HAL
open science

Bibliographies scientifiques : de la recherche d'informations à la production de documents normés

Gérald Kembellec

To cite this version:

Gérald Kembellec. Bibliographies scientifiques : de la recherche d'informations à la production de documents normés. Sciences de l'information et de la communication. Université Paris VIII Vincennes-Saint Denis, 2012. Français. NNT : . tel-00771553

HAL Id: tel-00771553

https://theses.hal.science/tel-00771553

Submitted on 8 Jan 2013

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 Paris 8 Vincennes - Saint-Denis

ÉCOLE DOCTORALE : Cognition, Langage et Interaction (CLI)

Thèse de Doctorat en Sciences de l'Information et de la Communication

Bibliographies scientifiques

De la recherche d'information à la production de documents normés

Gérald KEMBELLEC

Laboratoire Paragraphe

Thèse dirigée par Imad SALEH

Professeur des Universités

Date de soutenance le 3 décembre 2012

Jury :

Rapporteur	:	Mme Ghislaine CHARTRON, Professeur	Conservatoire national des arts et métiers
Rapporteur	:	Mr Stéphane CHAUDIRON, Professeur	Université Lille 3
Examinateur	:	Mr Mohamed HASSOUN, Professeur	ENSSIB
Examinateur	:	Mr Madjid IHADJADENE, Professeur	Université Paris 8, Saint-Denis
Examinateur	:	Mme Claire SCOPSI, MCF	Conservatoire national des arts et métiers
Directeur	:	Mr Imad SALEH, Professeur	Université Paris 8, Saint-Denis

Résumé

La bibliographie scientifique d'un chercheur rend compte de sa phase de documentation, de son positionnement et de ses choix argumentatifs. La recherche d'information scientifique constitue un des fondements du métier de chercheur, car ce dernier, en perpétuelle quête d'information, alimente sa réflexion tout en influant sur le processus de ladite information.

La production scientifique forme ainsi un cycle, dans lequel elle est conçue, tout en s'inspirant d'autres productions. Elle sera lue et commentée par des chercheurs, simultanément ou successivement, qui la citeront pour démontrer la pertinence de leur raisonnement. Cet écrit sera tour-à-tour une production finie, puis réactualisée en une source d'inspiration pour la recherche ultérieure.

Du fait que le patrimoine écrit est considérable et que le volume des données scientifiques accessibles en ligne augmente rapidement, la constitution d'un outil d'assistance à la recherche bibliographique s'impose. Du fait de la charge cognitive et temporelle de plus en plus importante en contexte d'investigation, de telles contraintes risquent d'amener à une perte de qualité pour la recherche d'information et de l'évaluation des corpus.

Dans l'écriture de chaque article, la formalisation d'une bibliographie est un exercice obligatoirement complexe, tant sur le fond que sur la forme, et chaque communauté scientifique a, en la matière, ses propres pratiques ainsi que ses codes typographiques. Ajouter a cela que les établissements scientifiques, enjeux de qualité obligent, se heurtent aux coûts forts et aux difficultés liées à la formation des professionnels de la recherche. Il peut s'agir des documentalistes, des chercheurs ou des étudiants en thèse en quête d'informations dans leurs centres ou sur les diverses plateformes qu'ils utilisent aux fins de réaliser des bibliographies normées.

La récolte des références, la veille scientifique, la recherche des documents primaires, la mise en forme des références, leur intégration dans les bibliographies ainsi que leur appel dans les documents de recherches sont autant de tâches aussi répétitives que décourageantes. Pour faciliter de telles tâches, la maîtrise des différentes étapes d'élaboration est nécessaire, comme par exemple le travail de coordination qui doit être d'autant facilité. Ou bien la maîtrise par les étudiants en thèse de la méthodologie d'utilisation des sources, qui doit être considérée comme acquise. Aussi la normalisation est-elle indispensable à l'écriture scientifique, ne serait-ce que pour assurer la cohésion des ouvrages collectifs, des revues ou des actes de conférences.

Quant au documentaliste qui est directement impacté dans la problématique de la gestion bibliographique, une facette importante de son métier, est un élément clé dans la chaîne de production de l'information documentaire. Il n'est en effet pas de produit documentaire (synthèse documentaire, revue de presse, etc.) qui ne contienne pas de bibliographie, souvent avec un style imposé par l'établissement de rattachement.

La recherche documentaire en pleine mutation est aujourd'hui informatiquement assistée, de la première étape d'établissement du périmètre de recherche jusqu'à la consultation des documents, sans oublier leur annotation. L'écriture même de la bibliographie est actuellement de moins en moins effectuée manuellement. Cette évolution des pratiques, qui va de pair avec celle des usagers de l'information (scientifique ou vulgarisation), induit un changement sensible dans les usages des professionnels, documentalistes ou éditeurs de logiciels. Les centres documentaires, notamment universitaires, ne sont pas en reste en informatisant systématiquement leurs catalogues et en les interconnectant.

Cette thèse a justement pour objectif de démontrer qu'il est possible de sérialiser le processus complet de recherche documentaire et d'offrir une méthode graphique associée. Nous postulons donc que la recherche documentaire peut être techniquement automatisée, de la première étape d'établissement du périmètre de recherche jusqu'à l'écriture de la bibliographie. Les étapes de sélection et de gestion documentaire peuvent aussi être facilitées par des outils et normes dédiés.

Après une introduction à la recherche d'information, nous examinerons la typologie des documents scientifiques, les modalités de stockage et de diffusion, ainsi que les normes et protocoles associés.

Nous décrivons plusieurs méthodes de recherche documentaire et différents outils d'interrogation des bases de connaissances.

Nous soumettons une étude centrée sur l'usager pour faire émerger des profils utilisateurs et les usages associés, puis nous soumettons une démarche conceptuelle et expérimentale d'accompagnement visuel à la recherche de documentation scientifique.

Nous synthétisons l'ensemble des procédures d'automatisation bibliographiques pour modéliser un outil de recherche alliant respect des normes, souplesse d'usage et considération des besoins cognitifs et documentaires de l'usager. Cette interface sert de support à une recherche naviguée dans les corpus documentaires avec l'intégration de services d'exposition de métadonnées.

Notre détaillons la modélisation d'un système de recherche documentaire, dédié à la recherche d'information en informatique scientifique, pour simplifier l'appropriation du domaine de connaissance par l'usager et automatiser au maximum les tâches subalternes. Nous proposons un panorama des différentes taxonomies représentatives du domaine informatique, dans l'optique d'un choix qualitatif pour une adaptation aux besoins correspondant à notre outil de recherche. Nous nous intéresserons ensuite aux méthodes existantes pour la visualisation optimale de ce type de classification de l'information. Enfin, nous modéliserons un système original de navigation et de recherche que nous implémenterons et évaluerons.

Pour conclure ce travail, nous reviendrons sur tous les aspects issus des différentes sciences qui nous ont permis d'envisager un modèle d'accès à l'information scientifique adapté aux populations cibles. Nous estimerons la réussite de ce projet et proposerons des perspectives pour nos futures recherches.

À ma femme Kim,
mes enfants Emma et Nathan.
À mes parents ...

Remerciements

Quand je me retourne sur ces années de recherches, de réflexion et de rédaction qui ont abouti à la présentation de ce manuscrit, je me rends compte que je n'aurai pas pu arriver seul à ce résultat. Je me dois de remercier mes collègues et amis qui m'ont montré le chemin ou alimenté dans ma réflexion.

En tout premier lieu, je dois remercier Kim, ma femme, pour son soutien indéfectible et sa patience depuis 2006, quand j'ai entamé ce travail de recherche. Merci à toute ma famille pour l'aide fournie au cours de ces années.

Merci à Quentin Kembellec, élève ingénieur à l'ENSEIRB-MATMECA de Bordeaux pour ses précieux éclaircissements en matière de mathématique.

Merci à Vincent BOYER, Maître de conférences en informatique à l'Université Paris 8, Saint-Denis pour les informations pratiques relatives à l'évaluation des bibliographies scientifiques et aux points d'entrées des bases de connaissances sur l'Internet. Ces informations me furent précieuses lors du début de ma thèse.

Merci à Catherine SAUVAGET, docteur en informatique à l'Université Paris 8, pour ses tests et critique sur mon outil de recherche. Sa réflexion et nos travaux communs de recherche ont permis d'orienter et d'alimenter ma propre réflexion pour affiner la plateforme de recherche présentée à la fin de cette thèse.

Merci à mes amies doctorantes et post-doctorante, Orélie DESFRICHES DORIA, Emma BESTER et Camille PALOQUE-BERGES qui partagent mon bureau au CNAM. Leur soutien moral, leurs avis et critiques furent précieux pour la rédaction de ce manuscrit.

Merci à l'ensemble de l'équipe de l'INTD et plus généralement du département Culture, Information, Technique et Société du CNAM, tant les équipes pédagogiques, administratives que documentaires.

Je désire remercier Jean-Baptiste YOUNES, Maître de conférences en informatique à l'Université Paris 7 Denis Diderot pour son aide précieuse à la rédaction de ce document avec LaTEXet la compréhension du format BibTEX.

Merci à Christophe ROCHE, Professeur à l'Université de Savoie et président de la société de terminologie française pour m'avoir aidé à saisir les subtilités de l'emploi du juste terme, en l'occurrence celui de taxonomie.

Je me dois également de mentionner le Bernhard RIEDER, Maître de conférences en sciences de la communication à l'Université d'Amsterdam. Son soutien durant mon MASTER et le début de mon doctorat me fut précieux. Ses enseignements techniques ont été très largement réinvestis dans l'outil que décrit la fin de cette thèse.

Je ne peux que terminer mes remerciements par mon directeur de thèse, le Professeur Imad SALEH pour son soutien indéfectible et ses conseils avisés depuis le MASTER jusqu'à la fin du Doctorat.

Table des matières

Table des figures **xi**

Liste des tableaux **xv**

Introduction **1**

Contexte de la recherche d'information scientifique 2

Objectifs 5

Positionnement 5

Problématique 8

La stylistique bibliographique 8

Citation et plagiat 10

Hypothèses 11

Hypothèse principale 12

Hypothèses secondaires 12

Méthodes 13

Plan de lecture 14

I Contexte et problématique de la bibliographie scientifique **17**

1 Recherche d'informations en contexte scientifique **19**

1.1 Contextualiser la RI pour les chercheurs 20

1.1.1 Petit point de méthodologie en RI scientifique 20

1.1.2 Le raisonnement scientifique 20

1.1.3 Le corpus scientifique 23

1.2 L'enjeu de qualité en RI scientifique 24

1.2.1 Bibliométrie 24
1.2.2 Scientométrie 24
1.2.3 Obsolescence et demi-vie d'un document 26
1.2.4 Indice d'immédiateté 29
1.2.5 Facteur d'impact 30
1.2.6 Le H-index 30
1.2.7 Pour conclure sur les indices qualité de l'information 34

2 Portails et entrepôts scientifiques 35
2.1 RI et exhaustivité, une utopie ? 36
2.1.1 Notion de visibilité et d'accès sur l'Internet 37
2.2 Les OPAC 39
2.2.1 Le Sudoc 39
2.3 Les moteurs de recherche scientifiques 40
2.3.1 Google Scholar 40
2.3.2 Microsoft Academic Search 41
2.4 Les éditeurs en recherche scientifique génériques 41
2.4.1 ScienceDirect 41
2.5 Les éditeurs de recherche scientifiques spécifiques 43
2.5.1 Le portail de l'Association for Computing Machinery (ACM) . . 43
2.5.2 Springer Verlag 44
2.5.3 IEEE 45
2.5.4 Emerald 45
2.6 Les bases de connaissance scientifiques 46
2.6.1 DBLP 46
2.6.2 Les bases documentaires à comité de sélection 46
2.7 Les archives scientifiques ouvertes 47
2.7.1 ArXiV et Hal 48
2.8 Les interfaces de recherche fédérées 49
2.8.1 Ebsco et Couperin 49
2.8.2 Isidore 50
2.8.3 Bielefeld Academic Search Engine (Base) 50
2.9 Portails de revues 51

2.9.1 Revues.org 51
2.9.2 Persée 51
2.9.3 Érudit 52
2.9.4 Cairn 53
2.10 Conclusion 53

3 La recherche d'information 55
3.1 Le paradoxe de la RI 58
3.2 Concepts, modèles et méthodes en RI 59
3.2.1 Portails de connaissance 60
3.2.2 Moteurs de recherche 61
3.2.3 Les méta-moteurs 79
3.3 Concepts avancés de recherche d'information 81
3.3.1 Interfaces de références virtuelles 81
3.3.2 Introduction aux systèmes de recommandation 86
3.3.3 Le concept de requête par l'exemple (QBE) 88
3.3.4 Services rendus par les systèmes de recommandation 91
3.3.5 Méthodologies des systèmes de recommandation 91
3.3.6 Conclusion 100

4 Qualité de l'information 101
4.1 RI et qualité de l'information 102
4.1.1 Bruit et taux de rappel 104
4.1.2 Silence et taux de précision 105
4.1.3 F-measure 106
4.2 Le PageRank 109
4.3 Conclusion 111

5 Les écoles de pensée en RI : Processus et Cognition 113
5.1 Le besoin d'information 114
5.2 Information et connaissance 117
5.2.1 Savoir ou savoir-faire 117
5.2.2 Connaissances générales et spécifiques 117
5.2.3 Connaissance et classification 118

5.3 Modèle de Guthrie 119
5.4 Le modèle « évaluation, sélection, traitement » (EST) 121
5.5 Le modèle « *Infomation Search Process* » 122
5.6 Le modèle « TIMS » de Dillon 125
5.7 Le modèle « *berrypicking* » de Bates et le phénomène de sérendipité . . 128
5.8 Le modèle « *information-seeking* » de Marchonini 129
5.9 Analyse comparative et critique des méthodologies 133
5.10 Conclusion 134

II Pratiques de recherche bibliographique 137

6 Bibliographies et métadonnées : normes, formats et styles. 139
6.1 Les métadonnées et notices bibliographiques 141
6.1.1 Nature et fonctions des notices 142
6.1.2 Nature et fonction des métadonnées 142
6.2 L'objet bibliographique 143
6.2.1 Définition de l'objet bibliographique 143
6.2.2 Nature et fonction de la bibliographie 143
6.2.3 Normes liées aux bibliographies 144
6.3 Les formats de notices bibliographiques et de classification 145
6.3.1 Les formats d'échange de données bibliographiques 145
6.3.2 Le Dublin CORE 145
6.4 Les formats d'échanges bibliographiques 149
6.4.1 Les formats de la bibliothèque du Congrès 149
6.4.2 BibTEX 152
6.4.3 bibWord, le format bibliographique de Microsoft Word 155
6.4.4 Le RIS 157
6.5 Les styles de bibliographies scientifiques 160
6.5.1 Exemple d'instruction bibliographique 161
6.5.2 Les principaux styles bibliographiques 162
6.5.3 Discussion conclusive sur les bibliographiques 163

7 Panorama des logiciels de gestion de bibliographie 165

7.1 Objectifs et protocole de test 167
7.2 Les logiciels « lourds » de gestion bibliographique 168
7.2.1 JabRef 168
7.2.2 EndNote 171
7.2.3 Mendeley 171
7.2.4 BibDesk 175
7.2.5 Zotero, le module du navigateur firefox 178
7.3 Les applications « en ligne » de gestion bibliographique 181
7.3.1 Zotero, le site internet 181
7.3.2 Refworks 182
7.3.3 Discussion 188
7.4 Conclusion 189

8 Étude d'usage de système d'information documentaire scientifique 191
8.1 Technologie et pratiques bibliographiques en milieu universitaire 192
8.2 Problématique 193
8.2.1 Constat 193
8.2.2 Postulat 193
8.2.3 Résultats attendus 193
8.2.4 Méthodologie d'enquête 194
8.2.5 Population cible et Panel 195
8.3 Les résultats 197
8.3.1 Contexte technologique 197
8.3.2 Usages d'outils de productions écrites en sciences 198
8.3.3 Les usages et formats de bibliographique 201
8.3.4 Les usages de logiciels de gestion bibliographique (LGRB) . . . 208
8.3.5 Critères pour le choix d'un LGRB 211
8.3.6 Recherche d'informations 213
8.3.7 Utilisation de sources pour les documentalistes et bibliothécaires 215
8.4 Analyse des résultats 217
8.4.1 Établissement de profils 218
8.5 Conclusion de l'étude 219

III Modélisation et développement d'un outil global de création de bibliographies en informatique scientifique 221

9 Urbanisation de systèmes d'informations 223

9.1 Historique 225
9.2 Valorisation des métadonnées hétérogènes 226
9.3 Méthodes orientées glanage 228
9.3.1 Dublin Core intégré dans les métadonnées HTML 228
9.3.2 Les méthodes basées sur le RDF « embarqué » 230
9.3.3 COinS 231
9.3.4 unAPI 232
9.3.5 Avantages et inconvénients du glanage 235
9.4 Méthodes orientées moissonnage 235
9.4.1 OAI-PMH 236
9.4.2 Z3950, SRU et SRW 240
9.4.3 RDF, SPARQL endPoints et triples stores 243
9.5 Conclusion 249

10 Panorama de taxonomies en informatique 251

10.1 Définition et usages de Taxonomie 253
10.1.1 Taxonomie ou taxinomie, historique terminologique 255
10.2 La Classification Décimale de Dewey 255
10.2.1 Historique de la CDD 256
10.3 Le système de classification ACM 257
10.3.1 Le principal document de l'ACM CCS 257
10.3.2 Les fichiers connexes à l'ACM CCS 258
10.4 IEEE ACM CCS étendu 260
10.5 Liste de sujets de l'ACM J. UCS 261
10.5.1 Fichier principal de la taxonomie J. UCS 262
10.5.2 La liste des termes supplémentaires du J.UCS 262
10.6 Le Système de classification HUJI 263
10.6.1 La classification des sujets HUJI 263
10.6.2 Liste alphabétique des mots-clés 263

10.7 La classification CORR Subject Areas d'ArXiV 264
10.8 Conclusion 266

11 OntologyNavigator 267
11.1 Proposition de définition du terme ontologie 269
11.2 État de l'art de la recherche d'informations par ontologies de domaine . 271
11.3 Ontologie de domaine informatique, conception d'un modèle exploitable 272
11.3.1 Notion de pertinence utilisateur par désambiguïsation et point de vue 273
11.4 Première ébauche 274
11.5 Visualisation d'un domaine de connaissance 281
11.5.1 La représentation 282
11.5.2 Visualiser pour chercher de l'information 283
11.5.3 Cartes cognitives 283
11.5.4 Cartes conceptuelles, ou concept maps 284
11.5.5 Synthèse des cartes cognitives 285
11.5.6 Pratiques comparées de visualisation de graphes 287
11.6 Deuxième approche 288
11.6.1 Objectifs 288
11.6.2 Le modèle 289
11.6.3 L'implémentation technique 291
11.6.4 Évaluation 293
11.6.5 Conclusion 293
11.7 Troisième approche 294
11.7.1 Méthodes de recherche proposées et présomptions de modèles exploitables 294
11.7.2 Traduction de l'ontologie en français 295
11.7.3 Mise en œuvre du modèle QBQ-S 298
11.7.4 Explicitation d'usage 300
11.7.5 Exemple de recherche contextuelle d'articles 301
11.7.6 Évaluation du modèle 304
11.7.7 Corpus ACM ré-indexé et enrichissement ontologique 310
11.8 Limites et perspectives de notre modèle 312

11.9 Conclusion 313
11.10Généralisation du modèle de recherche 313

12 Conclusion générale **315**

Conclusion **315**

Bibliographie **321**

Index **349**

IV Annexes **353**

A Publications relatives à la thèse **355**

B Chiffres ISI **357**

C Pratique bibliographique dans l'enseignement supérieur **359**

D Questionnaire usabilité de l'ACM CCS et d'OntologyNavigator **363**

E La classification informatique d'ACM **365**
E.1 Le contenu de la classificatoire 365
E.2 Nouvelle interface graphique d'ACM 405

F A propos de DBLP **407**

Table des figures

1.1 Répartition des citations par années pour l'article de Salton, Sources Google Scholar 26

1.2 Nombre de citations de l'article *Term-weighting approaches in automatic text retrieval* de G. Salton entre 2007 et 2011 27
1.3 Calcul de la demi vie d'article (exemple) 28
1.4 Représentation citations de G. Salton avec les données issues de CiteSeer 31
1.5 Représentation du H-Index de G. Salton avec les données issues de CiteSeer 32
1.6 Gadget de calcul du H-Index sur la base des statistiques de Google Scholar 32

2.1 Allégorie classique du web visible comme partie émergée de l'iceberg . . 37
2.2 Graphe de co-écriture navigable sur Microsoft Academic Search 42
2.3 Interface classique à l'URL `http://dl.acm.org/` 43
2.4 Recommandations basées sur les données 45

3.1 Le modèle « simpliste » Morville et Rosenfeld (2006) de système d'information 56
3.2 Modèle trivial de SRI proposé par Bates 57
3.3 Schéma fonctionnel d'indexation d'un moteur de recherche. 62
3.4 Les origines de la recherche à facettes informatisée 72
3.5 Schéma d'une recherche à facettes 73
3.6 Projet Flamenco 75
3.7 2 exemples de moteur à curseur : Social Search et FuzzFind 76
3.8 Parts de marché des moteurs de recherche commerciaux à travers le monde 78
3.9 Paramétrage d'un méta-moteur, ici Teardrop 80
3.10 Modèle original de ASKAL et son adaptation à la BU d'Angers 82
3.11 Capture d'une réponse sur « Rue des Facs ». 85
3.12 Concept de QBE sur le site commercial Amazon. 88
3.13 Exemple d'ontologie de la RS sur un site marchant Sieg *et al.* (2010) . 89
3.14 Exemple d'utilisation du système Clide. 91
3.15 Système de recommandation folksonomique : del.icio.us 100

4.1 Résultat de recherche d'informations documentaire 104
4.2 Illustration du PageRank 110

5.1 Diagramme « État-Transition » du modèle de Guthrie proposé par par Tricot 120
5.2 Interactions dans le modèle de TIMS, schéma original de Dillon 126

5.3 Berry Picking et sérendipité 129
5.4 « Information seeking », schéma original de Marchionini 130

6.1 Article décrit au format MODS, adaptation simplifée du marcXML . . 151
6.2 Bibliographie Word au format Office Open XML 156
6.3 Style Lavoisier 161
6.4 Principaux styles bibliographiques proposés par Mendeley 163

7.1 Création d'un fichier bibTEX à partir de JabRef 169
7.2 Export à partir de JabRef 170
7.3 Création d'un fichier bibTEX à partir de Mendeley Desktop 173
7.4 Import de notices bibliographiques avec bookmarklet vers Mendley via le navigateur web. 173
7.5 Annotation d'un document pdf par Mendley. 174
7.6 Détail d'édition d'un article dans Bibdesk 175
7.7 Politique de filtre de classement des entrée bibliographique de BibDesk, ici sur les mots clés 176
7.8 Détection automatique de notices bibliographiques BibTEX et COinS dans un site Web par BibDesk 177
7.9 Détection massive d'articles par Zotero 178
7.10 Détection d'un document auto décrit ou d'une notice bibliographique par Zotero 180
7.11 Zotero en ligne 181
7.12 Utilisation de Google Scholar avec RefWorks 184
7.13 Export RefWorks 188

8.1 Méthode d'écriture en SHS selon l'expérience 199
8.2 Méthode d'écriture en sciences dures selon l'expérience 200
8.3 Méthode d'intégration bibliographique dans un document pour l'ensemble de la population. 202
8.4 Formats de fichiers bibliographiques connus par l'ensemble de la population. 202
8.5 Formats de fichiers bibliographiques utilisés 205
8.6 Logiciels utilisés pour la gestion documentaire, parmi ceux qui en utilisent au moins un. 210

