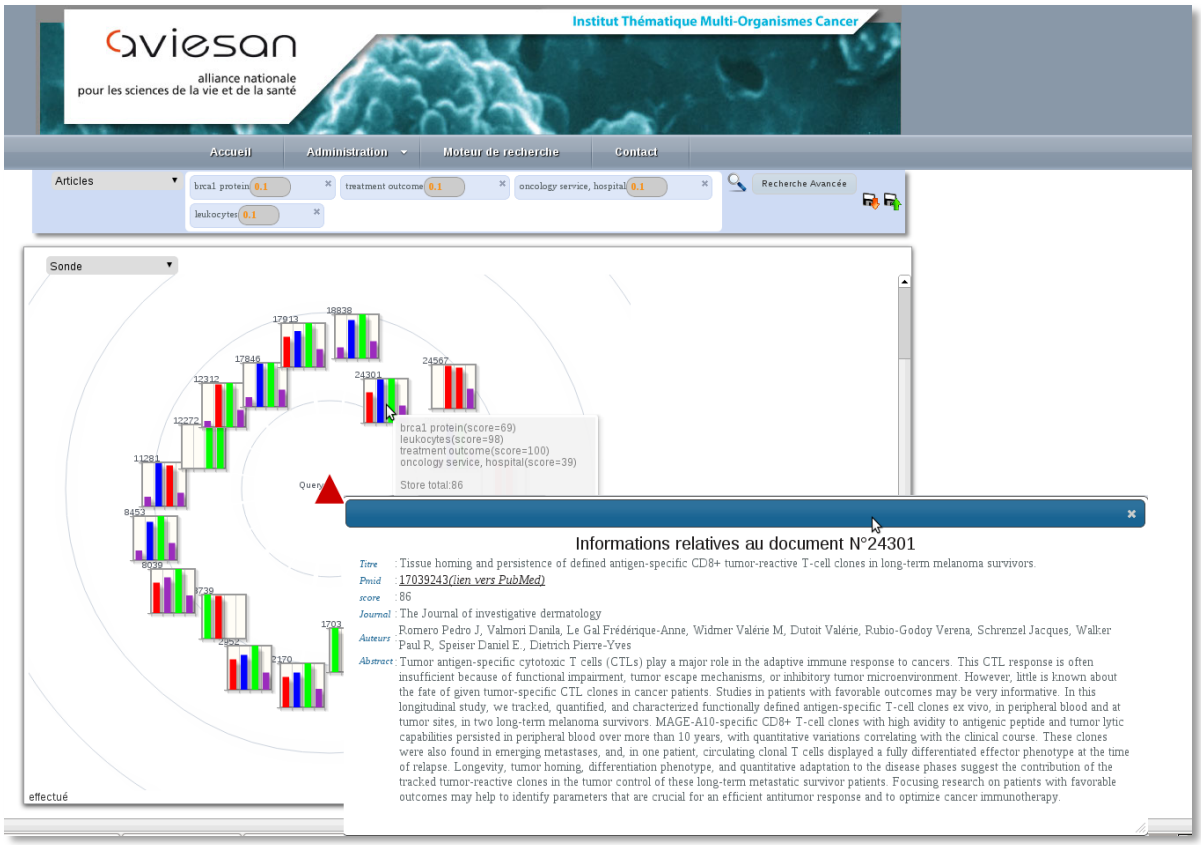


Figure 22 : Interface de l'outil OBIRS, porté sur la plateforme collaborative de l'AvieSan ITMO Cancer
Ici les résultats sont représentés sous forme radiale, avec la requête au centre et les pictogrammes représentant les résultats disposés en fonction de leur pertinence à la requête. Les éléments du corpus sont des publications scientifiques liées au cancer ; lors du passage du curseur de la souris sur un résultat, le détail est affiché dans une fenêtre pop-up.

IV.3.3. Représentation des éléments du corpus dans la carte

Nous avons vu dans la section III.2.1 du chapitre III que le calcul du RSV d'une entité du corpus est composé de trois phases : i) le calcul de la contribution d'un terme de l'indexation par rapport à un terme de la requête, ii) la *pertinence élémentaire* de tous les termes de l'indexation par rapport à un terme de la requête et enfin iii) la pertinence globale de l'entité par rapport à tous les termes de la requête. Si cette pertinence globale peut être estimée visuellement, nous l'avons vu, par le positionnement sur la carte, les pertinences élémentaires sont d'un grand intérêt pour l'utilisateur. Elles permettent, en effet, de comprendre les raisons qui ont conduit à la sélection d'un résultat. C'est pourquoi chaque élément du corpus constituant le résultat su SRI est représenté par un pictogramme synthétisant ces pertinences élémentaires. Ce pictogramme représente un histogramme dans lequel une barre est associée à chaque terme de la requête, la hauteur de la barre indiquant la pertinence élémentaire de l'entité par rapport à ce terme. Une information supplémentaire est fournie par la couleur de la barre. En effet, un terme de la requête et un terme de l'indexation appartiennent à l'ontologie de domaine sur laquelle est appliquée la mesure de proximité sémantique. Il peut y avoir entre les deux un lien[65] de généralisation (la barre est bleue si le terme de la requête subsume un

[65] Ce lien n'est pas forcément direct, il se peut qu'il y ait plusieurs intermédiaires entre les deux, ce qui explique les hauteurs de barre variable.

terme de l'indexation), un lien de spécialisation (la barre est alors rouge) et si la relation n'est pas directe entre les deux (*frères* ou *cousins*) la barre est représentée en violet (c.f. Figure 23). La barre verte est toujours au maximum, quant à elle, puisqu'elle traduit un match exact, autrement dit que le terme de la requête apparaît directement dans l'indexation du document. La carte étant interactive, ces informations sont détaillées lorsque l'utilisateur sélectionne un élément de la carte (par un clic de souris). Ainsi, dans le cadre du bas de l'interface (Figure 21) l'utilisateur peut voir les pertinences élémentaires et les concepts de l'indexation qui ont permis de sélectionner ce résultat. Enfin, si l'utilisateur souhaite avoir plus d'information sur un gène particulier, l'onglet *Show Description* fournit le nom du gène, sa description et un lien vers la base de données UniProtKB[66].

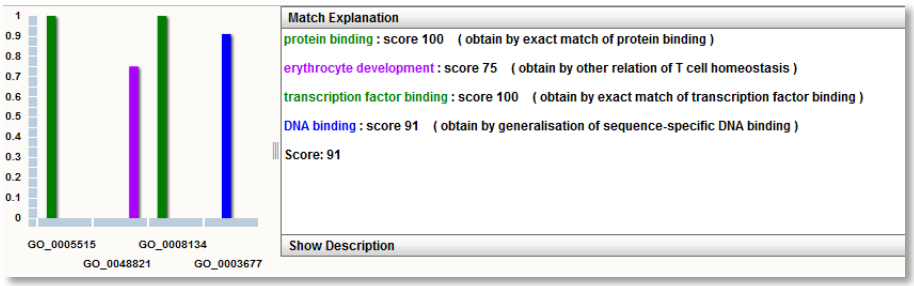


Figure 23 : Détail de l'interface d'OBIRS - Cadre d'explication de la sélection d'un élément du corpus
Les scores élémentaires de chaque terme de la requête sont indiqués et traduits par la hauteur de la barre correspondante dans l'histogramme. Les couleurs indiquent le type de lien entre le terme de la requête considéré et le terme de l'indexation qui lui est le plus proche : généralisation, spécialisation…

Un des avantages procurés par cet affichage détaillé sous forme de pictogrammes est de fournir à l'utilisateur des informations non seulement sur la pertinence globale mais également sur la contribution du document pour chaque terme de la requête. Ainsi, il peut différencier les gènes dont la pertinence globale résulte d'un score très élevé pour un seul terme de la requête, ou au contraire de l'agrégation de plusieurs scores moyens par rapport à l'ensemble des termes de la requête. On perçoit ici la notion de complémentarité qu'il est possible de dégager de ces scores élémentaires.