8.7 Évolution de la sensibilité aux logiciels libres ou propriétaires (en % des groupes sondés) chez les chercheurs. 212
8.8 Sources d'informations scientifiques pour la recherche dans l'enseignement supérieur. 213
8.9 Sources documentaires en sciences dures consultées par profil d'expérience. 215
8.10 Sources documentaires consultées en sciences humaines par profil d'expérience. 216
8.11 Sources de données en centre de documentation 216

9.1 Exemple de métadonnées HTML « embarquées ». 229
9.2 Exemple d'implémentation de COinS 231
9.3 Exemple d'implémentation d'unAPI 233
9.4 Utilisation d'unAPI par Zotero 234
9.5 Exemple de fiche au format XML dc-oai. 238
9.6 Exemple de réponse au format Z3950 SRW. 243
9.7 Un même triplet RDF présenté en littéral ou avec les URI (de gauche à droite) 244
9.8 Fichier RDF présenté en XML) 246
9.9 Fichier RDF présenté en graphe 247
9.10 Recherche sur le SPARQL *endPoin*t de l'entrepôt DBLP 248

10.1 Classification décimale de Dewey (extrait de la classe 000 dédiée à l'informatique) 256
10.2 Fichier principal de l'ACM CCS, en ASCII(1), HTML(2) et XML(3) . . 258
10.3 ACM CCS Discovered Terms 260
10.4 Portage en XML de la classification HUJI (extrait) 265
10.5 Taxonomie CORR (extrait) 265

11.1 Relations entre taxonomie, thésaurus et ontologie de domaine 272
11.2 Exemple de description d'un document avec la taxonomie ACM 275
11.3 Extrait du fichier XML des descripteurs implicites 278
11.4 Extrait de l'affichage de la première approche 279
11.5 Graphe de visualisation de la connaissance universelle de Leibniz (1666) 281

11.6 Visualisation en graphe de concepts, proposée par Roger *et al.* (2001) d'après Novak et Gowin (1984) 285
11.7 Lien transversal au sein de l'ACM CCS 289
11.8 Exemples tirés du thésaurus MotBis 290
11.9 Branche A. de la Taxonomie ACM au format GraphXML simplifié . . . 292
11.10 Focus + contexte dans notre interface 293
11.11 Modèle fonctionnel d'OntologyNavigator 296
11.12 Exemple de curation folksonomique 298
11.13 Exemple d'utilisation de JSON 300
11.14 Exemple d'utilisation de cette version d'OntologyNavigator 302
11.15 Processus de sélection et d'accès au document 303
11.16 Corpus bibliographique indexé et annoté (extrait) 311

B.1 Publication d'articles scientifiques dans des revues indexées par *Science Citation Index*(SCI) et *Social Sciences Citation Index* (SSCI) sur la période 1995 à 2008 358

E.1 Nouvelle version de navigation« tabulaire » de l'ACM CCS, *publiée en ligne le 22 septembre 2012* 405

Liste des tableaux

2.1 Quelques bases de références bibliographiques à comité de sélection (au 1er octobre 2012) 47
2.2 Hal et ArXiV au 1er octobre 2012 48

3.1 Exemple bilingue de racinisation 65
3.2 Résumé des fonctionnalités des deux principaux moteurs de recherche commerciaux en 2011 71
3.3 Bruit généré par les moteurs de recherche, Véronis 77

3.4 Indice de pertinence perçue par moteur de recherche dans l'étude de Véronis (2006) 78
3.5 Exemple d'échantillon de votes 92
3.6 Proximité des usagers basée sur la corrélation de Pearson 93

5.1 Processus de gestion cognitive du modèle de Guthrie 119
5.2 Processus de gestion cognitive et processus de base du modèle EST . . 122
5.3 Processus de base du modèle de ISP de Kuhlthau et sentiments associés 123
5.4 Processus de base du modèle TIMS 127
5.5 Processus de base du modèle de Marchionini 131

6.1 Les 15 attributs du Dublin Core 148
6.2 Tableau du Dublin Core qualifié 150
6.3 Typologie des documents décrits au Format BibTEX 154

8.1 Répartition des doctorants par expérience et domaine de recherche . . . 196
8.2 Contexte technologique des usagers, selon la répartition scientifique . . 197
8.3 Systèmes d'exploitation utilisés par profil utilisateurs en sciences dures. 198
8.4 Systèmes d'exploitation utilisés par profil utilisateurs en sciences humaines et sociales. 198
8.5 Outil de production de documents par profil d'utilisateurs. 199
8.6 OS en fonction de l'usage d'un compilateur de texte (LATEX,TEX). . . . 201
8.7 Formats bibliographiques connus en SHS 203
8.8 Formats bibliographiques connus en Sciences dures 204
8.9 Formats bibliographiques connus par les bibliothécaires et documentalistes.204
8.10 Formats bibliographiques utilisés en SHS 206
8.11 Formats bibliographiques utilisés en Sciences dures 206
8.12 Formats bibliographiques utilisés par les documentalistes et bibliothécaires en SCD. 207
8.13 Ventilation de l'usage des LGRB 209

9.1 Liste des verbes du protocole OAI-PHM. 239
9.2 Vocabulaire SRU/SRW (extrait) 242

10.1 Usage constaté des termes taxonomie et taxinomie sur Internet 255

10.2 Extrait de la liste des descripteurs implicites d'ACM 259
10.3 Comparaison quatre premiers niveaux de l'ACM & IEEE CCS 261
10.4 Extrait des mots clés HUJI 264

11.1 Mind Mapping vs Concept Mapping 286
11.2 Idiosyncrasie 286
11.3 Résultat de l'évaluation de l'outil (Kembellec *et al.*, 2009) 305
11.4 Choix *a priori* d'un modèle de représentation / accès 307
11.5 Choix *a posteriori* d'un modèle de présentation / accès. 307
11.6 Adéquation des résultats proposés avec le besoin d'information. 308

Introduction

If I have been able to see farther than others, it was because I stood on the shoulders of giants.

Sir Isaac Newton

... et accessoirement devise de Google Scholar

Contexte de la recherche d'information scientifique

La recherche de l'information scientifique constitue un des piliers du métier de chercheur (Guyot, 2011). Ce dernier est en perpétuelle quête d'information pour alimenter sa réflexion, ce qui influe sur son processus de production. À l'occasion de chaque événement scientifique auquel il, participe le chercheur note des idées et des références. Par certains aspects, les productions scientifiques forment une base de la recherche : au sein d'un cycle dans lequel chaque article, s'inspirant lui-même d'autres productions, sera lu et commenté par des chercheurs qui le citeront pour démontrer la pertinence de leur raisonnement, l'écrit sera tour à tour une production finie, puis réactualisée et une source d'inspiration.

La recherche documentaire est aujourd'hui informatiquement assistée, de la première étape d'établissement du périmètre de recherche jusqu'à la consultation des documents, sans oublier leur annotation. L'écriture même de la bibliographie est actuellement de moins en moins effectuée manuellement.

Les centres documentaires, notamment universitaires, informatisant leurs catalogues et les inter-connectant vont dans le sens de ce constat. Notons l'effort français de l'Agence Bibliographique pour l'Enseignement Supérieur (ABES), relayé par les services documentaires des universités, pour numériser et cataloguer les thèses dans le projet titanesque Thèses En Ligne (TEL) et le dispositif STAR. Malheureusement, cette mise à disposition, de quelque 26 488 thèses au 1er janvier 2012 n'induit pas une consultation systématique par les principaux intéressés, à savoir étudiants, enseignants et chercheurs. Si l'on en croit Markey (2007) les catalogues des bibliothèques sont tombés en disgrâce. Une étude de De Rosa (2006) menée en 2006 pour le compte de l'*Online Computer Library Center*(OCLC) [1] démontre que rares sont les étudiants, enseignants et chercheurs commençant leurs recherches par le catalogue de leur bibliothèque. Ils sont 89 % à lui préférer un moteur de recherche commercial, au premier rang desquels Google. Selon l'étude de 2005 de Swan et Brown (2005), 72 % des universitaires anglo-saxons utilisent le moteur Google.

1. http://www.oclc.org/

Lorcan Dempsey, directeur de la recherche à OCLC, qualifie notre époque d'ère « Amazoogle » (De Rosa, 2006), faisant allusion au moteur de recherche Google et à la librairie numérique Amazon. Selon lui, il faut aller à la rencontre des usagers dans leurs espaces virtuels de prédilection, c'est-à-dire sur les moteurs de recherche et sur Amazon. Selon le sociologue français Vourc'h (2010), le pourcentage d'étudiants prétendant se rendre à la bibliothèque universitaire au moins une fois par semaine a baissé de 54 % en 1997 à 49,9 % en 2006. Une enquête de Jouguelet et Vayssade (2010), commandée par le ministère de l'enseignement et de la recherche français et livrée en 2010 compare les comportements des étudiants face au numérique en France et à l'étranger. Un point commun découvert entre les étudiants français et ceux des pays étrangers visés par l'enquête était que les étudiants deviennent désormais des « internautes chevronnés, des natifs du numérique (*digital natives*), et fréquentent assidûment les réseaux sociaux de l'Internet ». Toujours selon cette étude, l'usage des assistants personnels, tablettes et autres périphériques nomades connectés a modifié la recherche d'informations et de documents. Dans le cadre de cette évolution des pratiques, « Wikipédia devient source de référence bibliographique et beaucoup d'étudiants croient que toute l'information se trouve gratuitement sur Internet ».

Ce constat d'évolution des pratiques des usagers de l'information (aussi bien scientifique que de vulgarisation) induit un changement dans les usages des professionnels de la documentation, qu'ils soient documentalistes ou éditeurs de logiciels. Selon de Kaenel et Iriarte (2007), « les dernières évolutions du web, avec l'entrée en jeu de XML, des nouveaux usages et nouveaux outils, ainsi que le déplacement du centre de gravité, qui s'est fortement rapproché des utilisateurs, ouvrent de nouvelles voies et de nouveaux champs d'application pour les catalogues en ligne ».

Le Coadic fait remarquer que les techniques d'étude utilisées en bibliothéconomie sont généralement quantitatives et posent des questions suivant un même modèle général : « Quel système ? Quelle personne ? Quels services ? » et du type « Où ?, Quand ?, Qui ? (Le Coadic, 2008) ». Actuellement, sous l'effet du nouveau paradigme orienté usager des sciences de l'information et de la communication (SIC), on s'intéresse de plus en plus à la façon dont un usager perçoit ses besoins. Les sujets abordés par les acteurs du domaine de recherche s'orientent alors davantage vers des questions du type : « Comment ? » : « Comment définissez-vous vos besoins ? Comment les présentez-vous

au système ? Comment utilisez-vous le système ? Comment vous aide-t-il (ou ne vous aide-t-il pas) ? » et « Pourquoi ? ».

Le propos de cette thèse s'inscrit directement dans cette démarche. Par une approche quintillienne (Simonsson et Johnson, 2006) du sujet, "*quis ? quem ? quid ? ubi ? quando ? cur ? quibus in consiliis ? quibus auxiliis ?*[1]", nous allons examiner soigneusement les nouvelles normes, méthodologies et outils qui permettent de simplifier la tâche des chercheurs dans leurs activités de recherche et de recherche et de gestion bibliographique. Nous tenterons de créer un lien entre les bases de connaissance scientifique de tous horizons et l'utilisateur final pour réconcilier l'usager avec les catalogues et les bases de connaissance.

1. « Qui, quoi, où, par quels moyens, pourquoi, comment, quand ? »

Objectifs

L'objet final de ce travail de thèse est de réaliser un outil intuitif pour la recherche d'articles scientifiques en adéquation avec les nouvelles pratiques de documentation scientifique. Comme notre périmètre de connaissance initial est centré sur l'informatique, nous avons circonscrit l'objet de nos expérimentations sur la recherche d'information (RI) aux champs scientifiques liés à l'informatique. Nous avons conscience que le choix de ce champ d'étude est structuré avec moins de nuances que les sciences humaines et sociales, ce qui peut avoir un impact sur les expérimentations. Nous proposons de produire un outil qui permettra, outre la recherche de l'information scientifique validée, d'appréhender le domaine de connaissance sous un angle facilitant la compréhension, et mettant en exergue les interconnexions entre les sous-domaines ; ces éléments relèvent d'un questionnement épistémologique. Cet outil se doit pour être utilisable par un plus grand nombre d'étudiants, en plus des chercheurs, d'être accessible en plusieurs langues, le choix revenant à l'utilisateur.

Positionnement

Cette thèse s'inscrit dans le champ des Sciences de l'Information et de la Communication (SIC) comme cadre principal. Les travaux sur l'usage des médias informatisés étudient les interactions entre la technologie et les pratiques des usagers. Notre vision du dispositif socio-technique étudié se fonde sur une approche compréhensive qui prend en compte les faits humains et sociaux. Nous allons ainsi tenir compte d'éléments théoriques issus de la sociologie, sur des populations très ciblées, liées principalement à l'enseignement supérieur et à la recherche. Les sciences psycho-cognitives apporteront une aide précieuse pour appréhender la complexité des phénomènes humains en jeu dans la recherche documentaire numérique.

Notre posture épistémologique est étroitement liée au domaine d'étude des sciences de l'information. Nous nous intéresserons plus particulièrement à la construction des systèmes d'information documentaire pour la recherche scientifique et à leurs usages. La diffusion des savoirs par des vecteurs technologiques récents retiendra notre attention, tout autant que les pratiques liées aux usages de ces dispositifs techniques.

La science de l'information, pour Le Coadic, est essentielle aux savoir-faire associés aux métiers liés à la documentation comme valorisation du cadre d'activités de nature

technique. Les notions propres à la science de l'information ne doivent cependant pas être définies de manières trop étroitement ciblées, par rapport aux pratiques d'une profession (Le Coadic, 1994).

Meyriat distingue deux approches informationnelles pour les sciences documentaires : une partie pratique relative aux technologies documentaires qu'il désignait sous le terme de « documentologie » et la théorie relative aux concepts informationnels liés au document qu'il proposait de dénommer « informatologie » qui proposait des discussions globales, comme la manière de donner du sens à une information, ou encore la problématique des conditions de réception de l'information (Meyriat, 1981).

Nous adhérons à l'approche documentaire « *vu, lu, su* » de Salaün, et plus globalement du collectif Pédauque qui propose trois parties distinctes (Pédauque, 2006, Salaün, 2012). En effet, nous pensons également que le processus de rédaction bibliographique requiert plusieurs habiletés centrées autour du document : le repérage dans la collection, le décodage du document pour l'extraction de l'information et enfin sa transformation en savoir. Ce dernier point étant l'apport social du document comme vecteur de connaissance.

Ainsi, les sciences de l'information étudient l'application et l'usage des connaissances dans les organisations, et l'interaction entre ces dernières, les individus et les systèmes d'information.

Pour participer à l'offre documentaire, l'idée d'une interface de recherche scientifique fédérée nous paraît une piste intéressante. Nous avons envisagé l'approche aristotélicienne (catégorisation de la connaissance) pour simplifier l'accès au savoir. Cette pratique, utilisée et éprouvée par les bibliothécaires du vingtième siècle, et les biologistes avant eux, peut cependant être valorisée par un artefact ergonomique qui soutiendrait et faciliterait l'acte psycho-cognitif complexe de recherche documentaire. Ainsi, diversifier l'accès aux entrepôts documentaires préexistants par l'hybridation électronique de l'ontologie d'un domaine scientifique avec un outil de recherche d'information documentaire nous a semblé être une expérience intéressante à mener.

Pour éclairer notre point de vue sur la problématique, il faut donner quelques éclaircissements sur notre parcours. Dans le monde de la recherche, nul ne peut se déclarer objectif face à son sujet de prédilection. Chercheur engagé professionnellement dans le champ de l'information-documentation, nous y sommes impliqué en tant que pédagogue. Enseignant en Information Scientifique et Technique (IST) à l'Institut

National des Techniques de la Documentation (INTD) du Conservatoire National des Arts et Métiers (CNAM), nous assumons un service d'enseignement au sein de la filière de la documentation auprès des élèves ingénieurs de l'EICNAM, des licences professionnelles de documentation audiovisuelle et d'entreprise, mais aussi du titre RNCP [1] de chef projet en ingénierie documentaire. Cependant, cette vocation n'est apparue que tardivement après avoir exercé le métier d'ingénieur en Technologie de l'Information et de la Communication pour l'Enseignement (TICE) dans les établissements universitaires d'Île-de-France. La pratique des flux informationnels nous ancre dans une dimension numérique très liée à l'informatique appliquée. Ces outils à notre disposition nous offrent une meilleure compréhension des solutions liées à l'information documentation qui se pratique actuellement. Cet héritage professionnel et culturel multiple est un atout pour notre recherche. Nous ferons nôtre la pensée de Brigitte Guyot sur la difficulté du positionnement en information-documentation : il s'agit d'« une discipline aux frontières mouvantes, tantôt attirée vers des préoccupations gestionnaires, tantôt vers des préoccupations techniques ou méthodologiques. Sciences hybrides à coup sûr, signalant l'état transitoire qui est celui du champ des pratiques et celui de la recherche (Guyot, 2004) ».

Dans sa proposition de posture épistémologique des sciences de l'information, Fondin n'interpellait-il pas dès 2001 sur sa dualité (Fondin, 2001) ? Comprendre le propos empêche-t-il de vouloir améliorer l'objet étudié ? Les deux aspects lui paraissaient indissociables. Fondin, soulignait que si le paradigme positiviste tend à approcher les sciences de l'information des sciences exactes (Fondin, 2001, p.121), il s'attache également à saisir la nature et les difficultés de la recherche d'information. Nous envisagerons donc les sciences de l'information comme étant structurellement attachées aux sciences humaines et sociales.

Les réalités contemporaines socio-économiques et technologiques produisent des mutations, des combinaisons des paradigmes de diverses disciplines au sein du vaste champ des SIC et plus particulièrement de l'information et de la documentation. En conséquence, nous espérons que le lecteur pardonnera l'éventail des sujets techniques abordés dans cette thèse dans le but concevoir un vecteur original d'accès à la documentation scientifique.

1. Répertoire National des Certifications Professionnelles.

Problématiques liées à la réalisation de bibliographies scientifiques

Le volume de données scientifiques accessibles en ligne augmente rapidement. La recherche bibliographique impose donc une charge cognitive et temporelle de plus en plus importante. Ces nouvelles contraintes risquent d'amener à une perte de qualité de la diffusion scientifique et de l'évaluation par les citations. Dans l'écriture de chaque article, la formalisation d'une bibliographie est un exercice obligatoire complexe tant sur le fond qu'en ce qui concerne la forme. Chaque communauté scientifique a ses propres pratiques en ce domaine et ses codes typographiques associés. Dans le cadre des établissements scientifiques, les enjeux de qualités se heurtent au coût et aux difficultés liés à la formation des documentalistes, chercheurs et étudiants de recherche d'information dans les centres documentaires et sur les diverses plateformes et à la réalisation de bibliographies normées. Si nous revenons sur ces points, la récolte des références, la veille scientifique, la recherche des documents primaires, la mise en forme des références, leur intégration dans les bibliographie ainsi que leur appel dans les documents de recherche sont autant de tâches répétitives et décourageantes.

La stylistique bibliographique

Dans son article polémique sur l'angoisse de la citation, Schick témoignait sur une réunion de crise organisée par son université pour essayer de trouver une solution à ce problème (Schick, 2011). Les bibliothécaires et membres du corps enseignant se sont concertés pour répondre à l'angoisse personnels des bibliothèques qui ne peuvent plus effectuer leur mission principale : aider les étudiants dans le champ la culture informationnelle (trouver, choisir et utiliser les sources). En effet, les élèves les harcèlent avec des questions sur la façon de formater les citations relatives de tous ordres dans les différents styles bibliographiques. Il semble même, selon Schick [1] que l'obsession de la citation bibliographique ait éclipsé le style et la thématique dans les préoccupations des pointilleux professeurs d'humanités de son université. L'attention des professeurs à la grammaire de la citation « parfaite » n'est pas sans conséquences sur le travail des étudiants. Ces derniers passent une quantité disproportionnée du temps et de l'attention

1. Kurt Schick enseigne la littérature à l'Université James Madison, en Virginie.

dédiés à leur travail à éviter de faire des erreurs de forme. La qualité d'écriture réside alors pour eux dans le suivi mécanique de règles plutôt que dans le développement d'idées originales. Schick poursuit son investigation en se demandant s'il faut juger l'écrit sur le contenu et son caractère plutôt que sur des caractéristiques de présentation. Cette question de rhétorique est qualifiée par l'auteur de « colossal gâchis ». Selon lui, le style de la bibliographie et des citations demeure l'élément le plus arbitraire, stéréotypé et normatif de l'écriture académique enseignée dans les lycées et les universités d'Amérique. Actuellement, le style bibliographique, qualifié de « jargon académique sacralisé », persiste malgré un ratio coût/bénéfice incroyablement élevé, à être enseigné aux étudiants dès le lycée. Les professeurs tatillons nuisent au développement de jeunes rédacteurs en les forçant à investir plus de temps et de réflexion dans des éléments moins importants que la rédaction. L'importance du formatage des citations est une préoccupation relativement nouvelle dans l'éducation. Les chercheurs expérimentés qui ont appris à effectuer leur recherche documentaire avant l'arrivée massive de Google comme source informationnelle prennent le temps d'une recherche complète et ne s'inquiètent du style qu'après l'écriture de leur production. De l'avis de Schick, pour les nouvelles générations, avec la qualité inégale des informations disponibles en ligne, « il est plus important lors d'un processus d'écriture de savoir comment évaluer la valeur de leurs sources que d'analyser les règles pédantes du formatage des sources et des notes de bas de page (Schick, 2011) ».

Le chercheur confirmé, qui souhaite répondre à un appel à contribution avec des contraintes temporelles fortes traitera probablement la stylistique bibliographie seulement après avoir alloué des efforts substantiels aux activités d'écriture plus complexes et rationnelles. Parmi ces activités citons l'affinage des sujets, la sélection, le choix et le traitement des sources, l'organisation des idées, la rédaction et la révision des manuscrits afin d'en améliorer l'orientation et la cohérence. Cependant, la présentation d'une bibliographie dans un format peu familier est toujours un exercice qui, sans être réellement complexe, va exiger du temps et de la concentration.

Citation et plagiat

Cette obsession n'est pas uniquement due à l'amour de la belle typographie : le problème annexe à la bibliographie scientifique (et donc à la citation) est l'absence de mention de l'auteur, qui assimile la citation à du plagiat. Théoriquement le plagiat est un problème d'appropriation de la pensée d'autrui, donc une faute morale. Ce cas de vol de propriété intellectuelle arrive, mais, le plus souvent, il s'agit d'un défaut de référence, voire d'une erreur de typographie comme l'oubli de marque de citation comme les guillemets.

Cette problématique est réelle, car elle lèse la personne non citée, surtout dans un contexte bibliométrique au sein duquel des indicateurs statistiques comme le H-index font et défont la renommée d'un chercheur. Ainsi, depuis quelques années, aux États-unis une chasse aux sorcières s'est organisée autour du phénomène de plagiat dans l'optique d'assainir les méthodes de citation. Cette hyper-vigilance méthodologique n'est pas sans conséquences sur les usages des populations de rédacteurs scientifiques, notamment les strates les plus jeunes, comme les doctorants.

Des recherches récentes menées dans le cadre du « Projet Citation [1] » corroborent le fait que l'obsession des enseignants anglo-saxons pour la citation a diminué les aptitudes numériques des étudiants relativement à la recherche et la gestion d'informations, ce que les Anglo-saxons appellent l'*information literacy*. Rebecca Moore et Sandra Howard Jamieson, les directrices du projet, blâment « l'hystérie du plagiat », qui oblige les enseignants à punir les mauvaises citations plutôt que de récompenser l'utilisation efficace des sources (Jamieson et Howard, 2011).

Notre questionnement

La problématique n'est évidement pas de se positionner contre la rigueur stylistique dans la citation. En effet l'étudiant doit apprendre à formater ses références et le jeune chercheur ne peut tout simplement pas publier sans une bibliographie normalisée. La normalisation est indispensable à l'écriture scientifique, ne serait-ce que pour assurer la cohésion des ouvrages collectifs, revues et actes de conférences. Le travail

1. *The citation project*, un projet inter-universitaire américain étudiant les différentes problématiques de la référence bibliographique en contexte d'enseignement et de recherche `site.citationproject.net`, accédé le 1er septembre 2012

de coordination, souvent assuré par un collègue chercheur, s'en trouve grandement facilité. Apprendre aux étudiants la méthodologie d'utilisation des sources les initie à un processus utile, même en contexte professionnel. Ajoutons que le documentaliste est également directement impacté par cette problématique de gestion bibliographique qui est une facette importante de son métier. Il n'est en effet pas de produit documentaire qui ne contienne pas de bibliographie, souvent avec un style imposé par l'établissement de rattachement.

Plutôt que de se focaliser sur la stylistique de la citation, n'est-il pas plus utile de se poser la question de l'automatisation de la chaîne de gestion bibliographique ? Nous ne pouvons en effet pas contrôler le ratio de temps et d'efforts que les étudiants et les jeunes chercheurs investissent dans un travail d'écriture. En revanche, nous pouvons influer sur la façon dont ils distribuent leurs énergies en proposant d'une part des formations à l'usage des logiciels de gestion de bibliographie (pour les enseignants-chercheurs) des interfaces de recherche compatibles avec lesdits logiciels (pour les documentalistes, bibliothécaires et personnels informatiques des services de documentation universitaire).

Ce qui nous amène à émettre quelques hypothèses pour satisfaire aux exigences de tous les acteurs de la chaîne liant le monde de l'enseignement et la recherche sans léser les auteurs.

Hypothèses

Nous allons exposer le contexte d'usage des bibliographies scientifiques, que cela soit dans le cadre d'un cursus d'enseignement ou de recherche, par des étudiants, des chercheurs ou des documentalistes. Nous allons également montrer la lourdeur de l'écriture scientifique, ce que ne vient pas alléger la complexité des modèles bibliographiques. L'enjeu connexe aux bibliographies et à la citation de références est le risque d'accusation de plagiat par défaut de mention à l'auteur.

Hypothèse principale

Dans le cadre d'un processus d'écriture scientifique, l'auteur doit se renseigner sur son sujet, ce qui l'engage dans un processus documentaire complexe.

Hypothèse principale

> Il est d'ores et déjà possible d'automatiser le processus complet de documentation scientifique par l'urbanisation des différents systèmes d'information et d'un outil d'inter-médiation dans un domaine scientifique spécifique.

Dans cette hypothèse, nous évoquons un processus complet de documentation. Nous allons établir un périmètre pour ce processus complet. Nous précisons que nous entendons par processus « complet de documentation » l'ensemble des méthodes, procédures et techniques permettant de rechercher, sélectionner l'information dans un ou plusieurs fonds documentaires spécifiques à un champ disciplinaire scientifique scientifique. Ce processus se poursuit par l'intégration des références bibliographiques dans un outil de gestion documentaire, l'intégration de références aux documents dans le traitement de texte, pour se terminer par la génération de la bibliographie dans le format souhaité. Toutes ces actions doivent donc être effectuées par des procédures simples de sélection sans que l'usager ait à recopier et formater les documents qu'il a sélectionnés.

Hypothèses secondaires

Nous pensons également que les étudiants, jeunes chercheurs, enseignants-chercheurs et les documentalistes sont prêts pour ce type d'outil. Il existerait donc une niche écologique favorable au modèle envisagé.

Première hypothèse secondaire

> Les populations cibles de la recherche documentaire scientifique – étudiants, enseignants-chercheurs et documentalistes – sont prêtes pour l'usage d'un tel processus.

Deuxième hypothèse secondaire

> L'outil que nous présentons peut se décliner sous forme d'un modèle abstrait afin d'être implémenté dans d'autres domaines scientifiques.

Troisième hypothèse secondaire

L'outil que nous présentons peut grâce, à un modèle visuel adapté, faciliter l'accès à des documents pertinents, mais aussi participer au processus cognitif de visualisation du domaine de connaissance.

Méthodes

Répondre à l'hypothèse principale

Dans un premier temps, pour répondre à la question de l'automatisation des tâches techniques de recherche documentaire, nous présenterons en détail les différents aspects de la recherche documentaire pour en exposer les parties technologiquement possibles à automatiser. Par la suite, nous observerons les protocoles documentaires d'inter médiation technique. Nous nous placerons enfin dans une démarche expérimentale d'automatisation des processus documentaires.

Répondre aux hypothèses secondaires

1. : Questionnaire et enquête préalables sur les pratiques des jeunes chercheurs et bibliothécaires.
2. : Modélisation comparée et évaluée des méthodes de recherche techno/sémantiques et des méthodes par glanage/moissonnage.
3. : Évaluation de l'usage de l'outil par enquête auprès des usagers témoins.
4. : Implémentation et évaluation d'un modèle visuel de recherche informationnelle.

Objectifs préliminaires

Nous concevons comme indispensable de commencer un travail de recherche à partir d'un domaine de connaissance après avoir fixé ses limites et analysé sa structure. Le premier objectif de ce travail est de théoriser le contexte et les principes de base de la recherche d'information, depuis les modèles pragmatiques simples jusqu'aux principes avancés d'aide à la recherche. Nous examinerons ensuite les modèles éprouvés de

recherche d'information sous l'éclairage des sciences cognitives pour saisir la charge informationnelle à laquelle est soumis un utilisateur de système de recherche d'information en condition d'usage. Nous avons synthétisé la littérature actuelle autour des projets similaires et recensé l'ensemble des sources d'information scientifiques dans le domaine de l'informatique.

Première partie

Nous commencerons par approcher le concept de recherche d'information. Notre approche sera issue des recherches en sciences cognitives appliquées à la bibliothéconomie. Nous aborderons cette problématique depuis le besoin d'information, notion que nous expliciterons, jusqu'aux diverses méthodes cognitives pour concevoir la recherche d'information. Ensuite, nous examinerons de manière pragmatique les aspects de la recherche d'information électronique. Nous observerons ces approches globalement, puis nous restreindrons notre investigation au cadre de la recherche scientifique. Cette partie sera l'occasion de s'intéresser d'un œil critique aux sources fiables d'informations scientifiques et techniques.

Deuxième partie

Dans notre deuxième partie, nous nous intéresserons aux bonnes pratiques bibliographiques en contexte scientifique. Pour ce faire, nous listerons les principaux formats d'échange de données et de métadonnées dans le cadre documentaire, ainsi que les styles de bibliographies scientifiques permettant de les représenter dans des écrits scientifiques. Nous décrirons ensuite les outils qui aident le chercheur dans son processus de documentation et d'organisation bibliographiques. Une étude exploratoire des pratiques de recherche d'information scientifique et technique, ainsi que de l'usage des outils associés, nous fournira ensuite de précieuses informations sur la culture informationnelle dans les milieux de l'enseignement et de la recherche en France. Fort des résultats de l'étude, nous déduirons l'importance de rendre les systèmes d'information interopérables, non seulement entre eux, mais aussi avec les outils de gestion documentaire choisis par les usagers. Enfin, nous proposerons un examen des méthodes actuelles d'urbanisation

entre systèmes d'informations (concept que nous expliciterons) pour établir quelles sont les plus adaptées au web documentaire actuel.

Troisième partie

Notre troisième partie sera consacrée à la modélisation d'un système documentaire dédié aux sciences informatiques, pour simplifier l'accès à l'information et automatiser au maximum les tâches subalternes. Nous commencerons par effectuer une étude comparative des différentes taxonomies représentatives du domaine informatique. Cette étude nous permettra de choisir une taxonomie, dans l'optique de se l'approprier et de l'adapter aux besoins correspondant à notre outil de recherche. Nous nous intéresserons ensuite aux méthodes existantes pour la visualisation optimale de ce type de classification de l'information. Enfin, nous modéliserons un système original de navigation et de recherche que nous implémenterons et évaluerons.

Conclusion

Pour conclure ce travail, nous reviendrons sur tous les aspects issus des différentes sciences qui nous ont permis d'envisager un modèle d'accès à l'information scientifique adapté aux populations cibles. Nous estimerons la réussite de ce projet et proposerons des perspectives pour nos futures recherches.

Première partie

Contexte et problématique de la bibliographie scientifique

Chapitre 1

Recherche d'informations en contexte scientifique

Tout ce que je sais, c'est que je ne sais rien..

Socrate

Introduction

Les enseignants chercheurs n'écrivent pas leurs articles scientifiques comme de la littérature classique. Il en va de même de la méthode et des points d'accès à la recherche documentaire qui sont également impactés par la spécialisation scientifique. La méthodologie d'acquisition de savoirs et de construction du discours scientifique repose sur des processus particuliers que nous allons expliciter brièvement. Il est possible de trouver de la documentation de ce type grâce aux filières classiques telles les moteurs commerciaux ou les annuaires sociaux. Cependant, le corpus scientifique est le plus souvent accessible au travers de portails spécialisés comme les bibliothèques universitaires

et les catalogues associés. La chaîne classique de l'édition imprimée est en train de muter et tend à disparaitre dans plusieurs disciplines dont les plus représentatives sont l'informatique, la médecine, la physique ou les mathématiques. Cette mutation est le fait de modèles économiques émergeant de la technologie internet, le plus souvent à l'initiative des communautés enseignantes elles mêmes. Nous allons, dans ce chapitre étudier la composition de la littérature scientifique et les facteurs de choix qui orientent les chercheurs dans leur constitution de bibliographie.

1.1 Contextualiser la RI pour les chercheurs

1.1.1 Petit point de méthodologie en RI scientifique

Avant de pouvoir contribuer scientifiquement, il est évident que le chercheur doit s'imprégner des travaux existants. Le travail de documentation est dense, puisqu'il faut chercher, sélectionner lire et assimiler les principaux écrits historiques et récents avant de pouvoir commencer à étayer une réflexion éclairée. Le chercheur ne part évidement pas de rien quand il aborde un sujet émergeant ou nouveau pour lui (Denecker *et al.*, 2000, p. 13). Il existe toujours un socle de connaissances générales interdisciplinaires. De plus Les connaissances plus spécifiques, étroitement liées à des contenus disciplinaires ou à des champs de connaissance particuliers viennent former une seconde couche de connaissance plus pointue, ce que nous voyons comme un « terreau » de connaissances du champs disciplinaire. Nous retrouvons en tout dernier lieu les écrits scientifiques émergents qui sont les « jeunes pousses » qui font la tendance scientifique.

1.1.2 Le raisonnement scientifique

L'information est une interprétation de la réalité à laquelle on accorde une certaine valeur. Cette valeur est variable selon les besoins et l'étendue des connaissances (Denecker *et al.*, 2000, p. 39). Pour s'approprier et ré-exploiter l'information, l'individu doit passer par une phase d'analyse que nous allons expliciter.

Denecker remarquait que dans un cadre de résolution de son besoin de connaissance, « l'individu exécute un certain nombre de calculs complexes et d'opérations logiques comme l'association, l'exclusion ou la sélection, avec des raisonnements simples comme l'analogie, la catégorisation ou l'inférence (Denecker *et al.*, 2000, p. 39) ».

Morin présente sa vision de ces opérations logiques sous la forme d'une opposition entre séparation et association avec une subdivision entre les processus et les réflexions associées (Morin, 1986, pp. 115–125).

Nous synthétiserons le processus mental que le chercheur va effectuer lors de sa recherche d'information de manière linéaire dans un plan d'expérience :

1. Inférer pour poser des hypothèses. Inférer consiste à admettre un nouvel énoncé à partir d'autres propositions déjà tenues pour vraies (Denecker *et al.*, 2000, pp. 39-40).
2. Réunir de l'information par un processus de recherche documentaire orienté par son vécu et borné par les hypothèses.
3. Traiter l'information recueillie avec les opérateurs logiques de cognition (Morin, 1986, p. 167)
4. Vérifier l'hypothèse. L'hypothèse est une « supposition, inspirée directement par ses connaissances antérieures (Denecker *et al.*, 2000, p. 40) ».

Rentrons dans le détail de l'étape de vérification d'hypothèses au sein du plan d'expérience. Durant l'exécution de son plan d'expérience, le chercheur va comparer (par analogie) ses hypothèses avec les éléments d'informations qu'il aura cherchés et analysés. Il va également chercher à classer ou catégoriser l'information dans un schéma taxonomique de la science étudiée.

Comparer : l'analogie

Définition : analogie (Denecker *et al.*, 2000, p. 41)

> L'analogie est une ressemblance, établie par l'imagination, entre deux ou plusieurs objets de pensée essentiellement différents (...). Le raisonnement par analogie conclut à une ressemblance générale à partir d'une ressemblance partielle.

Le Larousse en ligne définit, de manière plus simple, l'analogie comme un « point commun à des choses et qui crée leur ressemblance[1] ».

Dans le cadre scientifique, nous utiliserons la définition de Denecker. L'utilisation de l'analogie, de son point de vue n'est pas une simple mise en exergue de points communs

1. `http://www.larousse.fr/dictionnaires/francais/analogie`, accédé en ligne le 1er octobre 2012.

entre concepts, c'est une manière de s'extraire d'une situation cognitive complexe : « L'individu emploie un raisonnement analogique lorsqu'il ne peut appliquer simplement une règle ou interpréter des données. Il convoque alors des connaissances s'appliquant habituellement à des situations voisines (Denecker *et al.*, 2000, p. 41) ».

L'analogie favorise le transfert des connaissances (Tardif, 1998). En effet, à partir de la découverte d'une similitude, même partielle, entre objets elle autorise l'application à un nouvel objet les règles disponibles sur le premier (Weil-Barais *et al.*, 2011, p. 502).

La limite de l'analogie est liée à la compétence de l'usager du système d'information. Il doit être capable de juger de la pertinence de son jugement relativement aux objets mis en relation. Le réel degré de similarité dans la relation peut accélérer la cognition ou bien entrainer l'individu sur une fausse piste.

Classer : la catégorisation

Définition de catégorisation (Richard et Bonnet, 1990) :

> La catégorisation consiste à découper la réalité en classes d'objets similaires.

La catégorisation se divise en deux catégories (Denecker *et al.*, 2000, p. 43) :

1. Le niveau de base (la catégorie, approche par spécification/généralisation).
2. La typicalité (la représentativité de l'objet par rapport à son niveau de base).

Illustrons la typicalité par l'exemple d'un roman de poche aura une typicalité beaucoup plus prononcée dans le niveau de base de la catégorisation des livres qu'une bande dessinée ou un dictionnaire. Dans le cadre scientifique, un article de revue aura une typicalité supérieure à une monographie ou un brevet.

Nous estimons que l'évaluation de la typicalité est évidemment différente d'un individu à l'autre, mais c'est le sens commun qui prévaut.

Les opérations cognitives induites par le besoin d'information forment en quelque sorte les fondations du processus psycho-cognitif de la construction du savoir de l'individu. La structuration de ce socle de connaissances et compétences est un parcours individuel qui ne saurait être généralisé. Seule la méthodologie du « plan d'expérience » peut être généralisée, sans avoir de résultats identiques d'un individu à l'autre.

1.1.3 Le corpus scientifique

Pour recenser les documents dits scientifiques, il faut commencer par définir précisément ce que l'on entend par ce terme. Selon Pétroff (1984), un document scientifique et technique est un objet, mais aussi un acte scientifique dont le but est de faire passer le lecteur d'un état de connaissance dans un domaine donné à un autre état de connaissance. Il faut entendre dans ce contexte « état de connaissance » comme un réseau (mental) de relations entre éléments d'information. Toujours selon Pétroff, le discours du document scientifique est monosémique (à l'échelle nationale, voire internationale) du fait d'une préexistence codifiée des termes du langage. Chaque science possède les codes qui lui sont propres et qui peuvent être notés de manière littérale ou symbolique. Ainsi, un document scientifique en sciences dures pourra faire indifféremment référence à « micro » ou son équivalent μ pour désigner le millionième d'une unité.

Les accès courants à la documentation scientifique

La documentation scientifique est principalement composée d'articles, eux mêmes regroupés en journaux, actes de conférences et revues. Une autre source de documentation scientifique est l'ouvrage de recherche. Les productions écrites clés de validation des connaissances du chercheur apprenant, notamment les thèses et les mémoires de MASTER forment également une ressource non négligeable.

Types de documents scientifiques

Les documents rédigés en littérature scientifique peuvent être classés selon les catégories suivantes :

- Les livres publiés (*book*[1]) ou non publiés (*booklet*), chapitres (*inbook*) ou parties (*incollection*) de livres.
- Les articles (*article*) de journaux scientifiques.
- Les articles parus dans les actes d'une conférence scientifique (*inproceedings*).
- Les mémoires de Master recherche ou DEA (*masterthesis*), de doctorat ou d'habilitation (*phdthesis*)..
- Les documentations techniques scientifiques publiées (*techreport*) ou les manuels (*manual*).

1. Les traductions en italique sont celles qui apparaissent dans BibTEX

- Les inclassables comme les pages web personnelles de chercheurs ou de laboratoires, la littérature dite « grise », comme certains supports de cours sont classés dans « divers » (*misc*)
- Les documents en attente de publication (*unpublished*) d'auteurs reconnus par leurs pairs.

1.2 L'enjeu de qualité en RI scientifique

Dans le contexte très spécialisé de la communication scientifique, il est d'usage de tenter de mesurer la qualité des écrits, laquelle est sensément liée à celle de leurs auteurs. Si l'on définit littéralement le terme de bibliométrie par son étymologie, on obtient « mesure des livres ». Le Coadic, spécialiste du domaine, précise que si le terme de bibliométrie continue à être utilisé, celui d'infométrie gagnerait à l'être, car plus juste (Le Coadic, 2010).

1.2.1 Bibliométrie

Une première définition de la bibliométrie de Pritchard positionne cette discipline comme « l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication (Pritchard, 1969) ». Cette définition plutôt générique a été complétée par Price : « les recherches quantitatives de toutes les choses concernant la science et auxquelles on peut attacher des nombres » (de Solla Price, 1969), cette deuxième définition élargit le concept à la science en général et plus seulement aux communications. Cette définition pose donc les bases de la scientométrie.

1.2.2 Scientométrie

Pour ce qui est de l'infométrie, ce n'est que la mesure de l'activité scientifique et technique à travers une quantification des informations bibliométriques et du paratexte . La bibliométrie est fortement décriée[1] par une partie des enseignants chercheurs français. Polanco (1995) déclarait : « J'appelle réductionnisme bibliométrique le point

1. `http://www.sauvonsluniversite.com/spip.php?article1093`, accédé en ligne le 1er octobre 2012.

de vue par effet duquel l'article scientifique devient un outil de définition de la science et l'on fait de la publication écrite un indicateur privilégié de l'activité scientifique »

Ce débat sur le fait de réduire le travail d'un chercheur à ses écrits et aux indicateurs statistiques associés ne sera pas abordé. Cependant, ce que l'on appelle scientométrie, infométrie ou bibliométrie, est le fait d'étudier les statistiques liées à la publication. Ces informations apportent un certain nombre d'informations sur un document :

- Le nombre de citations depuis d'autres documents scientifiques ;
- Péremption ou demi-vie d'un article.

Eugène Garfield [1] a proposé l'idée d'utiliser les citations, renvoi d'un auteur vers un autre pour créer un recueil des articles les plus cités et cartographier la connaissance (Garfield et Small, 1985) (voir. Annexe B). Cette méthode est sujette à caution du fait des citations croisées entre auteurs. Ces citations sont dites « de complaisance », car elles ne sont qu'un échange de bons procédés dans l'optique de gonfler artificiellement le nombre de citations.

Comme il semble que le nombre de citations soit un indice de qualité d'un article, nous allons proposer de définir ce terme. Si nous reprenons la définition mathématique de Le Coadic (2005) du terme de citation, il faut distinguer la notion de citation de celle de référence :

« En science de l'information, quand un document (A) fait mention, se réfère à un document (B), on dit que le document (B) a été cité par le document (A). Dans ce cas, la référence et la citation sont interchangeables.

Une citation est un extrait, l'emprunt d'une phrase, d'un passage dans la production d'un vecteur de communication scientifique. Ce texte emprunté à un autre auteur et que l'on reproduit textuellement *in extenso* pour illustrer, appuyer ce que l'on veut dire est une citation. Une référence est une évocation : la simple indication d'un document auquel on se réfère pour exprimer une opinion sur un travail préexistant »

D'un point de vue pratique, dans un écrit scientifique, tout extrait d'un document tiers sans faire expressément mention des sources est considéré comme du plagiat. Pour ce qui est de faire allusion à un travail relaté dans un écrit, il n'est pas obligatoire de faire une mention des sources. Telle est la différence qu'il y a entre référence et

1. Eugène Garfield à créé l'*Institute for scientific information* (ISI) en 1966. Cette une société est maintenant connue sous le nom *Thomson Scientific* et propose un service (payant) de statistiques sur la documentation scientifique.

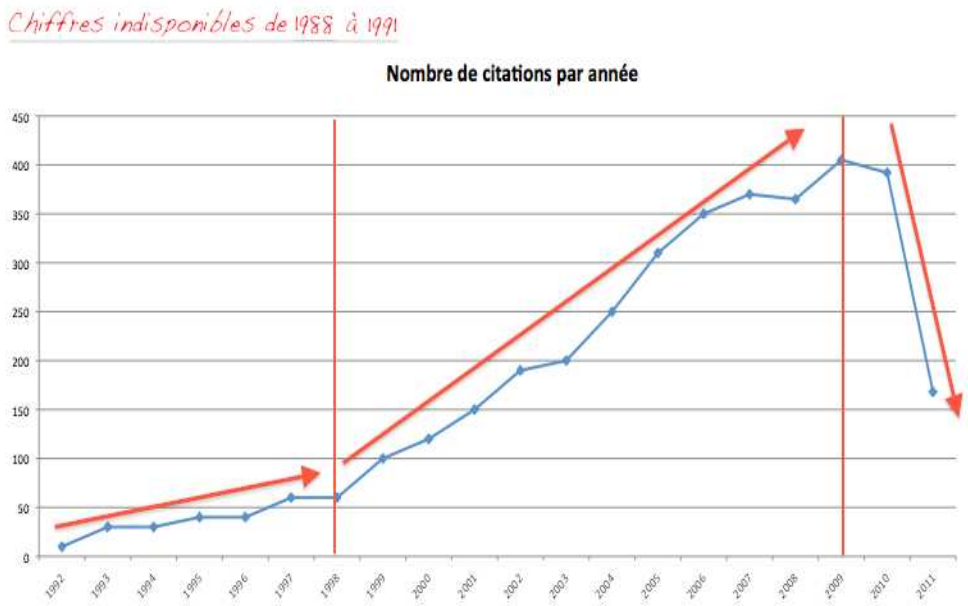


Figure 1.1: Répartition des citations par années pour l'article de Salton, Sources Google Scholar

citation. Dans l'usage courant, il est commun de considérer les deux termes (bien que différents) pour ainsi dire synonymes. Nous userons parfois de l'un ou de l'autre par abus de langage, considérant que les idées d'un chercheur sont aussi sacrées que les termes qu'il utilise pour les décrire dans une communication.