Par ailleurs, en regardant l'interface de la Figure 21, l'utilisateur peut remarquer que le terme 'protein binding' (identifiant GO_0005515) n'est pas discriminant puisqu'il est présent dans toutes les indexations des gènes fournis en résultat. Il lui est alors possible, dans le cadre qui est en haut à droite de l'écran, où l'on peut voir tous les termes de la requête accompagnés de curseurs, de modifier leur poids respectif dans le calcul du RSV en jouant sur ces curseurs. Le terme reste alors présent mais intervient moins fortement lors de la sélection des résultats. Dès que l'utilisateur modifie les poids, une nouvelle requête est lancée et l'affichage des résultats est mis à jour (la vue va donc changer). Enfin, suite aux tests effectués avec Armelle Regnault (Inserm, ITMO IHP), nous proposons la liste des résultats (dans la partie droite, en bas) afin de donner également les noms (labels) des différents gènes. En cochant la case 'Show labels only' ces labels prennent la place des pictogrammes sur la carte.

La principale contribution de ce travail est donc de fournir à l'utilisateur un outil performant en termes de recherche d'information, couplé avec des capacités d'interaction et de visualisation qui lui permettent de mieux comprendre les résultats qui lui sont présentés. Cette étape est primordiale dans un processus de recherche d'information itératif où il doit reformuler sa requête.

[66] Par exemple http://www.uniprot.org/uniprot/P10415.html

IV.3.4. Identification des passages pertinents

Nous avons couplé l'approche d'OBIRS avec une approche lexicale, dans un environnement hybride nommé CoLexIR permettant d'interroger une base de publications scientifiques (c.f. section 3.1.2 du chapitre 3). Dans ce cas les informations à afficher pour l'utilisateur sont non seulement celles présentées dans la section précédente (histogrammes, etc.) mais également des détails sur les publications sélectionnées, en particulier : leur identifiant dans la base de données PubMed (base de référence pour les publications du domaine biomédical), leur titre, leurs auteurs, le résumé ainsi que les passages dans le texte qui ont été identifiés comme traitant des concepts mentionnés dans la requête (c.f. Figure 24). Ces concepts sont des éléments du MeSH puisque la base de données PubMed s'appuie sur cette ontologie pour indexer les documents.

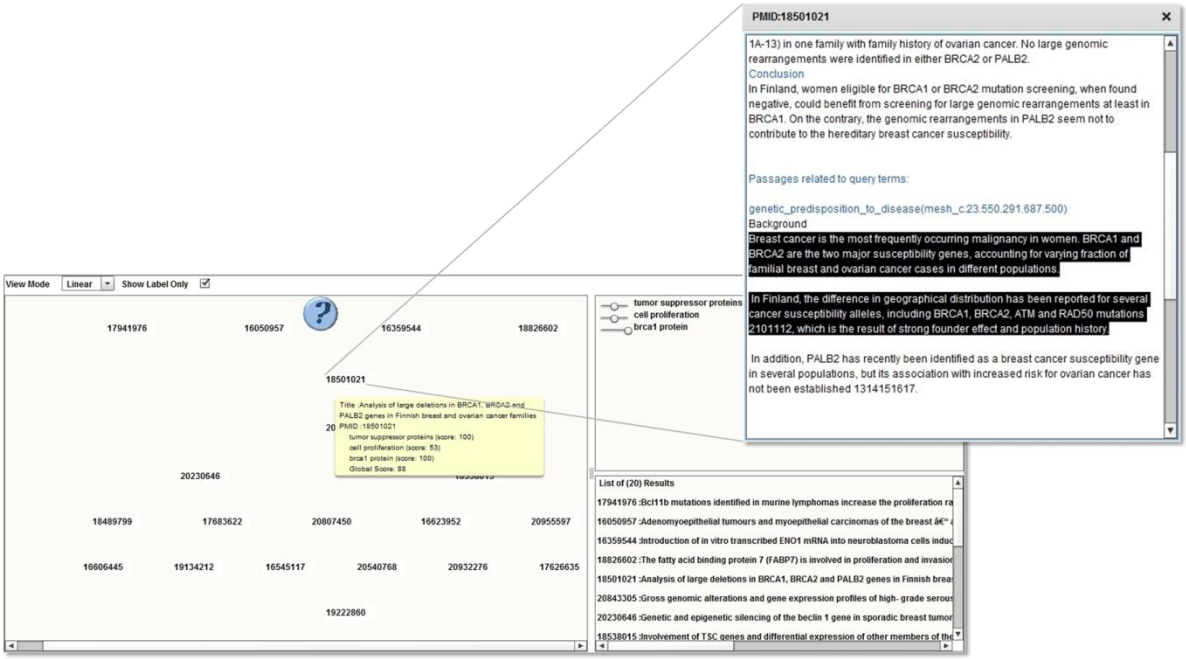


Figure 24 : Interface de présentation des résultats de CoLexIR

IV.4. Synthèse et perspectives

Au cœur de la démarche d'automatisation cognitive, faciliter l'accès aux informations pertinentes représente un enjeu de premier ordre. Submergé par de trop nombreuses informations, souvent hétérogènes, l'opérateur humain a besoin, en effet, d'outils qui synthétisent les informations les plus importantes et les explicitent afin de les rendre rapidement exploitables dans un processus décisionnel. Pour ce faire, les techniques de visualisation et d'interaction homme/machine sont incontournables. En effet, l'automatisation de certaines tâches cognitives à haute valeur ajoutée passe par le contrôle et l'optimisation de l'interaction homme/machine. Or la visualisation est un moyen fondamental pour appréhender l'information et susciter cette interaction. Elle n'est plus considérée seulement comme une simple vue sur des données, mais constitue une base au raisonnement grâce à certaines caractéristiques mises en exergue. Ce chapitre a détaillé différentes appréhensions des cartes de connaissances et de leurs composantes. Elles favorisent généralement une première compréhension rapide et globale d'un ensemble d'entités en aiguillant l'utilisateur sur les actions à mener s'il souhaite une vision plus approfondie. Les indications graphiques proposées explicitent certaines caractéristiques sémantiques des entités et l'interactivité permet à l'utilisateur d'ajuster certains paramètres (pondération, modification d'affichage, lentilles à utiliser…). Mais cette synthèse

visuelle d'information ne doit pas dégrader le contenu initial de l'information. Ainsi, dans la présentation des résultats de l'environnement OBIRS, nous nous sommes efforcés de conserver le plus d'information possible, afin de justifier à l'utilisateur la sélection de certains documents plutôt que d'autres. C'est lui qui doit rester maître de l'application. Même si dans certains cas, une certaine part d'initiative peut être laissée à la machine, celle-ci doit être limitée, comme nous l'avons vu dans le passé (Vaudry, Ranwez, Poulon, & Crampes, 2002).