1.2.3 Obsolescence et demi-vie d'un document

La demi-vie d'une citation, dans le cadre d'une revue, d'une conférence ou d'un journal rend compte de la longévité des articles, et donc de la persistance de sa notoriété à long terme. En théorie, un document a une durée de vie, une actualité pendant laquelle il a une certaine valeur en tant que citation. La demi-vie d'un document peut être définie comme le temps au bout duquel les citations qui y sont relatives tombent en deçà du seuil de 50 % de leur nombre total (Pinhas et Kordon, 1997). La demi-vie des citations est variable selon les sciences. Ainsi elle sera en moyenne de 5 à 6 ans en sciences dites dures, et plus en sciences humaines :

– 4,6 ans pour la physique ;

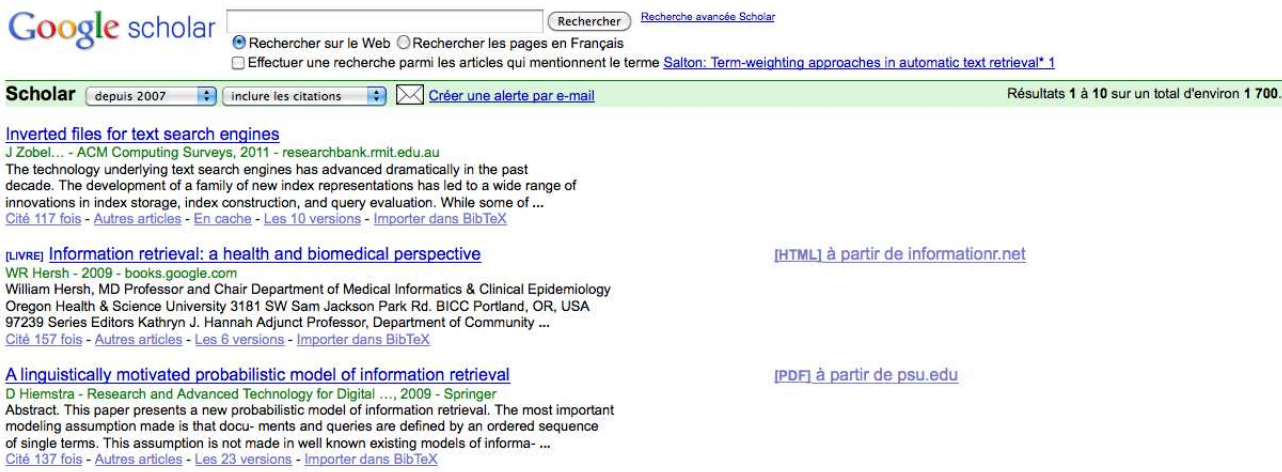


Figure 1.2: Nombre de citations de l'article *Term-weighting approaches in automatic text retrieval* de G. Salton entre 2007 et 2011

- 7,2 ans en philosophie ;
- 10 années en mathématiques ;
- 6 ans en biologie et en médecine.

La demi-vie des articles dépend de leur nature : en fonction du temps, le taux de citation suit une courbe en cloche (qui passe par un maximum entre 2 et 3 ans). La demi-vie peut être impactée par le fait que les articles « atypiques » ne commencent à être cités que bien après leur publication (nous reviendrons sur ce phénomène cf. 1.2.4). Pour modérer ce facteur, il est à rappeler que les articles « phares » ont une durée de vie nettement plus longue (Le Coadic, 2002) que la moyenne. Ceux servant de base à une théorie qui n'a pas été réfutée continuent d'être cités même s'ils datent de plusieurs décennies.

Par exemple, Salton et Shannon ont toujours une place prépondérante dans les bibliographies et sont encore régulièrement cités. L'exemple suivant nous permettra d'illustrer le concept de demi vie, mais aussi le fait que certains articles ont une demi vie plus longue que d'autres. Ainsi L'article *Term-weighting approaches in automatic text retrieval* de G. Salton, parue en 1988 apparait au 16 septembre 2011 comme ayant été cité 4 457 fois sur Google Scholar. Si l'on s'interroge sur les dates de ces citations, il apparaît que cet article a été cité 1700 fois ces 4 dernières années dont 560 fois durant la

Année	Année	Nombre de citations par année	Total cumulé	Demi-vie
1	1992	10	10	
2	1993	30	40	
3	1994	30	70	
4	1995	40	110	
5	1996	40	150	
6	1997	60	210	
7	1998	60	270	
8	1999	100	370	
9	2000	120	490	
10	2001	150	640	
11	2002	190	830	
12	2003	200	1030	
13	2004	250	1280	
14	2005	310	1590	
15	2006	350	1940	1820
16	2007	370	2310	
17	2008	365	2675	
18	2009	405	3080	
19	2010	392	3472	
20	2011	168	3640	1820
	Total	3640	3640/2	

Figure 1.3: Calcul de la demi vie d'article (exemple)

dernière année [1]. Si l'on observe l'historique des citations pour cet article (voir Fig. 1.1), l'article reste (relativement) peu cité entre 1988 et 1998 avec moins de 100 citations par années. Puis entre 1999 et 2009, la courbe devient ascendante avec un pic en 2009. Une tendance à la baisse se profile depuis 2010.

Calculons la demi vie de cet article :

Dans la capture de tableur 1.3, le total Tc de citations par année $c_(\lambda)_{n_k}$ de l'article λ est de 3640 sur une période de n années, ici 20 ans (k étant la variable itérative de l'année 1 à n. La formule générique pour calculer l'indice de demi vie DV est la suivante :

$$iDV_(\lambda) = \frac{1}{2}\sum_{k=1}^{n} c_(\lambda)_{n_k} \tag{1.1}$$

À partir de l'indice de demi-vie, on peut déduire la demi vie par simple lecture. Il faut encadrer l'indice de demi vie entre les années $n_(k-1)$ et $n_(k+1)$ dont les totaux

1. Il est à noter que CiteSeerX, qui se base sur les information extraites du très sérieux DBLP ne comptabilise qu'un total de 1079 citations pour cet article, dont 9 auto-citations. Cette différence s'explique par le haut niveau d'exigence des journaux, revues et conférences indexés dans DBLP.

cumulés de citation pour l'article λ sont juste supérieurs et juste inférieurs.
Pour expliciter l'exemple :

$$iDV_(\lambda) = \frac{1}{2}\sum_{k=1}^{20} c(\lambda)_{n_k} = \frac{3640}{2} = 1820 \tag{1.2}$$

Si l'on se reporte à la figure 1.3 il est simple de constater que les années d'encadrement sont 2005 et 2006, car $1590 < 1820 < 1940$. Ainsi, pour cet article la demi vie est de presque 15 ans (ce qui est beaucoup).

Mais, pour conclure sur la demi-vie d'un article et la notion d'obsolescence, les articles ne sont pas des denrées alimentaires périssables avec la mention « à citer de préférence dans les 7 ans ». Cependant, il est constaté que, de manière générale, un article n'est beaucoup cité que pendant une durée finie.

1.2.4 Indice d'immédiateté

Baudoin *et al.* (2004) décrivent l'indice d'immédiateté comme une mesure la rapidité avec laquelle les articles issus d'un journal ou d'une revue sont cités. L'indice d'immédiateté ii d'un journal λ pour une année n se calcule en divisant le nombre de citations Tc sur le nombre d'articles Ta publiés pour cette même année n.

$$ii_(\lambda)_n = \frac{Tc_(\lambda)n}{Tc_(\lambda)n} \tag{1.3}$$

Il est probable que pour des auteurs connus et prolifiques, la règle s'applique également. Pour reprendre l'exemple proposé par Baudoin *et al.* (2004), l'indice d'immédiateté du magazine *Nature* en 2002 représente le quotient du nombre de citations en 2002 et du nombre d'articles publiés en 2002 :

$$ii_(nature)_{2002} = \frac{6671}{889} = 7,504 \tag{1.4}$$

Pour la grande majorité des revues scientifiques, la valeur de cet indice se situe entre 0 et 1. Cet indicateur permet de repérer les journaux (et possiblement les auteurs) qui produisent les publications à fort impact dans un domaine de la communauté scientifique.

1.2.5 Facteur d'impact

Le facteur d'impact (IF) d'une revue scientifique mesure la fréquence avec laquelle, pendant une année donnée, l'article « moyen » d'une revue est cité dans les articles d'autres revues (Le Coadic, 2002). Le Facteur d'impact se définit comme étant le rapport entre le nombre de citations U reçues par une revue pendant l'année t et le nombre d'articles A publiés pendant les deux années précédentes. Il est donc évident que la publication dans certaines revues, ou certains actes de conférence sera un indice de qualité, si l'on se réfère à cet indice.

$$I_F(t) = \frac{U_{t-1}(t) + U_{t-2}(t)}{A(t-1) + A(t-2)} \tag{1.5}$$

1.2.6 Le H-index

La notoriété d'un auteur peut se mesurer d'après la quantité d'articles publiés ou encore par le nombre de citations qu'un auteur reçoit au cours de sa carrière. Cette notion de notoriété en matière de citations est à rapprocher de l'algorithme Google de « Pagerank » de Google. Cette manière d'établir la « valeur » d'un auteur à partir de tels indices est sujette à caution. La course à la publication, illustrée par la célèbre expression « publier ou mourir » (*publish or perish*) peut pousser certains acteurs scientifiques à multiplier des articles répétitifs. De plus, le réseau de connaissance peut biaiser les résultats liés au nombre de citations avec les citations de complaisance.

Le physicien Jorge E. Hirsch a élaboré en 2005 un algorithme statistique pour établir l'index (la valeur) d'un scientifique à partir de données bibliométriques (Hirsch, 2005). Le H-Index est un indicateur statistique destiné à mesurer la valeur scientifique d'un chercheur autrement que par des indicateurs tels que par une simple volumétrie des publications, le nombre moyen de citations et la somme de toutes les citations. Cet indicateur combine deux types de variables :

- Le nombre d'articles publiés et référencés dans les bases de données électroniques disponibles.
- Le nombre de fois où lesdits articles ont été cités (par d'autres).

Ainsi, un auteur prolifique n'étant jamais cité aura un H-Index à 0. Un scientifique n'ayant publié que peu d'articles, mais tous extrêmement cités, verra sont H-Index grimper.

Titre	année	articles	citations
Term-weighting approaches in automatic text retrieval	1998	1	1071
Improving retrieval performance by relevance feedback	1990	2	498
Approaches to Passage Retrieval in Full Text Information Systems	1993	3	146
The Effect of Adding Relevance Information in a Relevance Feedback Environment	1994	4	95
On the Use of Spreading Activation Methods in Automatic Information Retrieval	1988	5	64
Automatic Routing and Ad-hoc Retrieval Using SMART : TREC 2	1994	6	35
Automatic Text Decomposition and Structuring	1994	7	27
Parallel Text Search Methods	1998	8	22
Selective Text Utilization and Text Traversal	1995	9	21
Automatic Text Structuring and Retrieval - Experiments in Automatic Encyclopedia Searching	1991	10	16
Length Normalization in Degraded Text Collections	1995	11	11
Syntactic Approaches To Automatic Book Indexing	1988	12	11
On the Use of Term Associations in Automatic Information Retrieval	1986	13	7
Automatic Text Browsing Using Vector Space Mode	1995	14	6
Automatic Content Analysis in Information Retrieval	1968	15	4
An evaluation of term dependence models in information retrieva	1982	16	3

Figure 1.4: Représentation citations de G. Salton avec les données issues de CiteSeer

De l'avis de Bornmann et Daniel (2005) dans la revue *Scientometrics*, « Dans l'ensemble, les résultats suggèrent que le H-index est une approximation prometteuse de la qualité du travail scientifique ».

Pour établir manuellement un exemple, nous allons rester sur le même auteur G. Salton et noter ses statistiques bibliographiques depuis le site de références CiteSeer [1]. Dans le tableau1.4, nous n'avons pas pris en compte les auto-citations. Le H-Index est un système de plus en plus reconnu dans les milieux universitaires, notamment anglo-saxons.

Google Scholar, profitant des statistiques de la large base documentaire offre un outil de calcul du H-Index comme le montre la figure 1.6. Comment expliquer la différence entre le H-Index proposé par Google Scholar et celui calculé à partir des données de CiteSeer ? Sans offrir une réponse directe, nous savons que :

- Google Scholar est un outil grand public gratuit, avec des approximations qui rendent sa crédibilité contestable (Jacsó, 2010) ;
- CiteSeer est un site élitiste qui ne recense que les citations relatives à des articles publiés dans des structures à fort impact scientifique.

1. http ://citeseerx.ist.psu.edu/search ?q=%22G.+Salton%22&t=auth&sort=cite&start=10

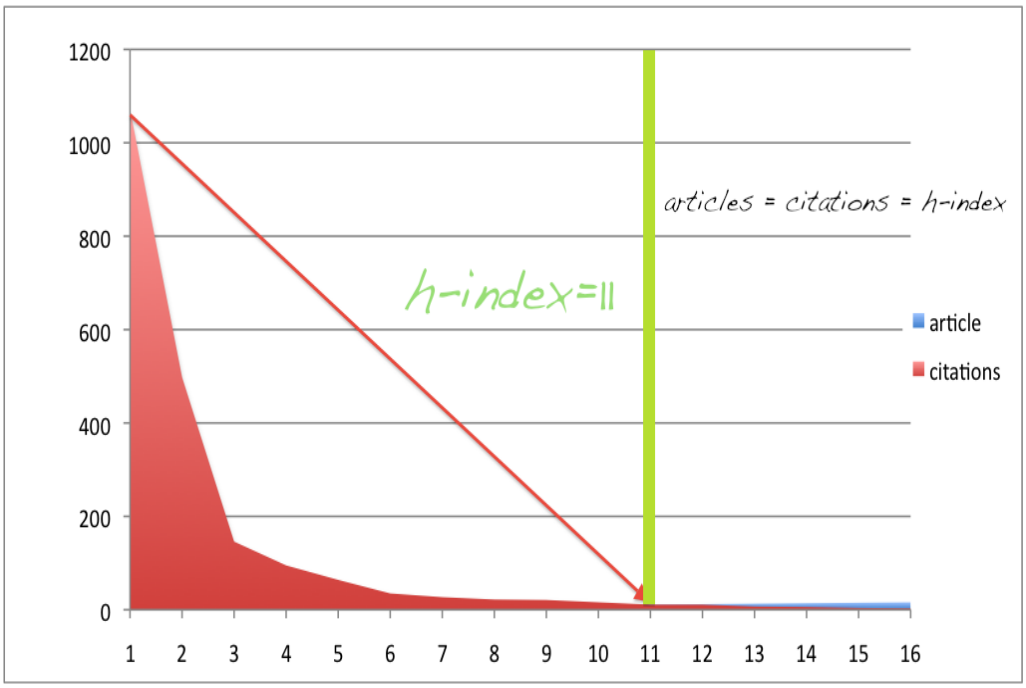


Figure 1.5: Représentation du H-Index de G. Salton avec les données issues de CiteSeer

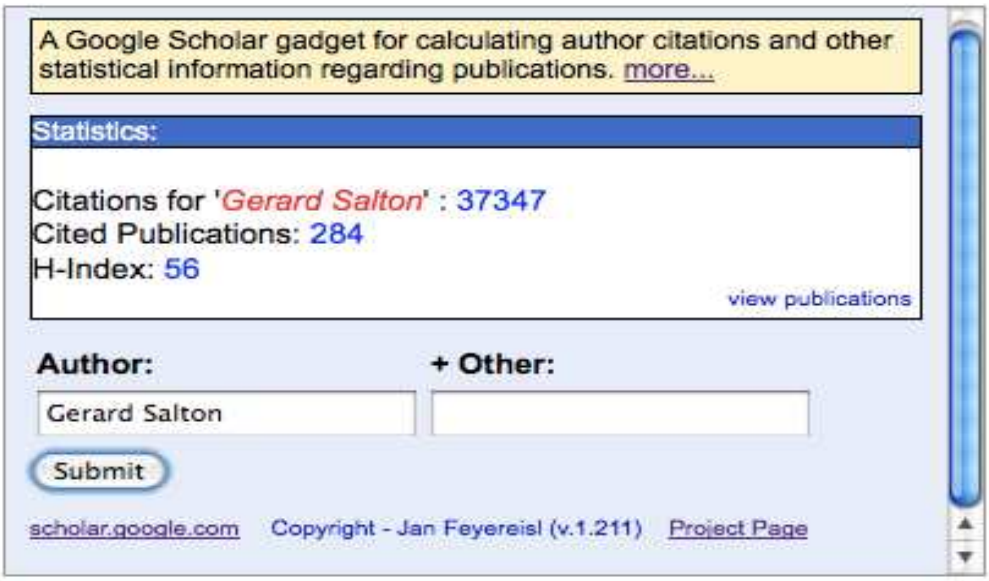


Figure 1.6: Gadget de calcul du H-Index sur la base des statistiques de Google Scholar

Le moteur d'indexation de Google Scholar indexe (théoriquement) les articles issus de périodiques avec comité de lecture, des thèses, des livres et des rapports scientifiques. En pratique, Google Scholar indexe tout ce qui a l'aspect d'un article scientifique. Rappelons l'expérience d' « Ike Antkare » qui est devenu un des scientifiques les plus cités du monde moderne, sans avoir d'existence réelle. Le chercheur français Cyril Labbe (2010) a créé en 2010 cet auteur et des articles générés aléatoirement qui lui étaient attribués. Il a ensuite écrit d'autres (faux) articles le citant copieusement. Google Scholar a plus tard indexé l'ensemble des documents et a attribué à « Ike Antkare » un H-index de 84. Cette valeur représente un indice bibliométrique supérieur à celui d'Albert Einstein.

De même, des chercheurs espagnols ont fait une expérience similaire, utilisant de faux articles afin d'alerter la communauté de la recherche sur la façon dont il est possible de facilement manipuler les données et les indicateurs bibliométriques avec Google Scholar. Ils ont créé six documents rédigés par un auteur fictif et les ont transférés vers le site personnel d'un chercheur de l'Université de Grenade. Ils ont également proposé un réseau de citations croisées entre articles entre différents articles. L'expérience a entraîné 774 citations pour cet auteur, ce qui a eu un impact significatif sur son indice de citation, mais aussi pour les revues qui étaient censées publier les articles (Delgado-López-Cózar *et al.*, 2012).

Dans ces conditions, comment calculer un H-Index de manière fiable ? Sur quelles statistiques faut-il se baser pour fiabiliser le H-Index [1] ? Cette question est pour l'instant sans réponse et elle mérite une étude approfondie. Le H-Index est un indicateur imparfait de l'impact scientifique d'un auteur sur une communauté. Cet impact est intéressant, mais dans notre optique, l'intérêt n'est pas de « ratisser » tous les articles d'une sommité scientifique, mais de repérer des articles ayant une crédibilité individuelle dans un contexte de recherche donné.

Le H-Index ne permet pas de se faire une idée de la valeur d'un article au travers de la notoriété de son auteur. Il est cependant un gage de crédibilité intéressant.

1. Autrement qu'en s'abonnant à *Thomson scientifics*, qui est un outil que peu de laboratoires peuvent s'offrir

1.2.7 Pour conclure sur les indices qualité de l'information

Dans le contexte de la recherche d'information scientifique et technique, la qualité d'un document se veut « mesurable ». Des indicateurs en batterie sont proposés pour juger de la valeur d'un article selon le nombre de citations y faisant référence, la notoriété de son auteur le contexte de publication. La course à la publication (publier ou mourir...) peut parfois entraîner des problèmes d'éthique plus ou moins graves, allant de l'auto-plagiat à la fraude scientifique comme le montre le tristement célèbre cas Poehlman (Brown, 2011 ; Wright et McDaid, 2011 ; Wager, 2011). Cependant, la volonté première de la scientométrie, comme le montre la section 1.2.2, est simplement d'offrir des indices de qualité dans le contexte scientifique, pas de juger de la pertinence d'un article. Comme nous l'avons montré avec le H-index de Salton (voir figure 1.6, page 32 et figure 1.5 page 32), et le montrerons dans le cas du H-index fourni par Google Scholar d'Ike Antkare (Labbe, 2010), le système a parfois ses failles (voir section 2.3.1 page 40). Mais s'ils sont établis avec de bonnes sources d'information, ces index forment un indicateur fiable.

Chapitre 2

Portails et entrepôts scientifiques

Introduction

La partie la plus longue et la plus fastidieuse du travail d'un chercheur est la recherche liée aux travaux relatifs à son domaine scientifique. Il est malheureusement possible lors de cette étape de passer à côté d'importants développements dans un domaine spécifique.

Il existe, comme vu précédemment, différentes écoles de recherche documentaire. Dans la méthode de Bates, celle que nous privilégions, la procédure consiste habituellement à commencer par la localisation d'un petit nombre de documents initiaux. L'examen approfondi des documents permet d'extraire des données clés. Ces éléments permettront de naviguer sur l'Internet pour faire émerger des documents contigus aux articles originaux. La proximité est établie par sérendipité, que ce soit par liaison bibliographique, co-auteurs, index bibliographiques d'archivages (CDD, ACM...) ou termes clés. Cependant, si cette méthode paraît simple et aisée à mettre en œuvre, un obstacle de taille se dresse souvent devant le chercheur en quête d'informations.

Ce chapitre va permettre de prendre connaissance des différents types de bases de connaissances scientifiques liées de près ou de loin à l'informatique. Nous allons, au cours de cette partie, évaluer les sources d'informations et les moyens d'accès à cette information.

Le passage à l'archivage et à la gestion électronique (GED) de la littérature scientifique a radicalement bouleversé les pratiques des auteurs et éditeurs. Ces changements

ont fortement impacté les usagers et les bibliothèques (Lardy, 2009). De nouveaux modèles économiques de publication et diffusion ont vu le jour (Chartron, 2007), entre accès libre à l'information et logique comptable.

2.1 RI et exhaustivité, une utopie ?

Claire Denecker faisait remarquer, dès l'entrée dans le 21e siècle, que : « les réservoirs d'informations n'ont jamais été aussi nombreux ni aussi facilement accessibles ; pourtant des contraintes très diverses entravent la lisibilité (Denecker *et al.*, 2000, p. 16) ». D'un point de vue général, dès 1998, le premier index Google recensait 26 millions de pages. Dix ans plus tard, en juillet 2008, l'index est passé à un trillion d'entrées (Alpert et Hajaj, 2008). La surabondance d'informations entraîne un risque de suffocation des chercheurs car il est devenu impossible de gérer une telle quantité de données, la profusion semblant parfois aller à l'encontre de l'ergonomie (Denecker *et al.*, 2000, p. 17). L'information arrive sous forme d'un mélange non trié de documents pertinents, de publicités et de documents futiles. Parmi cet amalgame, il est quasiment impossible de faire émerger directement une sélection de documents pertinents (Pochet et Thirion, 1999).

L'expression anglaise « *information overload* », qui décrit ce phénomène peut être traduite en français par infobésité [1]. Ce néologisme, s'il n'est pas très élégant, donne une image mentale précise de l'accès à l'information. Le domaine scientifique n'échappe pas à cette tendance à la surabondance (voir figure B.1 en annexe). Heureusement, les informations à disposition sont crédibles car validées par un comité de lecture composé de scientifiques. Toute la littérature scientifique n'est pas considérée comme ayant la même valeur. La bibliométrie est présentée par les instances d'évaluation scientifiques comme un facteur important pour effectuer son choix parmi les documents présentés (Coutrot, 2008, Filliatreau, 2009).