Si ces techniques de visualisation sont nécessaires pour assister l'opérateur dans son processus décisionnel, nous pensons qu'elles sont également d'un grand intérêt dans la communauté de l'ingénierie des connaissances. En effet, si, poussées par le développement du Web sémantique, les ontologies ont suscité un vif intérêt et se sont multipliées dans de nombreux domaines, les outils pour les visualiser restent très limités. Or que ce soit lors de leur conception, de leur mise à jour ou au cours de leur exploitation, il est essentiel de disposer de représentations synthétiques, focalisées sur certaines parties de l'ontologie ou mettant en relief certaines de leurs caractéristiques. Nos travaux préliminaires dans cette direction ont démontré la nécessité d'utiliser des techniques de filtrage efficace (Jalabert, Ranwez, Crampes, & Derozier, 2006; Jalabert, Ranwez, Derozier, et al., 2006). L'extraction de sous-ontologies que nous avons proposée est un premier élément de réponse (V. Ranwez, Ranwez, et al., 2012) qui, couplée avec une projection de type MDS (intégrant la prise en compte de distances sémantiques entre concepts), permettrait d'avoir plusieurs niveaux de visualisation des ontologies, comme nous l'avons évoqué dans (S. Ranwez et al., 2006). C'est une des perspectives développées dans le chapitre suivant.

CHAPITRE V

"Préparer l'avenir ce n'est que fonder le présent. […] Il n'est jamais que du présent à mettre en ordre. À quoi bon discuter cet héritage. L'avenir, tu n'as point à le prévoir mais à le permettre."

Antoine de Saint Exupéry, "Citadelle".

Conclusion, perspectives et projet de recherche

V.1. PERSPECTIVES DE RECHERCHE 112

V.1.1. INDEXER RAPIDEMENT UN GRAND NOMBRE DE RESSOURCES 112

V.1.2. PROPOSER DE NOUVEAUX MODES DE RECHERCHE D'INFORMATION 113

V.1.2.1. Rechercher des informations complémentaires 113

V.1.2.2. Approfondir les mesures de distance/similarité sémantique pour les exploiter plus efficacement en fonction du contexte applicatif 114

V.1.3. INSTRUMENTER L'AUTOMATISATION COGNITIVE : VISUALISATION, INTERACTION ET PERSONNALISATION AU CŒUR DE LA DEMARCHE 115

V.2. POSITIONNEMENT AU SEIN DE L'ECOLE NATIONALE SUPERIEURE DES MINES D'ALES ET DE L'INSTITUT MINES-TELECOM 116

V.2.1. PERSPECTIVES CONCERNANT LA TRANSMISSION DES SAVOIRS 116

V.2.2. INNOVER, CREER AU SERVICE DU DEVELOPPEMENT ECONOMIQUE ET SOCIAL. 117

V.3. UN DOMAINE D'APPLICATION PRIVILEGIE : LE DOMAINE BIOMEDICAL ET LA SANTE 117

Les quatre premiers chapitres de ce manuscrit ont présenté une synthèse de mes activités de recherche passées et des résultats obtenus. Il convient maintenant de se tourner vers l'avenir et de définir les perspectives ouvertes que je souhaite explorer en priorité, les contributions que j'envisage bénéficiant à différentes communautés : ingénierie des connaissances, recherche d'information et visualisation. Bien sûr, ces perspectives s'inscrivent pleinement dans les objectifs de l'équipe KID du centre de recherche LGI2P et s'intéressent aux différentes phases d'extraction, d'enrichissement et d'exploitation de connaissance dans un processus décisionnel (c.f. Figure 25). C'est pourquoi je détaille les perspectives que j'envisage en fonction des trois axes principaux qui ont structuré ce mémoire : indexation, recherche d'information, et visualisation/interaction homme-machine. Ces perspectives sont ensuite mises en contexte au sein de l'Institut Mines-Télécom.

Figure 25 : Chaîne de traitement de la donnée à la décision

V.1. Perspectives de recherche

Cette section détaille les perspectives à court, moyen et plus long termes, que j'envisage pour mes travaux de recherche.

V.1.1. Indexer rapidement un grand nombre de ressources

A la lecture du chapitre III, on peut remarquer que les systèmes de recherche d'information sont sensibles à plusieurs paramètres : la ou les mesures sémantiques utilisées entre concepts, les fonctions d'agrégation choisies pour fusionner ces mesures à un niveau d'abstraction supérieur, le tout dépendant fortement de la qualité de l'indexation.

Nos contributions en matière d'indexation ont concerné la définition de critères de qualité qui peuvent être considérés comme des exigences à respecter si l'on souhaite optimiser l'exploitabilité et la réutilisabilité d'une indexation. Nous avons ensuite proposé une méthode d'indexation par propagation dans différents contextes et pour différents types de données numériques. Cette indexation par propagation s'est avérée performante dans un système exploitant une représentation vectorielle où chaque ressource était représentée par son vecteur de satisfaction à différents critères.

Le principe de l'indexation par propagation répond à plusieurs de nos attentes, en particulier parce qu'il tire parti de l'interaction homme-machine pour produire rapidement un grand nombre d'indexations. Son application à l'indexation conceptuelle n'est pas immédiate et différentes limites ont été soulevées dans la section II.5. On envisage cependant de développer un environnement d'indexation permettant d'attribuer rapidement un ensemble de concepts d'une ontologie de domaine à différent types de ressources intégrant les propositions suivantes.

Concernant les données conceptuelles à propager, le principal problème consiste à ne pas propager trop de concepts pour éviter d'avoir la cardinalité de l'ensemble des concepts de l'indexation qui augmente de manière déraisonnable, réduisant ainsi sa pertinence. Dans ce qui suit, nous restreignons l'ontologie au *dag* induit par la relation *isa*. Une première idée consiste à extraire la sous-ontologie des *k*-plus-proches voisins et de ne conserver que les feuilles de cette sous-ontologie dans l'indexation (les ancêtres étant, de fait, contenus implicitement). Cependant on perd dans ce cas la fréquence d'apparition des concepts dans les indexations des *k*-plus-proches voisins (appelés concepts *indexant* dans la suite). On peut affiner cette méthode, en comparant ces différents concepts : si la distance sémantique d'un concept indexant à tous les autres est supérieure à un seuil, ce concept n'est pas pertinent et on lui attribue un poids faible dans l'indexation en cours de construction. Au contraire, si un concept est très proche (voire égal) à d'autres concepts indexant, il aura un poids renforcé dans cette indexation. Par ailleurs un indice de confiance peut être attribué à l'indexation obtenue (sur la même idée que l'*evidence code* proposé pour la Gene Ontology) : cet

indice étant proportionnel à la distance des *k*-plus-proches voisins. Différents tests devraient nous permettre de fixer une limite à la propagation afin de fixer la valeur de *k*.

Les qualités énoncées pour l'indexation permettent de fixer certains objectifs et donc certaines règles à respecter. Afin d'atteindre ces objectifs, une boucle de rétroaction peut être appliquée pour contrôler ces différentes règles.

La propagation telle que nous venons de la décrire propage les valeurs d'indexation en fonction de la distance sémantique des concepts *indexant* par rapport à l'ontologie du domaine. Un autre type de propagation peut être envisagé. Celui-ci n'implique pas d'interaction visuelle, il consiste, lorsque l'utilisateur saisit certains concepts pour indexer le document, à lui proposer d'autres concepts qui statistiquement apparaissent souvent en cooccurrence avec celui saisi. Même si ce type d'approche n'implique pas de traitement automatique élaboré, nous envisageons de l'inclure dans notre système d'indexation, comme alternative à l'interface visuelle. Cette alternative est notamment pertinente dans le cas où l'utilisateur n'a pas une bonne maîtrise des documents déjà indexés et a donc des difficultés à *déposer* un document sur un *paysage* de documents déjà indexés.

Enfin, l'indexation par propagation, malgré ses avantages, n'est pas applicable à de très grands volumes de ressources, du fait qu'elle nécessite une intervention humaine experte. Une alternative doit donc être imaginée. Pour les documents textuels, nous envisageons d'appliquer une première phase d'analyse textuelle afin de détecter certains champs lexicaux associés à des concepts dans ces documents et les indexer en conséquence, avec ces concepts. Cela suppose une phase d'apprentissage conséquente, qui permet d'associer ces champs lexicaux aux concepts, comme nous l'avons évoqué dans la section III.3.