1. À titre personnel, nous préférons très largement le terme plus poétique et respectueux de *déluge informationnel* proposé par Lévy (1998) à celui d'infobésité de Pochet et Thirion (1999). Cependant l'opposition proposée dans ce titre illustre de notre point de vue le concept de choix qualitatif : faire sa recherche en « gourmet ».

Figure 2.1: Allégorie classique du web visible comme partie émergée de l'iceberg

2.1.1 Notion de visibilité et d'accès sur l'Internet

Notre recherche met en exergue le fait que l'utilisation des moteurs de recherche commerciaux s'est généralisée dans la RI scientifique. Outre la question de la crédibilité des informations trouvées par ce moyen, une autre question est celle de la couverture de la recherche. En effet, si les moteurs de recherches indexent l'ensemble des informations ramenées par leur robots, ils n'ont évidement pas la possibilité de parcourir les sites qui leur sont inaccessibles.

Le Web visible ou accessible

La zone de recherche visible par les moteurs de recherche commerciaux classiques est appelée Web visible. Il s'agit de la plupart des sites web statiques dont l'accès en lecture n'est pas restreint. L'accès, et donc la visibilité d'un site sont grandement facilités par le fait que d'autres hypertextes y font référence par hyperlien.

Le Web invisible ou profond

Les sites dynamiques reliés à une ou plusieurs bases de données ne sont indexés que sur les parties statiques de leur structure. En effet, les contenus de ces sites sont construits par des interrogations des bases de données. Sans interrogation d'un agent, humain ou non [1], capable de saisir la syntaxe de requête, le contenu reste sommaire, voire sans objet. De plus, certains de ces sites sont soumis à des abonnements, payants pour la plupart, ce qui rajoute en opacité pour l'indexation par les moteurs de recherche. Selon Sherman et Price (2002) le web invisible peut être décomposé en quatre sous-catégories.

1. Le Web opaque (*The Opaque Web*) : Il s'agit de pages classiques, qui pourraient donc être indexées par les robots de moteurs de recherche. Elles ne le sont cependant pas faute de liens entrant (pas d'index). Du fait de cette absence de liens entrant, le Pagerank est faible et la page n'est pas indexée. Ces pages sont donc accessibles uniquement par URL et non par navigation.
2. Le Web privé(*The Private Web*) : Ces pages ont une audience volontairement réduite. Elles sont physiquement accessibles par le robot, mais exigent une authentification logicielle pour afficher le contenu. L'authentification peut se faire de manière logicielle par interaction avec l'usager et le contenu s'affiche ensuite.
3. Le Web propriétaire ou *The Proprietary web* : Ces pages ne sont pas accessibles par le robot. Il est aussi possible de proccéder à un accès restreint de manière native par le serveur web. Les méthodes sont nombreuses, qu'il s'agisse de filtrage de provenance (le cas des intranets), fichiers bloquant l'accès [2], ou empêchant l'accès d'un robot.

   Dans une moindre mesure, le spider des moteurs de recherche est sensé respecter les instructions inscrites dans le fichier robot.txt. Ce fichier est un simple fichier texte qui contient des instructions pour les robots afin de limiter l'indexation d'un site aux parties désirée par le webmestre.
4. *The Truly Invisible Web* ou web réellement invisible. Il s'agit des contenus qui ne peuvent être indexés pour des raisons techniques si le format du document est inconnu au spider ou que l'URL est mal formée. Par expérience, le contenu des

1. Grâce à des APIs, certains robots intelligents sont capables de parcourir plusieurs bibliothèques de données pour en extraire des contenus
2. Le serveur web Apache permet de préciser des droits d'accès récursif à un répertoire : le *.htaccess*

fichiers dynamiquement générés depuis une base de connaissances en utilisant la technologie AJAX ne sont pas indexés.

Traditionnellement, les bases de connaissances bibliographiques et les sites des éditeurs scientifiques sont classés entre le web opaque et le web privé. En effet, dans le cadre des éditeurs, les contenus sont généralement accessibles après connexion ; En ce qui concerne les catalogues, mêmes libres, les URL sont générées dynamiquement par le gestionnaire de contenus, ce qui empêche toute indexation. Présentons quelques un des vecteurs d'accès à l'information scientifique numérique.

2.2 Les OPAC

L'acronyme OPAC (*Online Public Access Catalog*) fait référence à la version accessible en ligne de l'interface du catalogue d'une bibliothèque. Ce catalogue permet de chercher un document à partir de ses métadonnées (auteur, titre, date ou mots clés) dans le progiciel de gestion intégré de bibliothèque (SIGB), mais aussi d'en connaître la localisation et/ou la disponibilité.

2.2.1 Le Sudoc

Le Sudoc [1], ou catalogue du Système Universitaire de DOCumentation, est un catalogue français réalisé par les Services Communs de Documentation (SCD), les bibliothèques des établissements de l'enseignement supérieur et de la recherche et l'Agence Bibliographique de l'Enseignement Supérieur (ABES). Selon le site officiel de l'Abes qui pilote le projet, au premier juillet 2010 le catalogue Sudoc comptait plus de 9 millions et demi de notices bibliographiques décrivant tous les types de documents [2]. Ce chiffre a été actualisé à 10 millions en 2012 [3]. Le catalogue Sudoc décrit également les collections de revues et journaux d'environ 2000 établissements documentaires hors enseignement supérieur.

Le Sudoc permet à ses utilisateurs d'obtenir la description bibliographique de documents dans l'optique de constituer une bibliographie via le téléchargement ou

1. `http://www.sudoc.abes.fr/`
2. `http://fil.abes.fr/2010/07/09/les-chiffres-du-sudoc-et-de-star/`, accédé en ligne le 1er octobre 2012.
3. `http://www.abes.fr/Sudoc/Sudoc-public`, accédé en ligne le 1er octobre 2012.

l'export de notices. Si le document référencé est en accès libre (zone blanche), un hyperlien est proposé. Dans le cas contraire, une référence permet de localiser un document dans une des bibliothèques du réseau Sudoc afin de pouvoir le consulter en ligne après authentification (le plus probable dans le cas d'un article scientifique) ou d'en demander le prêt s'il s'agit d'un livre non numérisé.

Le Sudoc offre des possibilités de navigation facilitée avec un hyperlien permettant d'afficher l'historique de consultation. Cette option propose de présenter, sur une même page, toutes les requêtes de la session de travail et les hyperliens dirigeant vers les résultats associés. Une autre fonctionnalité permet d'ajouter les éléments bibliographiques un à un dans un « panier » virtuel pouvant contenir jusqu'à cent notices. Le contenu de ce panier peut être exporté, tout ou partie, vers un affichage écran ou par le moyen d'un courriel. Les résultats sont contextuellement glanables par des plugins de navigateurs, tels Zotero.

Il est à noter que le Sudoc offre une interface mobile pour les tablettes et *smartphones*. Cette option peut paraître de l'ordre du gadget. Cependant elle est en pleine adéquation avec les usages constatés lors de l'étude commandée par le ministère (Jouguelet et Vayssade, 2010). Pour un utilisateur nomade, avoir accès à une source d'informations fiable depuis son périphérique mobile est un facteur d'usage important. Le Sudoc, inscrit dans une démarche de modernité, offre un accès de qualité aux notices bibliographiques. Il offre une localisation physique exhaustive des documents indexés au sein des universités françaises.

2.3 Les moteurs de recherche scientifiques

2.3.1 Google Scholar

Google Scholar est un moteur de recherche de productions scientifiques proposé en 2004 par l'ingénieur principal de Google, Anurag Acharya (Lardy, 2011). Notons que pour une question de visibilité le Sudoc est indexé par Google Scholar grâce à un partenariat entre l'Abes et Google (Bérard et Gibert, 2008). Ce partenariat offre également de l'information correctement indexée à Google Scholar, ce qui lui fait habituellement cruellement défaut (Beel, 2010, Jacsó, 2010). Nous avons noté au cours de nos essais des fonctionnalités de Google Scholar que le formatage des données bibliographiques

est approximatif, que ce soit en matière d'exposition (COinS, doi, embeded RDF) ou BibTEX. Les données elles-mêmes sont partielles, ce qui les rend inexploitables en l'état. Il faut notamment régulièrement réajuster le type du document. Le type « *Inproceedings* » (acte de conférence) est régulièrement remplacé par le générique « *article* ». Un article de revue n'est pas cité de la même façon qu'une communication avec actes. Cette remarque peut sembler de moyenne importance de prime abord. Cependant, la typologie BibTEX oblige à une certaine rigueur. Les revues et conférences scientifiques qui exigent une communication au format LaTEX désirent une bibliographie irréprochable.

Pour conclure sur Google Scholar, nous reprendrons la remarque de Lardy (2011) « On peut donc dire que Google Scholar est un bon point de départ mais qu'il n'a pas encore la maturité des outils de recherche documentaires commerciaux. »

2.3.2 Microsoft Academic Search

Ce moteur de recherche d'articles académiques (actuellement en version bêta) se place en alternative à Google Scholar. Il indexe plus de sept millions et demi de documents et permet une interrogation en texte intégral mais également par auteur, conférence, revue et date. Ces critères peuvent être croisés. L'innovation principale de ce moteur repose sur la détection d'entités nommées. La page de résultats par défaut présente le nombre de citations d'un article et propose un lien supplémentaire lorsqu'il est téléchargeable. Un élément appréciable sur ce moteur de recherche dédié à la science est le graphe de co-écriture pour un auteur (cf. Figure 2.2). Cet affichage permet de comprendre les partenariats d'écriture scientifique.

2.4 Les éditeurs en recherche scientifique génériques

2.4.1 ScienceDirect

ScienceDirect est le service en ligne de l'éditeur de revues scientifiques Elsevier. Il couvre beaucoup domaines de la recherche scientifique dont les sciences humaines et dures (qui nous intéressent plus particulièrement). L'accès à ce site est payant, mais il est souvent disponible au service commun de documentation de l'enseignement supérieur ou à travers des OPACS universitaires qui y sont abonnés.

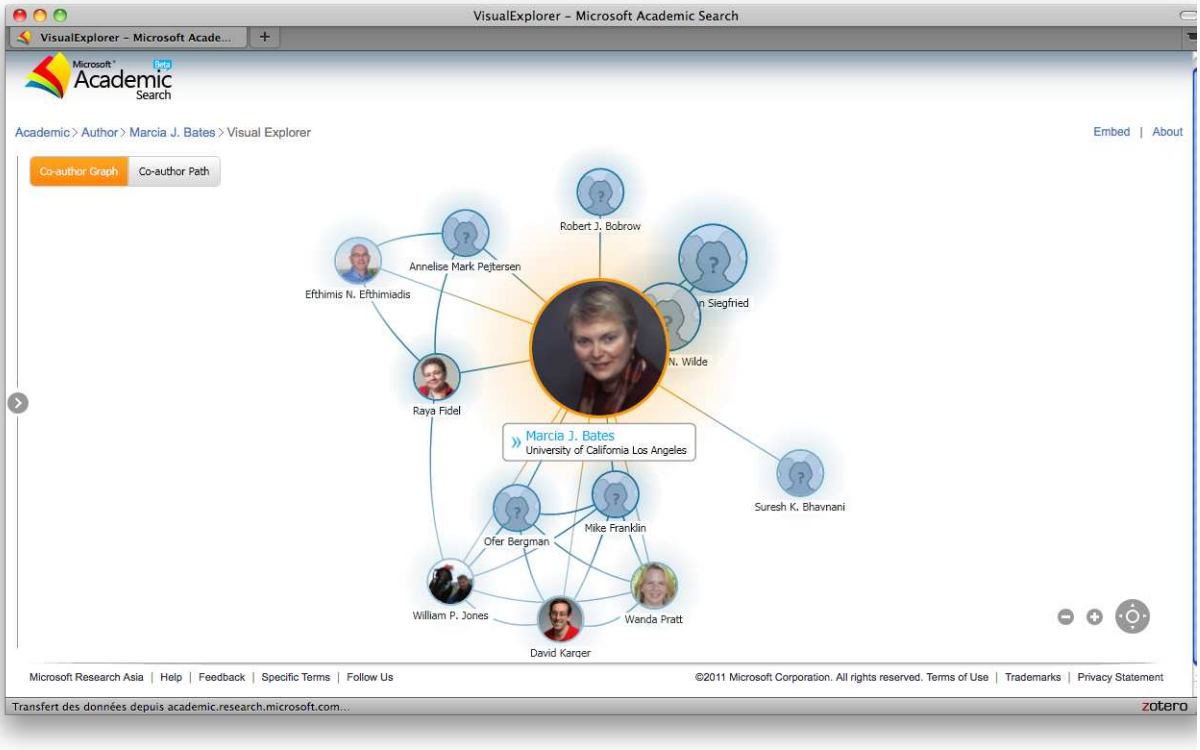


Figure 2.2: Graphe de co-écriture navigable sur Microsoft Academic Search

Les fonctionnalités de recherche simples sont à peu de choses près les mêmes que celles de n'importe quelle interface de recherche.

Une fonctionnalité intéressante est de sauvegarder les recherches ou de les transformer en « alertes » qui enverront un courriel de veille scientifique.

Une recherche avancée permet de spécifier les champs sur lesquels la recherche doit être faite et de spécifier un intervalle temporel. ScienceDirect permet, en mode e recherche avancée, de limiter le champ de sa recherche à des sujets scientifiques spécifiques et d'établir une liste personnelle de revues correspondant le plus à nos attentes.

L'export de citations offre un large panel de compatibilité pour enregistrer des notices : RIS, BibTEX ou en texte non formaté, mais aussi l'exposition de notices dans un format propriétaire compatible avec Zotero.

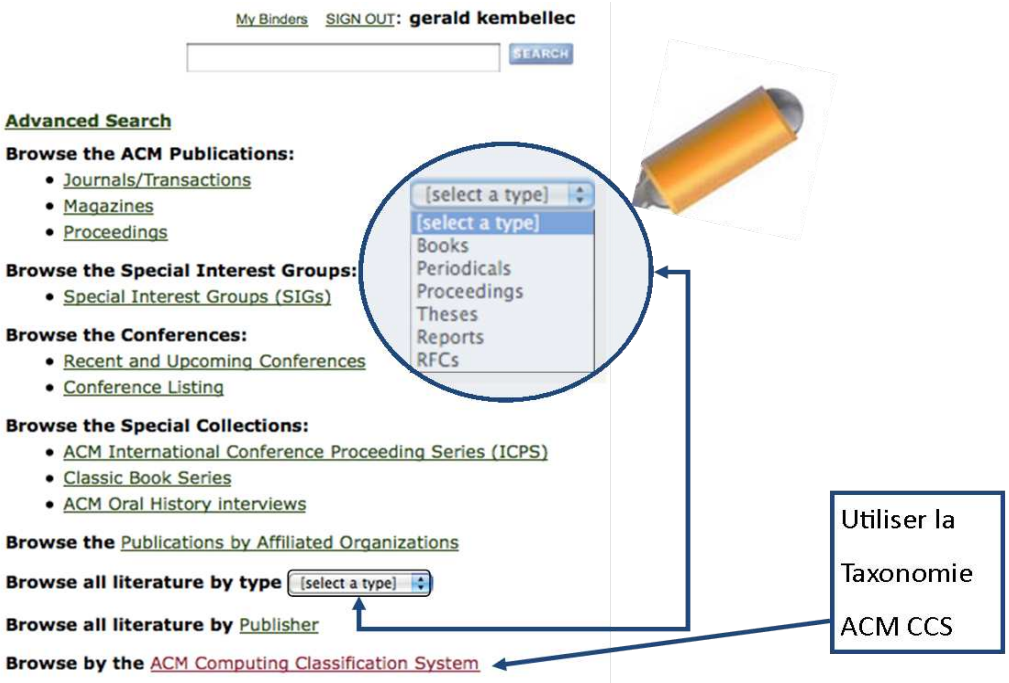


Figure 2.3: Interface classique à l'URL `http://dl.acm.org/`

2.5 Les éditeurs de recherche scientifiques spécifiques

Ce qui fait la force des éditeurs, c'est l'homogénéité et la qualité des productions. Cette qualité a évidemment un coût, payé au prix fort par les bibliothèques universitaires et autres services de documentation.

2.5.1 Le portail de l'Association for Computing Machinery (ACM)

Le portail ACM (Association for Computing Machinery) est l'outil officiel de diffusion de documentation scientifique et technique du consortium ACM. L'ACM s'auto-proclame « la plus grande société informatique scientifique pour l'enseignement et la recherche » sur sa page d'accueil [1] . Elle propose des ressources relatives à l'informatique en tant que science et profession. Dans le cadre de sa mission, l'ACM indexe et donne accès à ses nombreuses revues, actes de conférences et communications, mais aussi aux documents

1. `http://www.acm.org/`

d'autres éditeurs.

Malheureusement, l'interface d'ACM est décrite par le service de documentation de l'Université de Sufolk à Boston comme « étonnamment peu conviviale » [1]. Du point de vue de l'usage, il est courant de commencer par une recherche par mot clé en haut à droite de l'interface d'interrogation de la page de recherche classique que montre la figure 2.3 [2]. Toujours selon la même source (Bibliothèque de l'Université Sufolk de Boston), ce type de recherches aboutit régulièrement à des milliers de résultats, parfois hors-sujets. Pour optimiser la recherche, il souvent obligatoire de passer par la recherche avancée proposée par le portail. Il est également possible de parcourir les résultats d'une requête par type de document. Depuis le portail, le texte intégral de chaque article publié par ACM est accessible par abonnement ou par achat à l'unité. Des notices bibliographiques d'articles de références de chaque grand éditeur sont également disponibles. Nous examinerons plus en détail la classification ACM lors de l'étude comparative des taxonomies en informatique.

2.5.2 Springer Verlag

Springer Verlag (*Springer Science+Business Media*) est un groupe de presse scientifique allemand orienté dans le domaine général des Sciences, Technologies et Médecine (STM), sans spécialisation particulière.

Springer propose un catalogue de 2 000 revues d'un fonds de plus de 70 000 titres sous forme papier et électronique [3].

L'interface web de Springer offre la possibilité de recherche traditionnelle, mais intègre également des fonctionnalités contextuelles de recommandation basées sur le contenu. Ce système appelé *FingerPrint*, basé sur une indexation plein texte du corpus permet de proposer les dix documents considérés comme les plus proches dans un contexte donné (Van der Velde, 2012).

1. « *Sadly, theirs is a surprisingly un-user-friendly interface !* », source : `http://suffolk.libguides.com/content.php?pid=194180&sid=1627535`, accédé en ligne le 1er octobre 2012.
2. Accessible à l'URL `http://dl.acm.org/`, accédé en ligne le 1er octobre 2012.
3. Source : `http://www.springer.com/cda/content/document/cda_downloaddocument/SSBM_facts_figures_2012_F.pdf`, accédé en ligne le 1er octobre 2012.

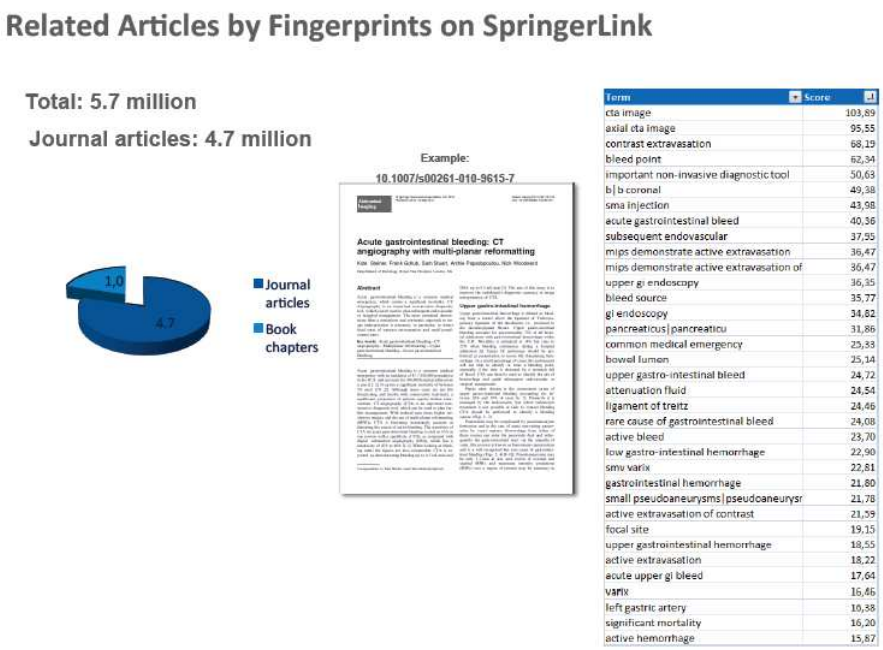


Figure 2.4: Recommandations basées sur les données

2.5.3 IEEE

Une autre base de données d'éditeur scientifique et technique est celle produite et proposée par l'IEEE (*Institute of Electrical and Electronics Engineers*) et la plus importante de leurs sous-société, l'*IEEE Computer Society*. L'IEEE, quant à elle, se proclame « première organisation au monde de professionnels de l'informatique ». Elle se positionne ainsi plus sur l'information technique que scientifique. Toujours selon le service de documentation de l'Université de du Sufolk, la recherche avancée y est un peu plus précise que pour le principal concurrent, l'ACM.

2.5.4 Emerald

Fondé en 1967 en tant que ramification de l'Université anglaise de Bradford [1], Emerald Group Publishing est un des éditeurs universitaires incontournables dans le domaine des affaires et de la gestion. Cet éditeur s'est diversifié avec une forte présence dans les disciplines des sciences sociales, de l'ingénierie, de la linguistique et de

1. http ://www.brad.ac.uk/external/

l'informatique. Toutes les revues de recherche estampillées Emerald sont évaluées par des pairs pour assurer la meilleure qualité (*Peer Review*). En 2009, plus de 21 millions d'articles de recherche et d'études ont été téléchargés sur le site d'Emerald [1].

Emerald propose un moteur de recherche et une interface de navigation taxonomique pour l'accès aux articles. Les articles sont présentés avec leur résumé et les métadonnées associées.

Les articles sont glanables par les outils adaptés à la détection contextuelle tels Zotero ou Mendeley. Un hyperlien offre également un accès direct la page Digital Objet Identifier (DOI) de l'article sur le resolveur officiel (ce qui permet de retrouver l'article même en cas de modification du site Emerald).

Il n'est en revanche pas proposé de format bibliographique comme le RIS ou le BibTEXdepuis la consultation initiale des notices. Il est cependant possible d'ajouter jusqu'à 10 articles dans un *panier* électronique pour la session et de les exporter ensuite au format RIS à l'affichage ou en téléchargement. Un lien hypertextuel propose offre une possibilité d'export vers RefWorks. Une dernière possibilité d'export est de recevoir la bibliographie par courriel. Emerald met à disposition des hyperliens d'étiquetage social pour un enregistrement et une mise à disposition aisée de la fiche d'un document.

2.6 Les bases de connaissance scientifiques

2.6.1 DBLP

La base de connaissance *Digital Bibliography & Library Project* (DBLP) est une base de connaissance issue d'un projet universitaire. Nous décrivons en détail les divers aspects de ce projet dans la présentation de notre outil (voir le paragraphe 11.7.3) et dans l'annexe F.

2.6.2 Les bases documentaires à comité de sélection

Le principe des bases documentaires à comité de sélection est de proposer des notices bibliographiques soigneusement rédigées et indexées sur les thématiques spécifiques.