V.1.2. Proposer de nouveaux modes de recherche d'information

Le deuxième axe de recherche envisagé concerne la recherche d'information. Nos contributions dans ce domaine concernent l'exploitation d'une ontologie de domaine pour retrouver des documents pertinents et l'utilisation de techniques d'agrégation particulières pour favoriser la justification des résultats. L'environnement OBIRS qui a été développé est utilisé, entre autres, pour la valorisation des résultats du projet ANR PhylARIANE (http://www.lirmm.fr/phylariane/). Par ailleurs, nous avons envisagé la reformulation comme une fonction objectif et la recherche d'un optimum de cette fonction. Ces travaux ont été validés et leur efficacité démontrée, cependant on se situe ici dans un cadre relativement classique de recherche d'information qui ne s'adapte pas forcément à tous les contextes. C'est pourquoi nous nous intéressons maintenant aussi à la recherche d'information complémentaire.

V.1.2.1. Rechercher des informations complémentaires

Dans la gestion de communautés scientifiques, par exemple, il arrive souvent que les instances de gouvernance veuillent constituer des groupes d'experts répondant à certains objectifs : réponse à un appel d'offre, création d'une commission d'évaluation, etc. Dans ce cas, il faut des profils d'expertise suffisamment différents pour ne pas être redondant, mais suffisamment proches pour pouvoir échanger, communiquer. C'est un des contextes envisagés pour la recherche par complémentarité.

La diversification consiste à proposer à l'utilisateur qui recherche de l'information des ressources qui répondent à la requête, mais qui sont les plus éloignés possible les unes des autres. De manière très synthétique, l'état de l'art montre que les différentes techniques employées (*distance based* et *coverage based*, par exemple) visent à maximiser le calcul du RSV et la distance entre les documents retournés. Ainsi la diversification est souvent envisagée comme une phase postérieure à la recherche d'information, qui consiste à réordonner les documents retrouvés. La complémentarité peut être envisagée comme une forme particulière de diversification, avec une contrainte supplémentaire sur les ressources sélectionnées, qui concerne leur recouvrement partiel. En effet,

pour être jugées complémentaires, deux ressources doivent avoir un certain taux de recouvrement : par exemple pour des experts, ils ne doivent pas être totalement étrangers aux thématiques des autres membres du collectif, sous peine de ne pas pouvoir communiquer ensemble et donc ne pas être productifs. Il nous semble intéressant d'envisager la complémentarité en amont de la recherche d'information, lors du calcul du RSV. Celui-ci implique des mesures sémantiques. Or, dans le cadre abstrait proposé par Sébastien Harispe, ces mesures sont toutes fonction d'un degré de ressemblance (*commonality* ou le néologisme *commonalité*) et de spécificité (*spécificity*). Le calcul du *RSV complémentaire* pourrait donc intégrer ces différentes composantes pour sélectionner les ressources, différentes approches pouvant être utilisées pour estimer la commonalité et la spécificité : par exemple la première peut être fonction de la cooccurrence des concepts dans l'indexation des ressources, la deuxième pouvant être déduite à partir du degré de similarité sémantique des concepts de l'indexation dans l'ontologie de domaine. Nous envisageons de définir et formaliser la complémentarité et de tester ce type d'approche en intégrant éventuellement plusieurs ontologies de domaine.

Enfin, en ce qui concerne les documents textuels, nous souhaitons utiliser les résultats de recherche de Benjamin Duthil (approche *Synopsis*), afin de rechercher des ressources qui n'auraient pas été indexées au préalable par des concepts. Cette notion est à rapprocher du peuplement d'ontologies et les résultats obtenus dans ce domaine pourraient donc également bénéficier à la communauté d'ingénierie des connaissances. Même si dans certains cas les temps de traitement sont limitants, une telle approche pourrait être envisagée en complément des approches citées ci-dessus. L'analyse lexicale n'étant pas ma spécialité, ce travail ne pourra être envisagé qu'en collaboration avec d'autres collègues de l'équipe KID.

V.1.2.2. Approfondir les mesures de distance/similarité sémantique pour les exploiter plus efficacement en fonction du contexte applicatif

Au cours de la chaîne de traitements allant de l'indexation à l'exploitation et à la visualisation d'informations, les mesures sémantiques jouent un rôle clé. Le cadre abstrait proposé par Sébastien Harispe a permis d'en caractériser les principales composantes et donc de sélectionner de façon plus efficace, les mesures les plus appropriées à un contexte donné. La librairie qui a été développée (SML[67]) suscite un intérêt dans la communauté et commence à être utilisée par d'autres équipes de recherche. Il convient maintenant d'envisager des études supplémentaires, afin d'étudier l'impact de ces différentes mesures et des différentes stratégies d'agrégation utilisées en RI lors du calcul du RSV. Dans cette étude, une réflexion concernant une possible pondération des concepts doit être menée. Ces poids peuvent être ceux attribués par l'utilisateur au moment de la formulation de la requête, afin de donner plus ou moins d'importance à un aspect de sa requête. Cette pondération est déjà prise en compte dans OBIRS. Mais une autre pondération peut être appliquée au niveau de l'indexation des ressources. Dans la section V.1.1, nous avons vu que cette pondération peut traduire un degré de fiabilité de l'indexation ou dépendre de la fréquence d'apparition d'un concept dans une ressource (textuelle principalement). La prise en compte de cette pondération est délicate. La première raison provient de son origine : s'agit-il de poids affectés lors de l'indexation par un opérateur humain, auquel cas, ils sont relativement subjectifs, ou bien de poids attribués de façon automatique ? Dans ce dernier cas, si les poids ont été affectés après analyse de l'ontologie de domaine (par exemple proportionnellement à un IC), comment les prendre en compte en particulier dans les cas où plusieurs ontologies peuvent être utilisées simultanément ? Cette question rejoint nos travaux actuels avec l'université Rovira i Virgili de Tarragone concernant les mesures de similarité sémantique basées sur plusieurs ontologies de domaine.

L'intégration de notre méthode de reformulation (section III.2.2) dans OBIRS a sensiblement amélioré la précision et le rappel et ceci sans dégrader significativement les temps de réponse

[67] http://www.semantic-measures-library.org/sml/

(quelques secondes, même dans le cas d'ontologies volumineuses). Ceci est dû à l'approche heuristique que nous avons adoptée, qui réduit l'espace de recherche à un voisinage autour des concepts de la requête initiale. Cependant cet axe doit être approfondi. En particulier, l'impact du choix des mesures de similarité appliquées à la reformulation de requête mérite d'être étudié, ainsi que la possibilité d'intégrer une pondération dans cette reformulation (en suivant l'idée de la stratégie DFR, *Divergence From Randomness*). Notre approche de la reformulation basée sur des ontologies de domaine définit une nouvelle famille de stratégies intégrant un modèle de préférences de l'utilisateur, dans lequel il est possible de paramétrer la stratégie d'agrégation choisie, et les contributions respectives des *bons* et *mauvais* documents. Les mesures sémantiques jouent là encore un rôle central.