1. Chiffres les plus récents fournis sur le site officiel dans la rubrique *about Emerald* :`http://www.emeraldinsight.com/about/index.htm`, accédé le 1er octobre 2012.

Bases	URL ou mode d'accès	Type	Références
FRANCIS	Bases de connaissances fédérées	SHS	2 500 000
URBADOC	`http://www.urbadoc.com/`	Urbanisme	705 000
PubMed	`http://www.ncbi.nlm.nih.gov/pubmed`	STM (médecine)	22 000 000
ERIC	`http://www.eric.ed.gov/`	Education	1 400 000
PASCAL	Bases de connaissances fédérées	STM	17 000 000
INSPEC	`http://www.theiet.org/resources/inspec/`	STM	13 000 000

Tableau 2.1: Quelques bases de références bibliographiques à comité de sélection (au 1er octobre 2012)

Parfois l'accès au document primaire est proposé en accès libre, d'autres fois les liens vers les documents scientifiques dirigent vers des portails d'éditeurs. Voici quelques bases documentaires parmi les plus réputées et les domaines associés : Les bases de connaissances sont parfois accessibles directement grâce à une interface dédiée, mais le plus souvent on y accède depuis une interface dédiée. Le contenus d'Eric, Francis et Pascal sont par exemple accessibles par une recherche fédérée au moyen de l'interface d'Ebsco [1] (voir 2.8.1).

2.7 Les archives scientifiques ouvertes

Une archive scientifique ouverte ou base de dépôt est un portail où les professionnels (ingénieurs, chercheurs, enseignants) en sciences et techniques peuvent mettre leur production à la disposition de la communauté scientifique en accès libre. L'objet de ces réservoirs (*repositories*) de connaissances n'est pas de juger la valeur scientifique et technique d'un document mais de rendre possible la diffusion électronique de la littérature. L'accès se veut complètement gratuit et sans restriction. Les plateformes de dépôts ouvertes n'effectuent donc par de relecture par des pairs. En effet, globalement, les documents archivés ont déjà subi cette étape, les documents sont déclarés *postprint* ou postpublication. Il est néanmoins possible de soumettre des rapports ou autres

1. Interface accessible à l'URL `http://search.ebscohost.com/`.

Archive	URL	Type	Nombre de documents disponibles
HAL	http ://hal.archives-ouvertes.fr/	Générique	203 000
ArXiV	http ://arxiv.org/	STM	705 000

Tableau 2.2: Hal et ArXiV au 1er octobre 2012

documents non publiés, ce qui sera spécifié. Ce type de documents n'ayant pas subit le jugement des pairs est nommé *preprint* ou prépublication.

La déclaration *Budapest Open Access Initiative* (Chan *et al.*, 2002) définissait l'accès libre à la littérature de la manière suivante : « Mise à disposition gratuite sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet ».

L'intérêt de ces systèmes est double. Le premier intérêt est l'interopérabilité des protocoles de mise à disposition (OAI-PMH). Un unique outil correctement configuré pourra donc tous les parcourir (parser) en un seul moissonnage. Le deuxième avantage de ces dépôts est la gratuité du service offert, même si l'intégralité du catalogue n'est pas obligatoirement en accès libre de droit. En effet, parfois il est illégal de déposer un article sur lequel il a été signé une renonciation de droit d'auteur (le *copyright transfert agreement*). Dans ce cas, une notice bibliographique peut être rédigée et soumise sur les archives ouvertes. Cette notice comprenant le titre et le résumé, il est possible de se faire une idée du document et de le chercher par d'autres moyens, notamment en SCD. L'auteur, de son côté, gagne en visibilité (Odlyzko, 2002) grâce à un public élargi (Antelman, 2004).

2.7.1 ArXiV et Hal

La première archive ouverte fut proposée par P. Ginsparg, physicien du laboratoire national de Los Alamos. Ginsparg avait eu l'idée de proposer un cadre collectif a tous ses collègues qui avaient pour habitude de créer des micro-systèmes éparses issus d'initiatives individuelles, le plus souvent en FTP (Bosc, 2005). ArXiV offre au 1er octobre 2012 un accès direct et gratuit à plus de 787 000 documents, principalement

dans les STM. Les documents mis à déposés sur ArXiV ne sont parfois pas encore soumis à un comité de lecture, ce qui peut être sujet de polémique sur leur validité scientifique. Cependant comme les auteurs peuvent faire évoluer leurs archives sur arXiv, la version validée par des pairs peut être déposée, si cela ne va pas à l'encontre du contrat de publication.

La problématique de Hyper articles en ligne (HAL) est sensiblement la même que celle d'ArXiV. HAL est une archive ouverte proposée et gérée par le Centre pour la Communication Scientifique Directe (CCSD) du CNRS qui propose un accès à 203 000 documents au 1er octobre 2012.

2.8 Les interfaces de recherche fédérées

2.8.1 Ebsco et Couperin

Pour regrouper et faciliter l'accès aux nombreuses bases de connaissances, des portails d'accès fédérés proposent un service à valeur ajoutée. Le leader du marché, Ebsco vend un service d'aide à l'acquisition et la gestion des abonnements périodiques aux entreprises, administrations, centres de recherche, bibliothèques académiques, médicales ou publiques [1]. Ce produit vaut aussi bien pour les abonnements papiers que pour l'accès aux bases de connaissances en ligne auprès de 95 000 éditeurs avec 375 000 titres de périodiques au niveau mondial.

Couperin est une association à but non lucratif dont l'objectif est de négocier l'acquisition de périodiques numériques aux meilleures conditions. Acteur majeur de la diffusion scientifique dans les milieux de l'enseignement supérieur et de la recherche, Couperin prose un accès à l'information scientifique et technique pour plus de 200 établissements d'enseignement supérieurs [2]. Au premier octobre 2012, Couperin offre un accès négocié à 122 fournisseurs en ligne, parfois directement, parfois via un intermédiaire comme Ebsco. [3].

1. Source :`http://www2.ebsco.com/fr-fr/app/AboutUs/Pages/abouteis.aspx`, accédé le 1er octobre 2012.
2. Information consultable à l'URL `http://www.couperin.org/fr/presentation/notre-organisation/les-membres-de-couperin`, accédé le 1er octobre 2012.
3. Information consultable à l'URL `http://www.couperin.org/fr/negociations/liste-des-negociations?filter_16=conclue`, accédé le 1er octobre 2012.

2.8.2 Isidore

Isidore est une plateforme de recherche fédérée spécialisée en sciences humaines et sociales (SHS) proposée dans le cadre du très grand équipement ADONIS du CNRS. La mise en œuvre de ce projet est le fruit du Centre pour la Communication Scientifique Directe (CCSD) qui gère également HAL. Son accès est libre et gratuit à tous. Les publics cibles sont ceux de la recherche, en particulier enseignants, chercheurs, doctorants et étudiants [1]. Nous ajouterons à cette liste les documentalistes liés à la recherche, comme ceux de centres communs de documentation universitaire.

Dans le cadre de cette thèse, nous notons avec intérêt la convergence de bonnes pratiques liées à la réalisation de ce système de recherche d'information. En effet, Isidore s'appuie structurellement sur les principes du web de données et donne accès à des données en accès libre. Nous discuterons plus largement de ces aspects (et d'Isidore en particulier) dans le chapitre 9 relatif à l'urbanisation des systèmes d'information en contexte documentaire.

2.8.3 Bielefeld Academic Search Engine (Base)

BASE est l'un des moteurs de recherche en IST indexant le plus de documents scientifiques au monde. C'est particulièrement vrai pour les ressources universitaires et les archives ouvertes. BASE est géré par la bibliothèque de l'Université de Bielefeld. Ce projet fut lancé par la bibliothèque de l'université allemande Bebielefeld en septembre 2004. L'index du moteur recensait alors presque 680 000 documents au travers de quinze sources [2]. En janvier 2012, plus de 33 000 000 documents étaient indexés pour 2 100 sources, dont ArXiV et PubMed (médecine) pour l'international et Les principales sources françaises comme Gallica, Cairn, HAL, revues.org et Persée . Base s'alimente également auprès de nombreuses universités à travers le monde. Comme le mouvement de libre accès à l'information se développe et prospère, les serveurs de dépôts sont de plus en plus nombreux à fournir un service normalisé OAI-PMH. L'accès aux textes intégraux est possibles pour environ 75% des documents indexés [3]. La charte morale revendiquée par Base est très axée sur la qualité. Le projet Base propose :

1. Source :`http://www.rechercheisidore.fr/apropos`, accédé en ligne le 1er octobre 2012
2. Source : `http://base.ub.uni-bielefeld.de/en/about_statistics.php`
3. Chiffres fournis par la bibliothèque de l'université de Bielefeld.

- de cautionner intellectuellement ressources sélectionnées ;
- de ne relayer que les serveurs dont les documents satisfont aux exigences spécifiques de pertinence et qualité académique ;
- un inventaire des métadonnées qui facilitent les recherches ;
- d'offrir un point d'accès aux ressources du « Web profond », celles là-mêmes qui sont ignorées par les moteurs de recherche commerciaux (ou perdues dans le déluge informationnel) ;
- un affichage des résultats de recherche comprend des données bibliographiques précises et fiables ;
- plusieurs options pour trier la liste des résultats (facettes).
- de parcourir par la classification décimale de Dewey pour accéder aux informations.

2.9 Portails de revues

La plupart des revues francophones de sciences humaines en ligne sont regroupées au sein de portails :

2.9.1 Revues.org

Le portail Revues.org se présente comme le doyen des sites français en sciences humaines et sociales (SHS) avec pour ambition première de promouvoir les revues francophones. Cependant, Revue.org accueille en 2012 aussi bien des revues éditées en France qu'à l'étranger en français, écrites pour partie ou en totalité en anglais, espagnol, portugais, russe et même basque.

Ce portail est animé par le Centre pour l'Édition Électronique Ouverte (Cléo). Basé à Marseille et Paris, cette unité associe le Centre national de la recherche scientifique (CNRS), l'École des Hautes Études en Sciences Sociales (EHESS), l'Université de Provence et l'Université d'Avignon.

2.9.2 Persée

Persée est un programme de publication électronique de revues scientifiques en sciences humaines et sociales. Les collections imprimées de revues sont numérisées et

mises en ligne sur un portail avec des possibilités avancées d'exploitation de ces corpus numérisés.

L'une des ambitions de Persée est d'offrir des services et des outils permettant une exploitation enrichie des documents.

Des accords de coopération sont en cours de formalisation avec les principaux portails francophones assurant la diffusion de la production courante de revues scientifiques. L'objectif est d'offrir aux lecteurs une continuité dans la consultation des fonds entre Persée et les autres sites.

Persée repose sur un certain nombre de normes et de standards ouverts qui garantissent la possibilité de réutilisation des données, une utilisation optimale du site web par tout internaute, l'interopérabilité du portail et des possibilités étendues de mutualisation avec d'autres outils du même type.

L'ensemble des développements réalisés dans le cadre du programme PERSEE sont « *Open Source* ». Ils sont en effet dotés une double licence CeCCIL et GPL et pourront être réutilisés dans le cadre d'autres projets de numérisation et/ou de diffusion de documents.

2.9.3 Érudit

Fondée en 1998, la plateforme Érudit est un consortium inter universitaire (Université de Montréal, Université Laval, Université du Québec à Montréal) et un organisme sans but lucratif qui donne accès à plus de 80 revues savantes, 27 revues culturelles, une cinquantaine de livres et actes, 30 000 mémoires et thèses, et près de 3 000 documents et données provenant de centres de recherche subventionnés par le Fonds québécois de recherche sur la société et la culture (FQRSC).

Les documents diffusés sur la plateforme sont produits par le Centre d'expertise numérique pour la recherche de l'Université de Montréal et par la Bibliothèque de l'Université Laval.

Les services d'édition sont basés sur des normes internationalement reconnues et sur des formats normalisés (Unicode, XML, XHTML, TIFF, PDF).

Érudit offre aux revues représentées une vitrine internationale grâce à une stratégie de référencement misant sur les partenariats et l'indexation dans des outils de recherche spécialisés et des bases de données disciplinaires.

2.9.4 Cairn

Cairn.info est né de la volonté de quatre maisons d'édition (Belin, De Boeck, La Découverte et Erès) ayant en charge la publication et la diffusion de revues de sciences humaines et sociales, d'unir leurs efforts pour améliorer leur présence sur l'Internet, et de proposer à d'autres acteurs souhaitant développer une version électronique de leurs publications, des outils techniques et commerciaux développés à cet effet.

En février 2006, la Bibliothèque nationale de France s'est associée à ce projet, de façon à faciliter le développement d'une offre éditoriale francophone, sous forme numérique.

Cairn.info réunit, en outre, différents investisseurs institutionnels, notamment Gesval, la société ayant en charge la gestion des participations de l'Université de Liège.

L'ambition de Cairn.info est d'aider les maisons d'édition, organismes ou associations ayant en charge des publications de sciences humaines francophones à gérer la coexistence des formats papier et électronique. Dans ce but, les services de Cairn.info couvrent à la fois la fabrication papier et électronique, la distribution papier (gestion des abonnements pour les revues, routage) et électronique (texte intégral en ligne, distribution des métadonnées auprès des sites et bases bibliographiques), ainsi que la diffusion et la promotion de ces publications auprès des publics auxquels elles s'adressent.

2.10 Conclusion

Notre but est la réalisation d'une interface de recherche en informatique afin de trouver de la littérature scientifique pertinente. Le portail ACM semble être particulièrement adapté à cette tâche. Nous pensons donc établir pour ce portail une librairie (au sens informatique du terme) pour automatiser la recherche depuis notre futur outil. Le projet DBLP offre également une intéressante collection de fiches bibliographiques. Cependant, connaître les potentielles sources de documentation scientifique ne nous donne pas d'information sur les méthodes de recherche, les attentes et les réels besoins de l'usager. Examinons ce qu'est réellement la recherche d'information, quels en sont les enjeux et les méthodes depuis plus basiques à celles les plus sophistiquées.

Chapitre 3

La recherche d'information

Tout ce que je sais, c'est que je ne sais rien..

Socrate

Introduction

Le terme de recherche d'information (*Information Retrieval*) est apparu pour la première fois en 1948 dans le mémoire de MASTER du MIT [1] de Mooers (Mooers, 1948). Dinet et Rouet définissent la recherche d'information (RI) comme « l'activité d'un individu qui vise à localiser et traiter une ou plusieurs informations au sein d'un environnement documentaire complexe, dans le but de répondre à une question ou de résoudre un problème (Dinet et Rouet, 2002) ».

Dans le premier chapitre de *An Introduction to Information Retrieval*, Manning *et al.* (2008) nous offrent la définition suivante de la recherche d'information : « *As*

1. *Massachusetts Institute of Technology*

Figure 3.1: Le modèle « simpliste » Morville et Rosenfeld (2006) de système d'information

an academic field of study, information retrieval might be defined thus :Information retrieval (IR) is finding material (usually documents) (...) that satisfies an information need from within large collections (usually stored on computers)[1]. »

Dans le troisième chapitre de leur livre *Information architecture for the World Wide Web*, Morville et Rosenfeld présentent la vision « grand public » d'une recherche d'information (cf. Fig. 3.1). Cette idée reçue est qu'il suffit de poser une question à un moteur de recherche et la « boîte magique » donnera la réponse ultime relativement à notre besoin de connaissances par rapport au sujet. Toujours selon Morville et Rosenfeld, ce type de résultat fait figure de cas isolé, pour ne pas parler d'exception (Morville et Rosenfeld, 2006, chp. 3). Une requête informationnelle est dépendante du fonctionnement du système d'informations et pas seulement de la personne qui l'initie. Posons les définitions suivantes proposées par l'association des professionnels de l'information et de la documentation (ADBS)[2] :

Définition de Recherche d'Information

> Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés ADBS (2012).

1. Proposition de traduction. « En tant que discipline académique, la recherche d'information peut être définie ainsi : La recherche d'information (RI) est le repérage au sein de grandes collections (généralement stockées sur les ordinateurs) de ressources (habituellement des documents) qui répondent à un besoin d'informations. »

2. `http://www.adbs.fr/vocabulaire-de-la-documentation-41820.htm`, accédé le 1er août 2012

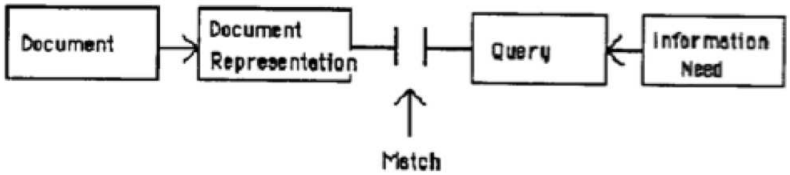


Figure 3.2: Modèle trivial de SRI proposé par Bates

Définition de Recherche documentaire

> Ensemble des méthodes, procédures et techniques ayant pour objet de retrouver des références de documents pertinents (répondant à une demande d'information) et les documents eux-mêmes ADBS (2012).

Définition de Recherche bibliographique

> Ensemble des méthodes, procédures et techniques ayant pour objet de retrouver les références bibliographiques de documents pertinents ADBS (2012).

De ces trois définitions, nous dégagerons une synthèse fonctionnelle de la recherche d'information, qui en tant que processus ne peut être réellement séparée des recherches documentaire et bibliographique. Notre vision sera donc pragmatique, nous voyons la recherche bibliographique comme la finalité de la recherche documentaire, elle-même résultante de la recherche d'information.

Marcia Bates, reprenant les concepts de bases posés par Robertson, proposait un schéma volontairement trivial (cf. Fig. 3.2) modélisant un système de recherche d'information (SRI) dans un contexte idéal (Bates, 1993, Robertson, 1977). Pour elle, il s'agit d'une représentation utopique, bien que répandue, de l'offre documentaire, avec le besoin d'informations formalisé par le biais d'une requête.

Nous ajouterons que le but de cette activité peut être purement cognitif, c'est-à-dire élargir son champ de connaissances dans un domaine (dans notre optique scientifique) pour tenter de le synthétiser pour mieux en saisir la substance. Il ne s'agit pas dans ce cas à proprement parler d'un besoin d'information, mais d'une forme de curiosité, l'envie de mieux appréhender un champ de connaissance.

3.1 Le paradoxe de la RI

Ce qui est paradoxal dans la RI c'est qu'elle peut traduire :

- Un besoin de références bibliographiques pour structurer et étayer sa connaissance et ses idées ;
- Directement un besoin de connaissances, c'est-à-dire combler une lacune.

Cependant dans ce dernier cas, la prise de conscience de ce besoin, ou manque informatif, découle d'une expérience du domaine [1]. Pour comprendre son besoin d'informations, il faut avoir déjà effectué un panorama du champ de connaissances (Boubée *et al.*, 2005). Si nous explicitons différemment les choses, pour comprendre son ignorance, il faut déjà avoir commencé à chercher. Y. F. Le Coadic définissait cet état de connaissance : « nous en savons assez pour savoir que nous avons un besoin d'information, mais nous n'en savons pas assez pour pouvoir poser les bonnes questions (Le Coadic, 2008) ». Cette réflexion peut être illustrée par la citation de l'en-tête de chapitre attribuée à Socrate : *Tout ce que je sais, c'est que je ne sais rien.* L'une des premières difficultés pour les usagers est d'identifier les sources pertinentes et d'avoir une vision claire des contenus.

Nous allons dans cette première partie étudier dans le détail les différents aspects de la recherche d'information dans les systèmes numériques. Nous étudierons les interfaces de recherche d'information et les mécanismes qui y sont associés.

1. Nous voyons le détail du besoin d'information et les aspects psyscho-cognitifs qui y sont liés dans le chapitre 5 « Les écoles de pensées en RI : Processus et Cognition »

3.2 Concepts, modèles et méthodes en RI

Internet offre une vaste collection de documents, et son utilisation comme source d'informations est évidente et est devenue très populaire. Comme l'ont souligné et analysé Dennis *et al.* (2002), il y a pléthore de technologies de recherche d'information en ligne, qui peuvent principalement être classées en quatre catégories :

1. Recherche par mots clés, sans aide. Un ou plusieurs termes de recherche sont entrés et le moteur de recherche renvoie une liste classée des résumés de documents hyperliés.

2. Recherche assistée. Le moteur de recherche produit des suggestions ou recommandations basées sur la requête initiale de l'utilisateur.

3. Recherche par classification. L'espace d'information est divisé en une hiérarchie de catégories, où l'utilisateur navigue du générique vers le spécifique.

4. Requête par l'exemple. L'utilisateur sélectionne un élément intéressant d'un hypertexte, qui est ensuite utilisé comme base d'une nouvelle requête.

Nous ajouterons à cela une cinquième catégorie, la recommandation qui propose de l'information potentiellement intéressante en fonction de l'usager et/ou du contexte de recherche.

Les outils de recherche d'information se divisent en deux catégories radicalement distinctes. Les portails de recherche, ou annuaires web sont des sites Internet qui proposent des liens vers un florilège de sites repérés par des experts d'un domaine pour leur qualité. Les moteurs de recherches sont des systèmes complexes permettant de trouver des ressources dans un corpus numérique. Ce corpus peut être l'Internet dans sa globalité, une base de connaissances ou un seul site web.

3.2.1 Portails de connaissance

Un portail de connaissance est un site de référence dans un domaine précis ou une page hypertexte dédiée à une communauté particulière. Ce site se présente sous la forme d'un ensemble de pages web hyperliées. Un portail peut être perçu comme un point d'entrée sur un panel de ressources autour d'un thème commun. Souvent, ces portails offrent une vingtaine de catégories pour le premier niveau de la classification. Le type des documents référencés et agrégés importe moins que leur spécificité commune : la thématique. Le plus souvent, ces indexations procèdent d'une intervention humaine. C'est le cas des portails spécialisés de Wikipédia . En effet, l'encyclopédie participative en ligne possède un portail d'accès pour chaque grande thématique. Ces thématiques sont animées par des groupes d'intérêt, qui compilent des hyperliens vers les articles au sein des portails. Un autre exemple de portait particulièrement intéressant, parce que géré manuellement, était l'annuaire Google dédié à l'informatique. Une hiérarchie de sujets prédéfinis comme l'informatique, le sport, l'art ou la musique est maintenue et enrichie de manière manuelle. D'après Eissen et Stein (2002), ces hiérarchies statiques ne sont pas satisfaisantes à deux égards :

1. elles nécessitent un effort de maintenance humaine considérable ;
2. pour des sujets particulièrement spécifiques, les catégories de navigation génériques, ou points d'entrée, sont inutiles et allongent considérablement le processus de recherche.

Ces deux points sont indéniables. À moins de justifier le recours à un comité d'experts par sa préexistence (certains sites internet sont créés par un comité d'experts, c'est le cas d'IEEE ou d'ACM), cette méthode est onéreuse et chronophage. De plus, l'accès à l'information pointue est ralenti, voire rendu malaisé si la classification n'est pas triviale. Ce dernier point peut contrevenir avec l'immuable règle des trois clics (Scapin et Bastien, 1997) qui situe le seuil de tolérance d'un usager de système d'informations à un maximum de trois clics pour trouver l'information désirée dans un hypertexte (Sottet *et al.*, 2005). Ainsi, Google a fermé ses services d'annuaire durant l'été 2011. Les services ont été repris depuis par leur initiateur historique Dmoz [1].