Notre équipe possède une bonne maîtrise des mesures de similarité sémantique qui utilisent la restriction des ontologies au graphe induit par la relation *is-a*. Or dans certains cas, il est nécessaire de prendre en compte d'autres types de relation. Ceci est particulièrement vrai dans le cadre des données liées (*linked data*) qui utilisent les technologies du Web sémantique pour décrire des bases de connaissance où concepts, instances et données sont mis en relation au travers de liens sémantiques typés. C'est ce qui nous a conduits récemment à nous intéresser à la définition d'un cadre pour définir de nouvelles mesures sémantiques pour la comparaison d'instances exprimées dans une base de connaissances RDF. Les premiers résultats de ces travaux appliqués aux systèmes de recommandation font l'objet d'une soumission. Nous souhaitons approfondir cette étude et l'appliquer à d'autres domaines comme la prédiction d'interactions protéiques dans le domaine biomédical. Cet axe de recherche est d'autant plus important que dans le cadre des plateformes collaboratives, il nous semble stratégique de nous tourner vers les technologies liées à l'open data plutôt que de nous limiter à des bases de données internes. En effet, les communautés identifiées dans ces plateformes sont géographiquement dispersées et les données concernant les individus sont réparties. Exploiter les liens entre ces différentes ressources permettrait de répondre à certains besoins exprimés par la gouvernance des différents instituts de l'AvieSan.

V.1.3. Instrumenter l'automatisation cognitive : visualisation, interaction et personnalisation au cœur de la démarche

En créant des outils de traitement automatique des données, nous n'avons pas la volonté d'exclure l'homme de la démarche, mais au contraire de lui faciliter certaines tâches en favorisant des mécanismes d'automatisation cognitive. La visualisation et l'interaction sont donc des éléments incontournables de notre approche. La visualisation que nous proposons n'est pas simplement une vue quelconque sur les données, elle constitue le support du raisonnement humain, en mettant en exergue certaines propriétés ou en suggérant des rapprochements ; elle intègre les différentes composantes des modèles sous-jacents. Ainsi les cartes sémantiques que nous avons proposées dans différents contextes ont montré leur efficacité que ce soit pour la navigation ou pour l'indexation par propagation, pour citer quelques exemples. Mais leur usage, s'il s'avère relativement intuitif, bouleverse néanmoins certaines habitudes de travail, comme nous avons pu le constater dans le projet MBox-*composer* pour les DJ. En particulier, la notion de distance sémantique peut être déroutante si le positionnement sur la carte n'est pas clairement spécifié. Nous souhaitons poursuivre nos travaux dans ce sens, en organisant des séances d'évaluation à plus grande échelle et en faisant intervenir différents profils d'utilisateurs. Cette évaluation permettrait d'améliorer le rendu visuel en l'adaptant à ces profils (les interactions avec Armelle Regnault (plateforme IHP de l'AvieSan), nous avaient permis d'améliorer considérablement l'interface d'OBIRS mais d'autres modifications sont envisageables). Or, à mon sens, ce n'est que par une forte utilisation des cartes sémantiques que nous pourrons récolter un maximum de retours utilisateurs pour identifier les manques et y répondre.

Les travaux sur la visualisation que nous envisageons ne concernent pas seulement les cartes sémantiques, mais également les cartes conceptuelles. Comme nous l'avons vu, ces cartes représentent les concepts modélisés dans les ontologies de domaine. Elles sont donc un moyen

d'appréhender ces ontologies, de manière différente des traditionnels arbres d'exploration "à la Windows". Cette visualisation pourrait intervenir dans différentes phases de l'ingénierie des connaissances comme la conception et la mise à jour des ontologies ou leur exploitation, par exemple en extrayant des sous-ontologies ou en proposant des vues personnalisées en fonction du profil des utilisateurs.

V.2. Positionnement au sein de l'école des mines d'Alès et de l'institut Mines-Télécom

Le paysage des écoles d'ingénieurs Françaises a subi une profonde mutation cette année, avec le regroupement des écoles des Mines et des écoles des Télécom dans un institut unique : l'Institut Mines-Télécom. Ce regroupement place l'institut au premier rang national des écoles d'ingénieurs et de management avec comme devise: "Former Innover Créer au service du développement économique et social". Cette section détaille mon positionnement dans cet institut.

V.2.1. Perspectives concernant la transmission des savoirs

Les orientations stratégiques de l'institut concernent en tout premier lieu la transmission des savoir : "former les ingénieurs des décennies à venir". Or les profils des étudiants qui accèdent aux formations proposées dans l'institut, et en particulier à l'école nationale supérieure des mines d'Alès, évoluent. L'impact des nouvelles technologies dans notre rapport à la connaissance, la modification de nos habitudes de vie, sont autant de facteurs qui nécessitent une mutation de nos modes d'enseignement. C'est pourquoi je participe activement à un groupe de réflexion sur les innovations pédagogiques au sein de l'EMA. Cette réflexion vise autant la définition du contenu de nos interventions, que les méthodes d'enseignement (en particulier la répartition des étudiants et l'intégration de l'autoformation et des supports accessibles en ligne) ou la façon d'évaluer les acquis. L'objectif est de favoriser un accompagnement fortement personnalisé mais sans surcharge de travail pour les enseignants. Il paraît nécessaire, pour cela, de mutualiser les efforts, de proposer aux étudiants des supports suffisamment riches pour qu'ils soient exploités efficacement en autoapprentissage, afin que le temps d'encadrement effectif des enseignants se concentre sur la mise en application et les *études de cas*. Ma participation depuis 2007 à l'enseignement de l'algorithmique et de la programmation en formation continue diplômante des écoles des mines, me procure une expérience qui va dans ce sens. Néanmoins, cela suppose que les supports soient particulièrement bien adaptés, conçus et rédigés spécifiquement pour un apprentissage en autoformation. Le temps nécessaire à cette préparation ne doit pas être négligé. Cette mutualisation pourrait bénéficier aux autres écoles de l'Institut Mines-Télécom et je participe également à une réflexion plus globale concernant les MOOCS (*Massive Open Online Courses*) au sein de ce groupe.

Par ailleurs, je m'investis depuis de nombreuses années dans des enseignements d'initiation en tronc commun. Plusieurs raisons à cela, la première étant sans doute que ma thématique de recherche était trop jeune et trop spécifique pour être intégrée à des cursus d'ingénieurs généralistes. Cependant comme nous l'avons vu dans le chapitre d'introduction, les ontologies font désormais l'objet d'une attention toute particulière dans le monde industriel. Les compétences liées à l'ingénierie des connaissances sont de plus en plus demandées. C'est donc tout naturellement que cette thématique a été intégrée au nouveau département de spécialisation des étudiants de formation initiale de l'EMA : EMACS (*Engineering and MAnagement of Complex Systems*). Je souhaite donc dans un futur proche me focaliser sur ces enseignements plus spécialisés et réduire mes interventions en tronc commun. Ces nouveaux enseignements, en offrant aux étudiants les fondements des ontologies et des techniques de base du Web Sémantique, me donneront, je l'espère, l'occasion de proposer également des sujets de projets étudiants sur ces thématiques. De ce fait, je rapprocherai mes thématiques de recherche de mes thématiques d'enseignement, afin qu'elles se nourrissent mutuellement. De plus, concernant la mise en place de ce nouveau département, nous essayons

d'avoir une approche transverse des différents modules, pour que les activités de recherche de l'équipe transparaissent dans la formation : envisager la décision depuis l'analyse et le traitement des données jusqu'à la visualisation (on retrouve ici le schéma de la Figure 25). Cette collaboration avec les autres enseignants est très stimulante et nous permettra, je l'espère, de communiquer cette motivation aux étudiants. Enfin, la formation par apprentissage InfRes (Informatique et Réseau) étant amenée à se développer, je souhaite également m'investir davantage dans les enseignements qu'elle offre aux étudiants, par exemple en proposant une adaptation du module Web Sémantique dans cette formation.

V.2.2. Innover, créer au service du développement économique et social.

Historiquement en lien étroit avec les entreprises, les laboratoires de recherche des écoles des Mines ont fait du transfert de technologie une priorité. Ainsi, outre les projets avec de plus grands groupes industriels, une partie de notre activité au centre de recherche LGI2P concerne l'accompagnement des PME locales et des créateurs lors du démarrage de leur activité. Dans son développement stratégique, l'Institut Mines-Télécom souhaite développer "l'innovation, en ingénierie et management, au service de l'industrie, de l'économie et de la société, dans ses grandes écoles ancrées dans leur territoire, ouvertes sur le monde et étroitement liées aux entreprises". Je souhaite donc poursuivre cette facette de mon activité, en apportant mes conseils aux créateurs accueillis et en réfléchissant avec eux à la proposition d'approches innovantes en matière d'intégration de l'ingénierie des connaissances dans les applications industrielles. Parmi les neuf secteurs stratégiques identifiés par l'IMT, je m'inscris dans "informatique et composants", "média et services", ainsi que sur l'axe" santé et autonomie" puisqu'il correspond au domaine d'application que nous privilégions depuis plusieurs années (c.f. section suivante).

V.3. Un domaine d'application privilégié : le domaine biomédical et la santé

Même si nous avons toujours adopté des approches génériques, les résultats des travaux de recherche que j'envisage trouvent un domaine d'application privilégié dans le domaine de la santé et du biomédical. Ceci se justifie d'une part par les besoins exprimés par nos partenaires. En effet, impliqués dans les actions de recherche en lien avec les plateformes collaboratives de l'AvieSan et de l'Inserm, les applications que nous avons développées (OBIRS, OBIRS-feedback, CoLexIR) s'inscrivent dans ce domaine. D'autre part, de nombreuses ontologies de domaine et bases de données indexées sont disponibles et accessibles librement. Les ressources contenues dans ces bases sont de plusieurs types (gènes, publications scientifiques, protéines, etc.) et l'application d'approches conceptuelles se justifie pleinement si l'on souhaite créer des liens entre elles.

Concernant la valorisation de la SML, nous étudions la possibilité de l'utiliser pour proposer un ensemble de services sur les ontologies présentes dans Bioportal (http://bioportal.bioontology.org/). Cette action sera menée en collaboration avec Clément Jonquet du LIRMM (Montpellier). Par ailleurs, l'application des mesures de similarité basées sur RDF et utilisant différent types de relations permettraient d'inférer de la connaissance à partir de graphes de grande taille. En étudiant les interactions entre concepts et instances on peut déduire certaines interconnections et trouver ainsi qu'une molécule peut être en lien avec telle ou telle maladie, tel médicament, etc.

Dans la même lignée, depuis quelques mois, nous sommes en cours de montage d'une chaire industrielle qui, au sein de l'Institut Mines-Télécom, rassemblerait chercheurs, industriels, et gestionnaires de santé, autour d'un projet commun de médecine numérique. Ce projet concernerait principalement l'exploitation de documents textuels, mais des interconnections avec les autres données biomédicales (gènes, protéines, molécules, etc.) ne sont pas exclues.

Au terme de ce manuscrit, et après un bilan des actions de recherche menées ces dernières années, les perspectives sont nombreuses. De par sa nature, l'ingénierie des connaissances est au centre de tout système complexe impliquant outillage informatique et opérateurs humains. Aussi, je projette de tout mettre en œuvre pour que la dynamique initiée autour de la gestion des connaissances et leur exploitation dans un processus décisionnel se poursuive dans l'équipe KID et en collaboration avec des chercheurs extérieurs dans le cadre de projets de plus grande envergure dans lesquels je serais amenée à prendre des responsabilités et à coordonner des équipes.

Références bibliographiques

Abdelali, A., Cowie, J., & Soliman, H. S. (2007). Improving query precision using semantic expansion. *Information Processing & Management*, *43*(3), 705–716.

Abel, M. H., Bach, T. L., Dehors, S., Dieng-Kuntz, R., Gandon, F., Luong, P. H., & Moulin, C. (2005). Ontologies pour le Web Sémantique et le e-Learning. In AFIA/INRIA (Ed.), *Journée thématique : Web sémantique pour le e-Learning*. Nice.

Abrouk, L., Gouaïch, A., & Raïssi, C. (2005). Annotation automatique de documents - Analyse des citations. *INFORSID 2006* (pp. 483–497). Hammamet, Tunisie.

Andreasen, T., Bulskov, H., & Knappe, R. (2003). On ontology-based querying. In H. Stuckenschmidt (Ed.), *18th International Joint Conference on Artificial Intelligence: Ontologies and distributed systems, IJCAI 2003*. Acapulco, Mexico.

Baar, S. (2004). *Préconisations pour la création d'un module d'indexation dans un outil de développement d'applications Web* (p. 84). Nîmes.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books)* (p. 944). Harlow, Essex: Addison-Wesley Professional.

Baziz, M. (2005). *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. Université Paul Sabatier.

Baziz, M., Aussenac-Gilles, N., & Boughanem, M. (2003). Exploitation Des Liens Sémantiques Pour l'Expansion De Requêtes Dans Un Système De Recherche d'Information. *Actes du XXIe Congrès INFORSID* (pp. 121–134). Nancy, France.

Baziz, M., Boughanem, M., Pasi, G., & Prade, H. (2005). A Fuzzy Set Approach to Concept-based Information Retrieval. *4th Conference of the European Society for Fuzzy Logic and Technology and the 11ème Rencontres Francophones sur la Logique Floue et ses Applications (Eusflat-LFA 2005 joint Conferences)*. Barcelona, Spain.

Belkin, N., Ingwersen, P., & Pejtersen, A. M. (1992). Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (p. 352). Copenhagen, Denmark: ACM.

Bender, M. A., Pemmasani, G., Skiena, S., & Sumazin, P. (2001). Finding least common ancestors in directed acyclic graphs. *Proceedings of the Twelfth Annual Acm-Siam Symposium on Discrete Algorithms*, 845–854.

Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications (Second Edition)* (Springer S., p. 614). New York: Springer.

Buja, A., & Swayne, D. F. (2002). Visualization methodology for multidimensional scaling. *Journal of Classification*, *19*(1), 7–43.

Calvier, F.-E., Plantié, M., Dray, G., & Ranwez, S. (2013). Ontology Based Machine Learning for Semantic Multiclass Classification. *Terminology & Ontology : Theories and applications - TOTH 2013*. Chambery, France.

Chabert-Ranwez, S. (2000). Composition Automatique de Documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes Intentionnelles de l'Utilisateur. *Informatique*. Nîmes: Montpellier II University.

Chen, W., & Mizoguchi, R. (1999). Communication Content Ontology For Learner Model Agent in multi-Agent Architecture. *Workshop on Ontologies for Intelligent Educational Systems, 9th International Conference on Artificial Intelligence in Education, AI-ED'99*. Le Mans, France.

Chevillotte, S. (2005). Les langages documentaires. Lyon: enssib.

Crampes, M., De Oliveira-Kumar, J., Ranwez, S., & Villerd, J. (2009). Indexation de photos sociales par propagation sur une hiérarchie de concepts. In F. Gandon (Ed.), *actes des 20es Journées Francophones d'Ingénierie des Connaissances* (pp. 13–24). Hammamet, Tunisie.

Crampes, M., De Oliveira-Kumar, J., Ranwez, S., & Villerd, J. (2009). Visualizing Social Photos on a Hasse Diagram for Eliciting Relations and Indexing New Photos. *Ieee Transactions on Visualization and Computer Graphics*, *15*(6), 985–992.