1. L'*Open Directory Project*, ou *Directory Mozilla* qui donne son nom au site, Dmoz.org, est un répertoire de sites web créé en 1998, sous licence *Open Directory*. Il est géré par une vaste communauté d'éditeurs bénévoles provenant du monde entier, chacun étant responsable de vérifier l'exactitude et la

3.2.2 Moteurs de recherche

Un moteur de recherche est un outil dont l'interface permet de localiser une information dans une base de données à partir d'une requête. Originellement, les moteurs de recherche étaient des programmes installés localement sur les ordinateurs et ils consultaient, soit des bases locales, soit des bases distantes à travers des protocoles tels le FTP ou Gopher. Avec l'émergence du protocole HTTP, de nouveaux types de moteurs de recherches apparurent. Il s'agit d'une base de données indexant le contenu des pages référençables sur le web visible. Cette base est alimentée par un robot qui parcourt en permanence l'Internet.

Modèle fonctionnel

Curt Franklin expliquait qu'un moteur de recherche est composé de quatre parties principales (cf. figure 3.3 [1]) qui sont détaillées ci-dessous (Franklin, 2000) :

1. Un robot collecteur (web crawler) qui parcoure l'Internet de page en page. Un robot d'indexation est également appelé *crawler* ou *spider*. Il s'agit d'une analogie à une araignée qui parcourt inlassablement la toile, se déplace dans le web de site en site pour en collecter le contenu. Le contenu et les URL sont intégrés à une base de connaissances pour être traités.
2. Un indexeur chargé d'archiver les pages et d'en extraire les termes clés. Les URL des pages et les mots clés seront ensuite intégrés dans une base de données associative. L'indexation des pages repose sur les métadonnées collectées et celles calculées (mots clés calculés par traitement automatique du langage). Le système confronte les mots clés proposés par le webmestre éditorial et les termes émergeant de calcul statistiques. Un système de pondération permettra de faire l'équilibre entre ces données et de proposer des termes représentatifs du contenu de la page.
3. Un index inversé qui est une deuxième entrée sur la base de connaissance qui associe à chaque concept l'ensemble des URL des documents qui sont pertinents.

catégorisation des sites dans une ou plusieurs catégories. Ce service est toujours accessible en français à l'url `http://www.dmoz.org/World/Fran%C3%A7ais/`

1. Crédit image : `http://computer.howstuffworks.com/internet/basics/search-engine1.htm`, accédé le 1 août 2012

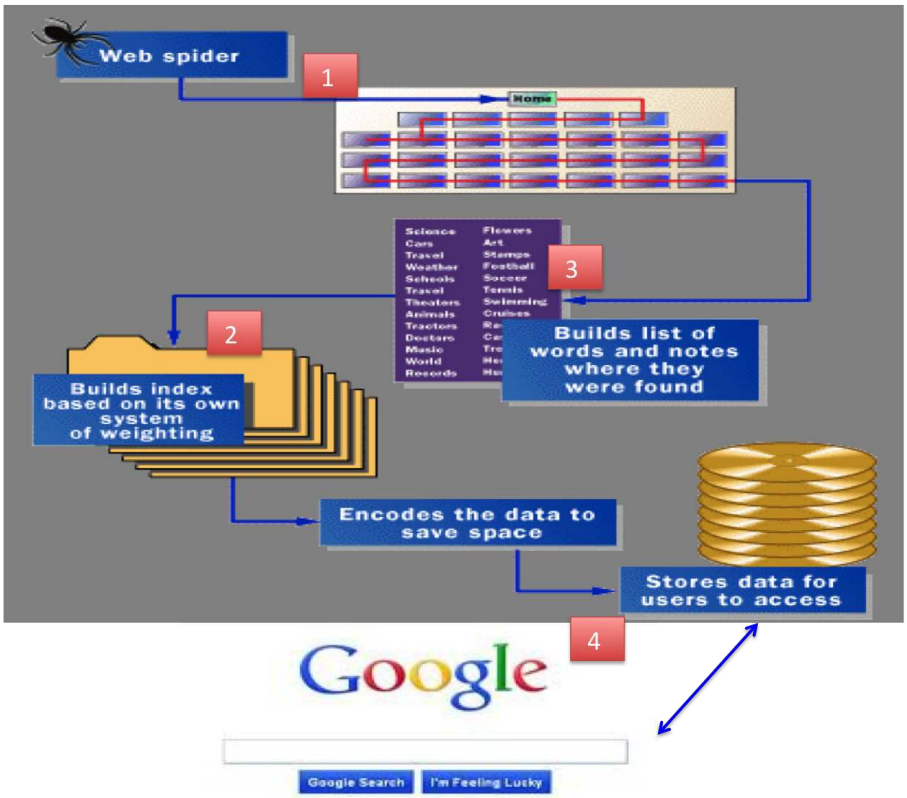


Figure 3.3: Schéma fonctionnel d'indexation d'un moteur de recherche.

4. Une interface de consultation qui a pour rôle d'aider les utilisateurs à interroger facilement la base de données sans connaître le langage d'interrogation de la base.

Les principaux robots sont Googlebot, Slurlp de Yahoo et ExaBot du français Exalead. Grâce aux dernières moutures de son algorithme d'indexation (*Panda* et *Penguin*) et aux robots qui parcourent inlassablement internet, le leader Google ambitionne de ne proposer que de la qualité à ses utilisateurs dans les premiers des résultats de recherche. Le moteur va éliminer tous les sites ayant un contenu de piètre qualité, ou résultant d'un copier-coller depuis un autre site. Cette expérimentation est en cours en France depuis la mi-août de l'année 2011 pour *Panda* et avril 2012 pour *Penguin* et nous n'avons pas encore le recul pour juger des résultats, si ce n'est par les baisses de classement des sites trop sémantiquement pauvres. Cependant, comme Google avait appelé dès le début de l'année 2011 à adopter de bonnes pratiques, entre autres l'usage de métadonnées descriptives, il est à espérer une amélioration des résultats.

Traitement du langage naturel

Quand un utilisateur entre une requête dans le formulaire d'interrogation d'un moteur de recherche, il ne se doute pas de la transformation que va subir sa requête. L'utilisateur va peut-être utiliser le moyen d'expression qui lui est familier, à savoir une phrase construite complexe. Cette phrase aura une syntaxe propre et peut être même une orthographe qui lui sera particulière. Nous appellerons ce mode d'expression le langage naturel.

Un des problèmes majeurs auxquels font face les concepteurs de systèmes de recherche d'information (SRI) est la correspondance entre l'information désirée par l'utilisateur avec celle contenue dans les documents indexés. Cette problématique est d'autant plus complexe que la requête exprimée est une étape intermédiaire qui s'intercale entre l'information désirée par projection mentale et les informations disponibles. En 1950, Alan Turing prédisait qu'il ne serait pas possible de communiquer de manière naturelle avec une machine avant la fin du 20e siècle (Turing, 1950) [1]. Malheureusement, à l'heure actuelle aucune solution matérielle ou logicielle n'est à même de réaliser cet exploit (Saygin *et al.*, 2000).

Cependant, grâce à la lemmatisation, à la détection des expressions composées, aux thésaurus et aux réseaux sémantiques, il est possible pour un automate de traiter convenablement une requête en langage naturel pour en extraire les concepts dominants.

De manière générale, avant de commencer à comparer les éléments de la requête avec la base de connaissances, le travail suivant est effectué :

1. Le correcteur orthographique va aligner les termes saisis avec une orthographe cohérente et éliminer les risques d'erreurs de saisie.
2. Les termes composés de plusieurs mots vont être détectés pour être traités comme une seule entité.
3. Le texte va être segmenté en termes (Kan *et al.*, 1998).
4. Une désambiguïsation va tenter de régler les problèmes de polysémie.
5. Un lemmatiseur ou un stemmer va réduire les termes à leur racine.
6. Les mots non discriminants vont être exclus de la requête.

Revoyons en détail ces différents points de traitement de la requête initiale.

1. `http://www.loebner.net/Prizef/TuringArticle.html`, consulté le 22 juin 2012

La correction orthographique

Pour contextualiser la correction orthographique ou grammaticale de termes dans le cadre d'une recherche d'information, Sitbon *et al.* (2007) rappellent que le traitement est une réécriture en vue d'un traitement automatique et non pas une correction complète dans le but de fournir une correction complète tant du point de vue de la grammaire que de l'orthographe. Nous distinguerons la correction grammaticale et celle orthographique. Dans le cadre d'une correction purement orthographique, l'outil compare les mots d'une requête avec ceux d'un dictionnaire. Les problèmes liés à la mauvaise écriture d'un mot dans une requête peuvent avoir des causes multiples :

- Un mauvais usage du périphérique d'entrée, comme une faute de frappe.
- La dysorthographie qui est un trouble de la production écrite généralement associé à la dyslexie ou à l'inattention.

Selon Sitbon *et al.* (2008), dans le cas d'une dysorthographie, les troubles les plus courants sont :

- Segmentation en mots erronée (Gillon, 2004b), exemple :
  re-cherche d'un fort ma Sion au lieu de recherche d'information.
- Erreurs de conversion entre graphème et phonème, exemple :
  Unphormassion au lieu d'Information.
- Confusions de phonèmes, exemple :
  Monné au lieu de Monnaie.

Le traitement de ces troubles est organisé dans le cadre d'un correcteur orthographique par des algorithmes qui calculent la présence d'un mot dans un dictionnaire et à défaut propose une solution de remplacement. Ce peut être par exemple l'utilisation distance de Levenshtein qui calcule la distance minimale entre le mot dysorthographié et un autre mot du dictionnaire (Dice, 1945, Levenshtein, 1966).

Le rapprochement et la segmentation

Le rapprochement est la première étape de la segmentation, ou *tokenisation*. La *tokenisation* consiste à séparer les lemmes entre eux. La détection d'expressions composées de plusieurs mots dans une requête peut être traitée de plusieurs manières. Pour segmenter un texte, Sitbon et Bellot (2005) proposent de le séparer en chaines lexicales cohérentes au niveau du sens, ce qui n'est pas applicable dans ce contexte. En effet, dans le cadre d'une requête, le texte est trop court pour être segmenté en chaines

préfixe 1	lexème 2	suffixe
over	clock	ing
sur	cadence	ment

Tableau 3.1: Exemple bilingue de racinisation

lexicales. Cependant, si cette méthode sert principalement à résumer automatiquement un texte, une recherche d'entités nommées peut être utilisée pour analyser le contenu d'une requête. On appelle entité nommée dans un texte tout ce qui fait référence à un concept unique. En se basant sur les travaux de Chinchor et Robinson (1997), Sitbon et Bellot proposent d'utiliser trois types d'entités nommées à partir d'un lexique fermé : listes de noms de personnes, noms de lieux et noms d'organisations (Sitbon et Bellot, 2005). Originellement, Chinchor proposait un traitement textuel qui offrait en sortie un texte reformaté au format XML et traitait également les unités temporelles (dates, horaires), et les quantitatives (valeurs monétaires et pourcentages). Si un tel travail est fort utile dans une quête de sens, les moteurs de recherche font rarement de la classification, juste de la segmentation.

La lemmatisation, racinisation, troncature et désambiguïsation

En poursuivant notre étude sur les méthodes de traitement d'une requête dans un moteur de recherche, nous allons distinguer trois types de manières de « traiter » les termes qui composent la requête. Après le rapprochement des mots dans les termes composés, le passage de la liste de mots vides, les termes restants vont être traités pour éliminer les variations morphologiques. Chaque terme est réduit à une forme terminologique minimale par la racinisation, lemmatisation ou troncature.

Le processus de racinisation, ou stemming

Dans un cadre de racinisation les deux termes du tableau 3.1 seront réduits au lexème (racine) alors qu'en lemmatisation le mot entier forme une seule entité nommée référencée dans un dictionnaire. La racinisation repose sur une liste d'affixes de la langue et sur un ensemble de règles de dé-suffixation construites *a priori* (Moreau et Claveau, 2006). La base de données Postgres propose Snowball, une solution logicielle basée sur le projet de Martin Porter, inventeur du populaire algorithme de *stemming* en

anglais. Ce système est composé d'un dictionnaire de données et d'un ensemble de règles. Snowball propose maintenant des algorithmes stemming pour un grand nombre de langues, dont le français. L'algorithme sait comment réduire et normaliser les variantes d'un mot (flexions) vers sa base, ou *stem*.

Le processus de lemmatisation

Le processus de lemmatisation est une tâche complexe de réduction des termes à leur forme minimale, ou forme canonique. Un lemme peut être :

- simple : un seul mot ; par exemple : « fûtes » aura pour forme canonique, ou lemme, le verbe « être » quel que soit le contexte.
- composé : un mot composé (mot formé de plusieurs mots) ; par exemple : *peer-to-peer*
- complexe : un syntagme ou expression (groupe de mots placés dans un sens précis et s'organisant autour d'un terme central) ; par exemple : *peer reviewed paper*.

La Bibliothèque du CNAM propose un dictionnaire français de lemmatisation [1].

La lemmatisation peut intervenir, dans sa forme avancée, en contexte. Cette technique identifie la fonction grammaticale d'un mot pour en déduire son lemme. A partir du moment où la fonction grammaticale a pu être détectée, le lemmatiseur recherche dans sa base de connaissance le mot puis retourne le lemme associé à la fonction grammaticale. La solution la plus connue de lemmatisation est *TreeTagger* développée par l'Université de Stuttgart et dont les ressources linguistiques sont disponibles pour de multiples langues dont le français [2].

Le processus de troncature

Une autre méthode de flexion est la troncature, c'est à dire de simplement couper un mot pour n'en garder que le début. La méthode proposée par Enguehard consiste à ne garder de chaque terme que la sous-chaîne de caractères commençant au début du mot jusqu'à atteindre deux voyelles non consécutives (Enguehard, 1992). Cette heuristique qui permet de dé-suffixer les termes par approximation est très rapide et peu coûteuse. La troncature est complètement hors de propos dans un cadre d'indexation, mais peut trouver sa place dans une algorithmique de SRI quand la vélocité est à privilégier.

1. `http://abu.cnam.fr/DICO/mots-communs.html`, accédé le 1er août 2012
2. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`, accédé le 1er août 2012

Les analyseurs morphologiques (raciniseur et troncature) sont généralement plus faciles à mettre en œuvre que les systèmes complexes d'analyse grammaticale. De plus, il fonctionnent plus rapidement, du fait de leur simplicité. Par rapport à une analyse morpho-syntaxique (lemmatisation) des termes dans le cadre d'un processus d'analyse de requête, la précision est donc forcément réduite par analyse uniquement morphologique (racinisation ou troncature). Cependant, pour une intégration dans un moteur de recherche, quand la rapidité doit être privilégiée, la racinisation ou la troncature peuvent être préférées.

La soustraction des mots vides

En informatique appliquée à la recherche d'information documentaire, il existe des mots dits « vides » qui ne doivent ni être indexés dans le cas d'une indexation ni recherché dans le cadre d'une requête, car non discriminants. Ces mots « vides », perturbent le score de recherche en introduisant du bruit [1]. Les mots vides (*stop words* en anglais) sont alors souvent regroupés dans un « anti-dictionnaire » (*stop-list* en anglais). Ces mots sont les déterminants, prépositions, conjonctions et adverbes (Ibekwe-SanJuan, 2007).

Mots clés et critères booléens

Identifier de l'information pertinente pour l'individu dans le déluge informationnel de réponses à une requête est une tâche qui, une fois de plus, requiert une certaine pratique de la recherche d'information. La réduction des réponses ne faisant pas sens dans le contexte, ce que l'on appelle le bruit passe par l'usage de filtres par opérateurs booléens. Or, selon Spink *et al.* (2001) seules 5 % des requêtes comportaient en 2001 au moins un opérateur booléen. Cette pratique est indispensable pour encadrer sa demande informationnelle, mais elle n'est pas forcément maîtrisée par les usagers. Le terme d'algèbre booléenne, ou logique, vient du nom du mathématicien, logicien et philosophe anglais George Boole. Ce dernier publia ses travaux relatifs opérateurs binaires (Boole, 1854) au 19e siècle. L'algèbre booléenne correspond à une grammaire basée sur trois opérateurs : ET, OU, SAUF (en anglais AND, OR, NOT). Ils permettent d'interroger efficacement un outil de recherche d'information.

1. Voir section 4.1.1.

Ces termes sont communs à tous les moteurs de recherche présents sur Internet, d'où la nécessité de bien les maîtriser. Les termes ET et SAUF sont parfois représentés sous la forme de + et -. Il est à noter que dans certains moteurs de recherche, l'opérateur booléen + est implicite. C'est le cas de Google : si l'on ne mentionne rien, un opérateur AND est intercalé à chaque blanc (espace vide).

L'opérateur AND (ET en français) ou +. L'opérateur ET implique que les termes de votre recherche soient contenus dans les pages de résultat. Il faut également garder à l'esprit que les termes ne sont pas forcément contigus, ni même dans le même ordre. Prenons l'exemple d'une requête visant à récolter de la documentation sur le thème de la recherche d'information en bibliothèque :

Information **AND** *Retrieval* **AND** *Library*

Les réponses à cette requête contiendront obligatoirement les mots information, *retrieval* et *library*. Cette méthode permet, en ajoutant progressivement des termes clés, d'affiner la requête et d'obtenir des résultats moins nombreux et plus adaptés (voir section 4.1.1 le concept de bruit).

L'opérateur OR (OU en français), parfois noté « | [1] », offre la possibilité de sélectionner l'un ou l'autre des termes d'une recherche dans les résultats. Pour chercher un terme composé de plusieurs mots, il faut intercaler la séquence recherchée dans des guillemets de type anglo-saxon ou *double quote*. Reprenons notre exemple de requête sur les recherches en bibliothèque.

"information retrieval" **OR** *"information seeking"* **AND** library

Cette deuxième requête aboutit sur des réponses qui contiendront soit « *Information Retrieval* », soit « *Information Seeking* », soit les deux. Nous avons ajouté le terme

1. | se prononce *pipe*, tuyau ou tube en anglais.

« *Library* » pour limiter la recherche aux sciences de la documentation. L'usage de OU inclusif permet de faire une recherche impliquant des synonymes. Il est ainsi possible de couvrir un maximum de documents portant sur des concepts identiques (voir plus loin la notion de silence).

L'opérateur NOT (SAUF en français).

L'opérateur **SAUF** propose d'exclure un terme de la liste des réponses proposées par le système de recherche d'information. SAUF peut être noté également « - », selon les outils de recherche d'information. Si nous poursuivons notre exemple, nous pourrons l'adapter ainsi :

"Information Retrieval" **AND** Library **NOT** *"information seeking"*

Dans l'exemple, le résultat portera sur tout les éléments indexés par le système comprenant « *Information Retrieval* » et *Library*, mais pas « *information seeking* ». L'opérateur booléen NOT est ici représenté par le signe mathématique « - ». Une requête comprenant l'opérateur NOT permet de préciser une recherche (voir partie 4.1.1 page 104, le concept de pertinence) en réduisant le nombre de résultats. Il est très utile dans le cadre de la polysémie, de soustraire des termes issus du champ lexical qui ne nous intéresse pas (voir section 4.1.1 la notion de bruit). Pour illustrer notre propos, prenons l'exemple de l'*Association for Computing Machinery* (ACM). Cette association présente une conférence annuelle sur la sécurité des systèmes d'informations. Ce rassemblement, une référence dans le domaine, est sobrement baptisé *ACM conference on Computer and Communications Security*. L'acronyme de cet événement est ACM CCS. Dans notre champ d'intérêt, l'ACM propose également une taxonomie de l'informatique. Ce document est unanimement reconnu comme la référence dans le monde de l'informatique. Les domaines de la recherche et classification d'informations relatives à l'informatique scientifique l'utilisent comme base de classification. Ce système de classification est nommé ACM *Computing Classification System*, son acronyme est également ACM CCS. Si nous mettons en œuvre une recherche simple sur le terme « ACM CCS », au 12 septembre 2011, le moteur de recherche Google offre 110 000 réponses. En utilisant

l'opérateur booléen NOT le résultat est d'environ 26 300 résultats grâce au formulaire de recherche avancée de Google.

- dans un langage spécifique ;
- grâce à l'utilisation de mots-clés.

La requête peut être exprimée dans un langage de requête booléen en langue naturelle au travers de l'interface de recherche.

Les critères avancés de recherche

Des opérateurs spécifiques permettent de construire une requête en délimitant certains aspects. Ces opérateurs ont en général pour objectif de spécifier ou d'élargir une recherche pour encadrer au maximum les phénomènes de bruit et de silence [1].

La distance

Les opérateurs de distance permettent de rechercher des documents au sein desquels les termes t_1 et t_2 seront distant d'un maximum de n mots. Ainsi, plusieurs mots liés par NEAR (moteur Bing) AROUND (Google) doivent apparaître ensemble à une distance limitée (généralement, un maximum de 10 mots). Avec le moteur de recherche Google, l'operateur AROUND peut même devenir une fonction qui prend en argument une valeur numérique de n.

"information retrieval" **AROUND(10)** *"ontology"*

Google offre la possibilité de restreindre sa recherche à un nom de domaine, à un sous élément d'un domaine, voir même à un seul site web. La commande *site :* est l'opérateur de ce type de requête. Il s'agit d'une option accessible à travers le formulaire de recherche avancé de Google ou directement en mode requête.

La commande *define :* offre la possibilité de demander à Google de rechercher une définition d'un terme sur l'Internet. Les résultats retournés sont souvent des définitions

1. Voir section 4.1.1

Fonctions/Moteurs	*Google*	*Bing*
ET	*espace vide,+,AND*	*AND,&,&&*
OU	*| , OR*	*| , OR*
SAUF	*- , NOT*	*- , NOT*
PROCHE	*AROUND(n)*	*NEAR :*
SYNONYME	*~*	
LIMITER à un espace	*site :*	*site :„domain :,url :,ip :*
DÉFINIR	*define :,définir :*	*define*
Dans le TITRE	*intitle :*	*intitle :*

Tableau 3.2: Résumé des fonctionnalités des deux principaux moteurs de recherche commerciaux en 2011

issues de Wikipédia ou de dictionnaires en ligne. Il arrive également de dénicher des définitions de spécialistes sur des blogs ou des sites de recherches.

L'option Google ~ (*titlde espagnol*) permet d'ajouter un mot et ses synonymes à une requête. Cet outil est à double tranchant, car si il permet d'éviter des résultats nuls et le silence, il génère également très rapidement du bruit [1]. Il est donc préférable d'utiliser le *tilde* avec parcimonie quand un complément d'information est nécessaire suite à un silence trop important.

1. Voir section 4.1.1

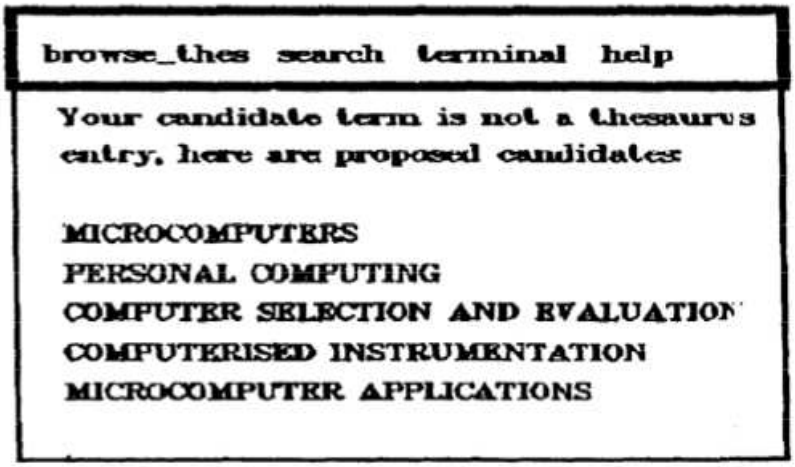


Figure 3.4: Les origines de la recherche à facettes informatisée

Recherche par facettes

Selon Boutin (2008), l'époque où l'on exprimait son besoin d'information uniquement par une requête générique décrivant la thématique générale des documents est révolue. Il s'agit aussi de caractériser son besoin d'information par des dimensions complémentaires (appelées facettes) qui ne renvoient pas seulement au contenu thématique des documents. Nous en avons identifié plusieurs et retenu cinq dans l'implémentation que nous avons proposée : le niveau de polarité d'une page web, le niveau de subjectivité d'une page web, le niveau d'accessibilité d'une page web, le niveau de lisibilité d'une page web et la centralité d'une page web dans son contexte. Chacune de ces dimensions a fait l'objet de développements théoriques et pratiques dans des domaines scientifiques d'appartenance, par exemple la linguistique computationnelle ou la psychologie cognitive. Notre objectif a consisté à aller chercher ces concepts et à étudier dans quelle mesure ils étaient transposables à l'analyse de corpus web. Le concept de recherche à facettes, ou par facettes est nommé ainsi par analogie avec un objet qui dans un monde en 3 dimensions possède de multiples facettes ou angles de visualisation. Comme nous l'explique Boutin (2008), le concept initial de facettes pour une classification bibliographique est imputé à S.R. Ranganathan Ranganathan (1963). L'idée de relier les éléments de métadonnées à l'affichage de résultats dans une RI est attribué à Belkin et Marchetti (1989). À l'époque, l'outil proposé liait un thésaurus à une interface de recherche par terme clé.