Crampes, M., Gout, O., Mille, N., Villerd, J., & Ranwez, S. (2007). Cross-Linking Music and Pictures Through Moods. *Proceedings of the 2007 International Conference on Multimedia Systems and Applications, MSA 2007, June 25-28, 2007, Las Vegas Nevada, USA*, 137–144.

Crampes, M., Ranwez, S., Plantié, M., & Vaudry, C. (2003). Qualités d'une indexation portée par XML et une ontologie au regard d'un standard. . (E. Bruillard & B. de La Passardière, Eds.)*Sciences et techniques éducatives, Hors série 2003, "Ressources numériques, XML et éducation"*, 105–134.

Crampes, M., Ranwez, S., Velickovski, F., Mooney, C., & Mille, N. (2006). An integrated visual approach for music indexing and dynamic playlist composition - art. no. 607103. *Multimedia Computing and Networking 2006, 6071*, 7103.

Crampes, M., Ranwez, S., Villerd, J., Velickovski, F., Mooney, C., Emery, A., & Mille, N. (2006). Concept maps for designing adaptive knowledge maps. *Information Visualization*, *5*(3), 211–224. doi:10.1057/palgrave.ivs.9500127

Crouch J., C., & Yang, B. (1992). Experiments in automatic statistical thesaurus construction. *15th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 77–88). Copenhagen, Denmark: ACM.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Dewey, M. (1876). *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library* (p. 42).

Duthil, B., Roche, M., Montmain, J., & Poncelet, P. (2011). Towards an automatic characterization of criteria. *DEXA'2011*.

Duthil, B., Trousset, F., Dray, G., Montmain, J., & Poncelet, P. (2011). Vers une caractérisation automatique de critères pour l'opinion-mining. *Les Cahiers du numérique*, *7*(2), 41–62.

Eades, P. (1984). A heuristic for graph drawing. *13th Manitoba Conference on Numerical Mathematics and Computing* (Vol. 24, pp. 149–160).

Farah, M., & Vanderpooten, D. (2007). L'Agrégation en Recherche d'Information. Une revue critique des principaux modèles théoriques de Recherche d'Information. *CORIA 2007* (pp. 125–136). Saint Etienne, France.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: Mathematical foundations*. Berlin-Heidelberg: Springer.

Gomez-Pérez, A., Fernandez-Lopez, M., & Corcho, O. (2004). *Ontological engineering - with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. *Advanced information and knowledge processing* (p. 403). London ; Berlin ; Heidelberg: Springer-Verlag.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Haav, H.-M., & Lubi, T.-L. (2001). A Survey of Concept-based Information Retrieval Tools on the Web. *5th East-European Conference, ADBIS 2001* (Vol. 2, pp. 29–41). Vilnius, Lithuania.

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (n.d.). A unifying theoretical framework for the study and definition of semantic measures - application to bio-ontologies. *BMC Bioinformatics (submitted)*.

Hernandez, N. (2005). Ontologies de domaine pour la modélisation du contexte en recherche d'information. *Informatique*. Toulouse: Paul Sabatier, Toulouse.

Hirst, G., & St Onge, D. (1998). Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database and some of its applications (Language, Speech, and Communication)*. Cambrige, MA, USA: The MIT Press.

Hubert, G., Mothe, J., Ralalason, B., & Ramanonjisoa, B. (2009). Modèle d'indexation dynamique à base d'ontologies. *Actes de la sixième édition de la COnférence en Recherche d'Information et Applications, CORIA'09* (pp. 169–184). Toulon, France.

Iksal, S., & Garlatti, S. (2001). Documents virtuels et composition sémantique : Une architecture fondée sur les ontologies. *Nîmes TIC2001* (pp. 91–96.). Nîmes, France: Ecole des mines d'Alès.

Jalabert, F. (2007). Cartographie des connaissances : l'intégration et la visualisation au service de la biologie. Application à l'ingénierie des connaissances et à l'analyse de données d'expression de gènes. Montpellier: Montpellier II.

Jalabert, F., Ranwez, S., Crampes, M., & Derozier, V. (2006). I2DEE : intégrer et visualiser des données biologiques pour concevoir une ressource termino-ontologique. In M. Harzallah, J. Charlet, N. Aussenac-Gilles, & M. Lewkowicz (Eds.), *17èmes journées francophones d'Ingénierie des connaissances, IC2006* (pp. 141–150). Nantes, France.

Jalabert, F., Ranwez, S., Derozier, V., & Crampes, M. (2006). (IDEE)-D-2: An integrated and interactive data exploration environment used for ontology design. *Managing Knowledge in a World of Networks, Proceedings* (LNCS., Vol. 4248, pp. 256–271). Berlin/ Heidelberg: Springer.

Jansen, B J, Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, *36*(2), 207–227.

Jansen, Bernard J. (2000). The effect of query complexity on Web searching results. *Inf. Res.*, *6*(1).

Jimeno-Yepes, A., Berlanga-Llavori, R., & Rebholz-Schuhmann, D. (2010). Ontology refinement for improved information retrieval. *Information Processing & Management*, *46*(4), 426–435.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B. (1979). Citation Classic - Multidimensional-Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Current Contents/Social & Behavioral Sciences*, (39), S12–S12.

Lazaridis, M., Axenopoulos, A., Rafailidis, D., & Daras, P. (2012). Multimedia search and retrieval using multimodal annotation propagation and indexing techniques. *Signal Processing: Image Communication*. doi:10.1016/j.image.2012.04.001

Lee, W. N., Shah, N., Sundlass, K., & Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. *AMIA Annu Symp Proc*, 384–388.

Lefèvre, P. (2000). *La Recherche d'informations : du texte intégral au thésaurus* (p. 253). Paris: Hermès science.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Fifteenth International Conference on Machine Learning* (pp. 296–304). Morgan Kaufmann Publishers Inc.

Lucas, W., & Topi, H. (2004). Training for Web search: Will it get you in shape? *Journal of the American Society for Information Science and Technology*, *55*(13), 1183–1198.

Maedche, A. D., & Staab, S. (2001). Ontology learning for the Semantic Web. *Ieee Intelligent Systems & Their Applications*, *16*(2), 72–79.

Maniez, J. (2002). *Actualités des langages documentaires : fondements théoriques de la recherche d'information*. (É. de l'Association des professionnels de l'information et de la documentation (ADBS), Ed.) *Études et techniques* (Vol. Collection, p. 395). Paris.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (p. 496). Cambridge University Press.

Marchiori, M. (1998). The limits of Web metadata, and beyond. *Computer Networks and ISDN Systems*, *30*(1-7), 1–9.

Mbao, M. (2007). Distance Sémantique et Carte Conceptuelle. *Ecole des Mines d'Alès, Université Montpellier II*. Nîmes.

Menzies, T. (1999). Cost benefits of ontologies. *intelligence*, *10*(3), 26–32.

Minack, E., Demartini, G., & Nejdl, W. (2009). Current Approaches to Search Result Diversification. *Proceedings of The First International Workshop on Living Web at the 8th International Semantic Web Conference (ISWC)*. Westfields Conference Center, Washington DC., USA.

Mizoguchi, R., & Bourdeau, J. (2004). Le rôle de l'ingénierie ontologique dans le domaine des EIAH. *STICEF - Sciences et Technologies de l´Information et de la Communication pour l´Éducation et la Formation*, *11*.

Morrison, A., Ross, G., & Chalmers, M. (2003). Fast multidimensional scaling through sampling, springs and interpolation. *Information Visualization*, *2*(1), 68–77.

Munzner, T. (2009). A Nested Model for Visualization Design and Validation. *IEEE transactions on visualization and computer graphics*, *15*(6), 921–928.

Otlet, P., & La Fontaine, H. (1896). Création d'un Répertoire Bibliographique Universel: note préliminaire. *IIB Bulletin I* (pp. 15–38).

Pakhomov, S. V, Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., & Chute, C. G. (2010). Towards a framework for developing semantic relatedness reference standards. *J Biomed Inform*.

Pastorello Jr., G. Z., Daltio, J., & Medeiros, C. B. (2008). Multimedia Semantic Annotation Propagation. *2008 Tenth IEEE International Symposium on Multimedia* (pp. 509–514). IEEE.

Petrakis, E. G. M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. (2006). X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. *JDIM - Journal of Digital Information Management*, *4*(4), 233–237.

Platt, J. C. (2005). FastMap, MetricMap, and landmark MDS are all Nyström algorithms. *International Workshop on Artificial Intelligence and Statistics (AISTATS)* (pp. 261–268).

Prévot, L., Borgo, S., & Oltramari, A. (2010). Interfacing ontologies and lexical resources. In C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, & L. Prévot (Eds.), *Ontology and the Lexicon, a Natural Language Processing Perspective* (Studies in., pp. 185–200). Cambridge: Cambridge University Press.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(1), 17–30.

Ranwez, S., & Crampes, M. (1999a). Conceptual Documents and Hypertext Documents are two Different Forms of Virtual Document. In M. Milosavlejevic, F. Vitali, & C. Watters (Eds.), *Workshop on Virtual Documents Hypertext Functionality and the Web at the 8th International World Wide Web Conference*. Toronto, Canada.

Ranwez, S., & Crampes, M. (1999b). Méta-description en XML de documents vidéo. *ISKO'99*. Lyon, France.

Ranwez, S., Duthil, B., Sy, M. F., Montmain, J., Augereau, P., & Ranwez, V. (2013). How ontology based information retrieval systems may benefit from lexical text analysis. In A. Oltramari, L. Qin, P. Vossen, & E. Hovy (Eds.), *New Trends of Research in Ontologies and Lexical Resources* (Theory and., pp. 209–230). Springer.

Ranwez, S., Ranwez, V., Sy, M.-F., Montmain, J., & Crampes, M. (2010). Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information. In S. Desprès (Ed.), *21es Journées Francophones d'Ingénierie des Connaissances IC 2010* (pp. 247–258). Ecole des Mines d'Alès, Nîmes, France: Presse des Mines.

Ranwez, S., Ranwez, V., Villerd, J., & Crampes, M. (2006). Ontological distance measures for information visualisation on conceptual maps. *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Pt 2, Proceedings, 4278*, 1050–1061.

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., & Douzery, E. J. P. (2007). OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC evolutionary biology*, *7*, 241. doi:10.1186/1471-2148-7-241

Ranwez, V., Janaqi, S., & Ranwez, S. (2012). An O(n.m) algorithm for calculating the closure of lca-type operators. *Ars Combinatoria*, *104*.

Ranwez, V., Ranwez, S., & Janaqi, S. (2009). Extraction de sous-ontologies autonomes par fermeture des opérateurs hyponymie et hyperhonymie. In G. K. et P. T. Ladjel Bellatreche (Ed.), *JFO'09 : 3ème édition des Journées Francophones sur les Ontologies* (pp. 45–56). Poitiers, France: ACM.

Ranwez, V., Ranwez, S., & Janaqi, S. (2012). Sub-ontology extraction using hyponym and hypernym closure on is-a directed acyclic graphs. *IEEE Transactions on Knowledge and Data Engineering*, *24*(12), 2288–2300.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130.

Richy, H., & Després, S. (2007). Métadonnées, ontologies et documents numériques. *Techniques de l'ingénieur*, *H 7 155v2*, 1–17.

Rocchio, J. J. (1971). Relevance Feedback in Information Retrieval. (G. Salton, Ed.)*The SMART Retrieval System - Experiments in Automatic Document Processing*, 313 – 323.

Rondeau, I., Ranwez, S., & Crampes, M. (2005). Approche multidimensionnelle du texte pour le balisage des ressources pédagogiques. In R. Dieng-Kuntz, M. Grandbastien, & D. Herin (Eds.), *Journée thématique: Web sémantique pour le e-Learning - AFIA 2005*. Nice, France.

Sánchez, D., Solé-Ribalta, A., Batet, M., & Serratosa, F. (2012). Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. *Journal of biomedical informatics*, *45*(1), 141–55. doi:10.1016/j.jbi.2011.10.005

Seidenberg, J., & Rector, A. (2006). Web ontology segmentation: analysis, classification and use. *Proceedings of the 15th international conference on World Wide Web*. Edinburgh, Scotland: ACM.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martínez-Cruz, L. A., Corrales, F. J., et al. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, *2*(4), 330–8.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *1996 IEEE Symposium on Visual Languages* (pp. 336–343). IEEE Computer Society.

Sowa, J. (1984). *Conceptual Structures : information processing in mind and machine*. (A. Wesley, Ed.) *the system programming series* (p. 481).

Staab, S., & Studer, R. (2009). *Handbook on Ontologies - Second edition. International Handbooks on Information Systems* (p. 811). Springer.

Sy, M. F., Ranwez, S., Montmain, J., & Ranwez, V. (n.d.). Efficient conceptual relevance feedback using semantic neighborhood connectivity. *IEEE Transactions on Knowledge and Data Engineering*.

Sy, M. F., Ranwez, S., Montmain, J., & Ranwez, V. (2012). OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine. *COnférence en Recherche d'Information et Applications - CORIA 2012*. Bordeaux, France.

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2011). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, *13*(Suppl 1), S4. doi:10.1186/1471-2105-13-S1-S4

Tversky, A. (1977). Features of Similarity. *Psychological Review*, *84*(4), 327–352.

United Nations International Scientific Information System (UNISIST). (1975). *Principes d'indexation*. Paris: Unesco.

Vaudry, C., Ranwez, S. R., Poulon, A., & Crampes, M. (2002). Initiative mixte dans les DVP : de la Pertinence à l'Adaptation. In S. Garlatti & M. Crampes (Eds.), *Documents Virtuels Personnalisables, DVP 2002* (pp. pp. 141–154). ENST Bretagne, Brest, France.

Villerd, J. (2008). *Représentations visuelles adaptatives de connaissances associant projection multidimensionnelle (MDS) et analyse de concepts formels (FCA)*. Montpellier II.

Villerd, J., Ranwez, S., & Crampes, M. (2008). Navigation sur des Cartes de Connaissances supportées par un Treillis de Galois. *CARTO 2.0, "Où en êtes-vous de la mise en scène de vos informations ?"* (pp. 105–113). Champs Sur Marne, France.

Villerd, J., Ranwez, S., Crampes, M., & Carteret, D. (2007). Using Concept Lattices for Visual Navigation Assistance in Large Databases: Application to a Patent Database. In P. E. and M. L. Jean Diatta (Ed.), *CLA 2007, the fifth International Conference on Concept Lattices and Their Applications* (pp. 88–99). Montpellier, France.

Villerd, J., Ranwez, S., Crampes, M., & Carteret, D. (2009). Using concept lattices for visual navigation assistance in large databases. *International Journal of General Systems, 38*(4), 405–425.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Dublin, Ireland: Springer-Verlag New York, Inc.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England), 23*(10), 1274–81.

Wolff, C. (2010). *Les technologies de l'information et de la communication contribuent-elles à l'essor d'une société de l'information accessible au plus grand nombre ? Entre mythe et réalité, quelle appropriation de l'Internet par les 16-25 ans en France.* (p. 121). Nîmes.

Yager, R. R. (1979). Possibilistic decision making. *IEEE Trans. on Systems, Man and Cybernetics*, (9), 388–392.