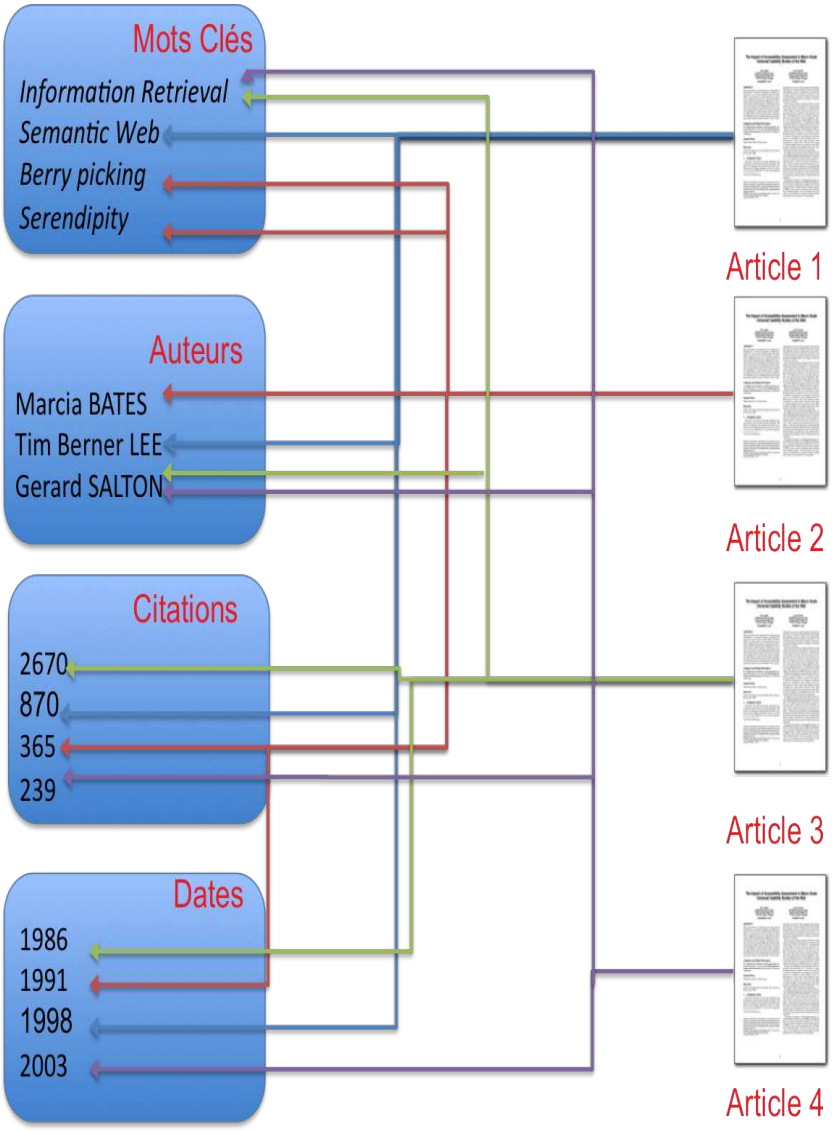


Figure 3.5: Schéma d'une recherche à facettes

Dans l'exemple proposé dans l'article, Belkin et Marchetti (1989) montraient comment rechercher les termes liés à *computer* dans le thésaurus (cf. Figure 3.4). Ce premier exemple illustre les différents aspects d'un même terme.

Ainsi la présentation à facettes permet une classification multiple pour un objet. Chaque facette correspond typiquement à la valeur possible d'une propriété commune à un ensemble d'objets. Dans la figure 3.5, nous proposons d'illustrer par l'exemple la recherche par facettes. Cette recherche est non contractuelle et dans un souci de lisibilité, nous avons limité le nombre des articles et celui des facettes. Si une recherche aboutit à l'affichage de 4 articles, il est possible dans le cas présent de les classer selon 4 critères :

- l'auteur ;
- la date ;
- les mots clés associés ;
- un élément de bibliométrie : ici le nombre de citations.

Pour chaque article, une couleur différente est utilisée pour faciliter la visualisation des facettes en deux dimensions. Il est possible de sélectionner les articles en fonction d'un aspect bien particulier issu des métadonnées. Une sélection par auteur, sans précision retournera ainsi 4 entrées classées par ordre alphabétique alors que si l'on sélectionne l'auteur Salton, seuls deux documents répondront au critère de cette facette. Le critère de sélection peut aussi être appelé contrainte. Une des plus belles réussites dans la présentation et la navigation à facettes dans un corpus multimédia est le projet Flamenco. Les facettes sont préexistantes dans la base de connaissances, il peut s'agir de présenter les résultats d'une requête par auteur, mots clés, la langue, disponibilité ou format de fichier. Un fil d'Arianne (*breadcrumb*) rappelle à l'usager les contraintes de recherches imposées aux facettes.

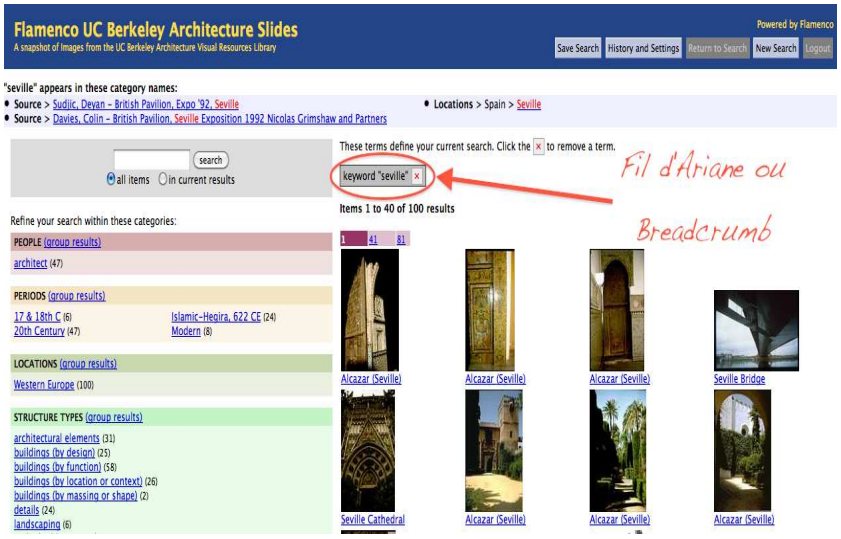


Figure 3.6: Projet Flamenco

Moteur à curseur

Pour aller plus loin dans le concept de facettes, Boutin déclare que « l'expression du besoin pourrait être affinée par l'internaute à travers l'expression de dimensions complémentaires au sujet de la recherche (Boutin, 2008) ». Son propos sort la recherche à facettes de la simple réorganisation de la présentation des résultats sur des critères purement objectifs comme ceux fournis par les métadonnées. Boutin déclare qu'un contenu de l'hyperespace est également possible à mettre en exergue sous des aspects subjectifs tels la tonalité du discours, le degré de subjectivité, son niveau de langage. Cette catégorisation supplémentaire offre à l'usager un moyen original de spécifier sa recherche pour limiter le bruit [1]. Cette méthode s'illustre le plus souvent par des curseurs qui permettent de régler avec précision la focale d'une recherche sur un ou plusieurs aspects. Boutin citait en exemple les moteurs clush.com et le mindset de Yahoo. L'initiative de Yahoo offrait un curseur horizontal axé de la publicité (*shopping*) vers l'information (*researching*). Ce curseur est à la disposition de l'usager qui peut choisir entre des documents à caractère plus ou moins commerciaux. Clush permettait de privilégier des contenus contenant du texte ou des hyperliens grâce à un curseur. Ces deux initiatives ne sont plus en ligne, nous avons donc sélectionné des sites actuels avec

1. Voir section 4.1.1

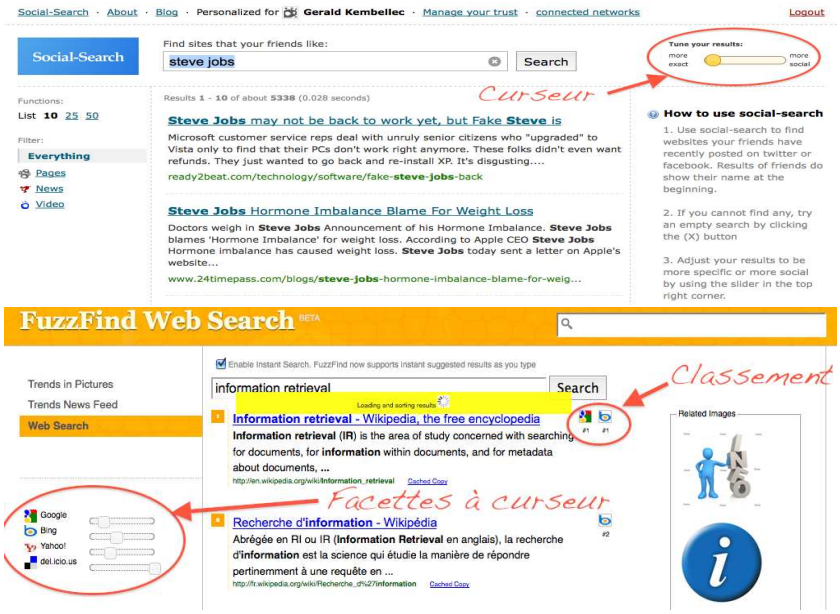


Figure 3.7: 2 exemples de moteur à curseur : Social Search et FuzzFind

des fonctionnalités toujours d'actualité. Les deux illustrations de moteurs à curseurs montrent comment atteindre un équilibre dans une recherche sous des aspects divers.

La première partie de l'illustration propose une recherche d'information avec une possibilité de choisir d'orienter sa recherche entre information traditionnelle et sociale sur les réseaux sociaux [1].

La deuxième moitié de l'illustration propose un modèle d'outil de recherche qui offre un accès aux informations de plusieurs moteurs (méta-moteur) [2]. L'originalité de cet outil est de pourvoir doser le crédit accordé à chaque moteur pour proposer un affichage personnalisé. Comme le montre la capture d'écran, chaque résultat est noté par rapport au classement qui lui est attribué sur les moteurs de recherches commerciaux. Ce classement est pondéré par rapport à la valeur attribuée à chaque moteur par le curseur. Cet outil, associé à une bonne connaissance des politiques de classement et des objectifs commerciaux des outils de recherches, offre un méta-moteur particulièrement efficace pour croiser le meilleur des algorithmes de recherche.

1. Social Search : `http://www.social-search.com`
2. Fuzz Find : `http://www.fuzzfind.com/v2/`

Moteurs	*Dir*	*Exalead*	*Google*	*MSN*	*Voilà*	*Yahoo*
Toutes positions	46,5%	34,5%	24,8%	31,2%	49,1%	21,7%
Première position	43,3%	29,7%	16,2%	29,0%	72,3%	17,9%

Tableau 3.3: Bruit généré par les moteurs de recherche, Véronis

Pour conclure sur les moteurs de recherche

Si l'on reprend l'étude française menée par Véronis (2006) pour comparer l'usage et les performances des moteurs de recherche en France, les résultats sont sans appel. Il est étonnant de croiser quelques chiffres issus de cette étude. [1] Premièrement, le fort taux de documents non pertinents retournés par le système (bruit [2]), quel que soit le moteur de recherche. La proportion de bruit généré est élevée puisqu'elle atteint pratiquement la moitié pour certains moteurs, et le cinquième pour Yahoo qui réalise la meilleure performance sur ce critère.

Deuxièmement, il est à noter le degré de satisfaction très médiocre des utilisateurs. Pour quantifier la pertinence perçue, Véronis a demandé de noter la pertinence des résultats retournés de 0 à 5, en fonction du rang occupé par la réponse lors de l'affichage (premier rang ou tous rangs confondus). Pour les meilleurs moteurs (Yahoo, Google), la note moyenne pour les dix premiers résultats affichés se hisse péniblement à 2,3 sur l'échelle suivante :

0. Entièrement insatisfait du résultat ;
1. pas satisfait du résultat ;
2. plutôt pas satisfait du résultat ;
3. généralement satisfait du résultat ;
4. satisfait du résultat ;
5. entièrement satisfait du résultat.

À l'heure actuelle, fin 2011, les principaux moteurs de recherche commerciaux sont l'incontournable Google, Yahoo et Bing (cf. Fig 3.8. [3]). MSN search est devenu Bing,

1. Pour information *Dir* est un moteur expérimental proposé par le groupe Iliad : `http://fr.dir.com/`
2. Voir section 4.1.1
3. Illustration issue du site web du spécialiste international en référencement Greenlight à l'URL : `http://www.greenlightsearch.com`

Moteurs	*Dir*	*Exalead*	*Google*	*MSN*	*Voilà*	*Yahoo*
Toutes positions	1,4	1,8	2,3	2	1,2	2,3
Première position	1,5	2,2	2,9	2,3	0,5	2,8

Tableau 3.4: Indice de pertinence perçue par moteur de recherche dans l'étude de Véronis (2006)

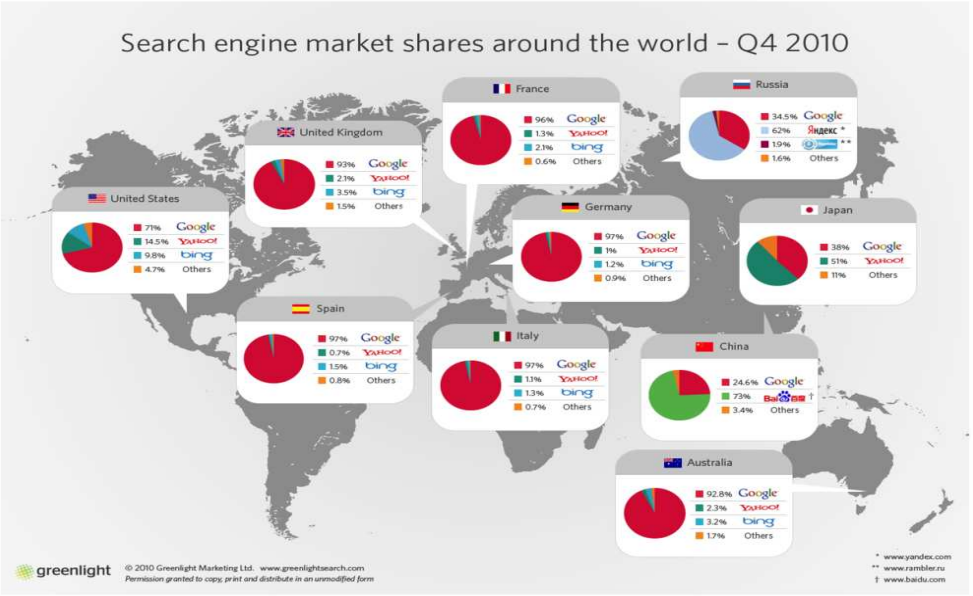


Figure 3.8: Parts de marché des moteurs de recherche commerciaux à travers le monde

mais depuis 2006, l'utilisation des moteurs de recherche n'a pas fondamentalement changé, et les statistiques d'utilisation restent comparables en terme de part de marché. Avec l'émergence de la Chine comme puissance économique, la donne va peut-être évoluer. Si l'on considère la forte implantation de moteurs locaux, il est possible d'imaginer une ouverture de ces moteurs vers l'étranger. Plus probablement, la Chine et ses centaines de millions d'usagers de l'Internet vont rejoindre massivement un moteur à forte connotation capitaliste, comme pour les biens de consommation classiques. Il faut cependant, pour cela, que l'accès à l'information s'assouplisse.

3.2.3 Les méta-moteurs

Un méta-moteur ou un méta-chercheur est une interface de recherche dont l'aspect est souvent identique à celui d'un moteur classique. La différence majeure entre un méta moteur et un moteur de recherche réside dans le contenu indexé. En effet, là où un moteur de recherche indexe des millions, voire des milliards de pages, un méta-moteur se contente d'interroger les moteurs. Le principe du méta-moteur est d'utiliser les résultats fournis par les moteurs classiques, d'en retraiter le contenu et d'en faire une présentation personnalisée, éventuellement épurée de tout aspect commercial. Pour instaurer un canal de communication avec un moteur classique, un méta-moteur utilise soit une API, soit un *wrapper* développé à cet effet.
Un *wrapper* est une « rustine » logicielle développée pour exploiter une application tierce lorsque l'on a une visibilité réduite sur son fonctionnement. Le plus souvent, une URL est « forgée » et encodée en HTML avec l'adresse d'un moteur de recherche, mais aussi les variables recherche et le contenu de la requête pour simuler une requête » manuelle »
Il est donc plus avantageux d'utiliser une , comme celles fournies par Google et Yahoo qui permettent de parser des arguments personnalisés à la requête, mais également d'avoir un flux de retour normalisé, le plus souvent en XML. De plus, cette option légalise le processus d'utilisation, ce qui s'accompagne malheureusement d'une publicité ciblée sur la requête. Lors d'une requête effectuée par un utilisateur, le processus de recherche se décompose ainsi :

1. analyse syntaxique de la requête (séparation des expressions booléennes et des termes de la recherche.
2. Liste de mots vides ou *Stop list* (l'étape de lemmatisation / racinisation est effectuée par les moteurs)

De manière concrète, le méta-moteur envoie ses requêtes à plusieurs moteurs de recherche, et retourne les résultats de chacun d'eux, en les classant, tout en éliminant les doublons. Il est souvent possible de paramétrer le méta-moteur de façon à sélectionner en amont ses moteurs favoris. De même, il est possible de choisir la manière dont les flux d'information seront traités et affichés.

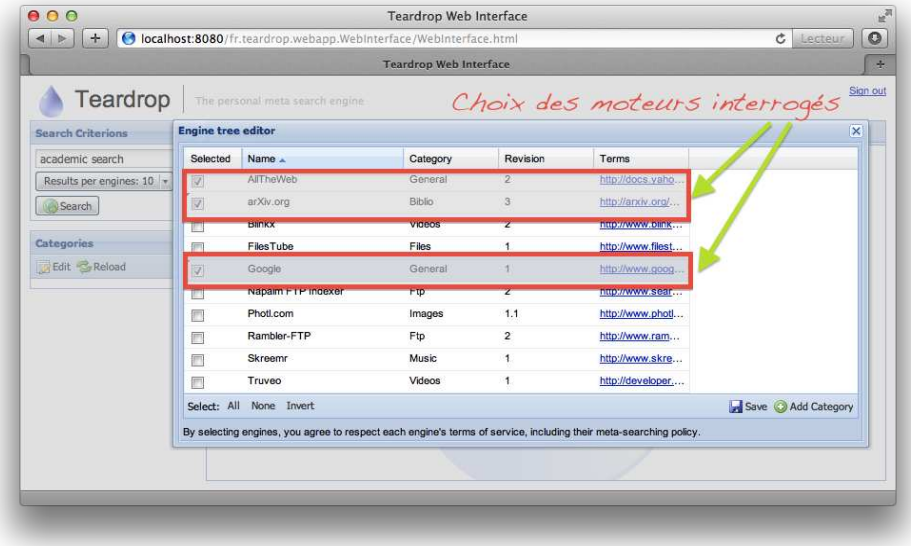


Figure 3.9: Paramétrage d'un méta-moteur, ici Teardrop

Voici quelques exemples de méta-moteurs :

1. Copernic agent, un logiciel pour Windows, technologie rachetée et utilisée par le site mamma.com [1]
2. Teardrop, un logiciel java pour toutes les plateformes (cf. figure 3.9) [2].
3. HooSeek [3], un méta-moteur solidaire (finance des associations).
4. Ixquick [4], un méta-moteur qui ne conserve pas les adresses IP des utilisateurs.
5. Seek [5], un méta-moteur francophone.
6. Seeks [6], un méta-moteur libre, sous licence GPL [7].

Le principe de méta-moteur offre un certain nombre d'avantages qui sont malheureusement contrebalancés par quelques inconvénients de tailles. Quand une requête est soumise au métamoteur, ce dernier interroge simultanément d'autres moteurs et

1. `http://www.copernic.com/fr/` et `http://www.mamma.com/`, accédés le 1er août 2012
2. `http://www.teardrop.fr/`, accédé le 1er août 2012
3. `http://www.hooseek.com/`, accédé le 1er août 2012
4. `https://www.ixquick.com/fra/`, accédé le 1er août 2012
5. `http://www.seek.fr/`, accédé le 1er août 2012
6. `http://www.seeks.fr/`, accédé le 1er août 2012
7. Gnu Public Licence : `http://www.gnu.org/licenses/agpl.html`, accédé le 1er août 2012

reformate les résultats par ordre de pertinence. Par exemple, comme les méta-moteurs ne possèdent pas leur base de connaissances propre, il n'est pas possible d'utiliser des technologies de suggestion (sauf à procéder à un enregistrement systématique des requêtes utilisateurs). Les requêtes à facettes ne sont également pas utilisables, faute de métadonnées indexées. Plus de réponses ne signifient pas plus d'informations exploitables, mais plus de données retournées. Le rappel est fortement impacté par le bruit généré en raison de l'abondance de résultats [1]. L'ordre d'affichage est le résultat d'une moyenne des classements des moteurs de recherche pour les résultats proposés. L'intérêt principal du méta-moteur est de réordonner les résultats fournis par les moteurs de recherche classiques en supprimant les doublons.

3.3 Concepts avancés de recherche d'information

3.3.1 Interfaces de références virtuelles

L'offre de références bibliographiques par le biais de l'outil informatique est un phénomène antérieur à l'avènement de l'Internet. Ce type d'offre est tout simplement l'extension électronique des offres postales ou téléphoniques. Ce service a suivi une évolution parallèle à la révolution de l'Internet des années 90. Nicolas Morin (2003) explique que le premier modèle recensé de ce type de service date du milieu des années 1980, avant même l'émergence du réseau Internet (Howard et Jankowski, 1986). Des échanges entre usagers et documentalistes avaient lieu par courriel. En 2002, des bibliothèques de toute la Floride ont mis au point un projet collaboratif de Service de Référence Virtuel, sobrement intitulé « *Ask a Librian* ». Ce projet permet de poser des questions directement à un bibliothécaire au travers d'outils tels que les formulaires ou les chats. La charte de ce projet se proposait de répondre sous 72 heures. Depuis la réussite de ce projet, de nombreuses institutions ont repris le concept, parmi lesquelles la bibliothèque du congrès ou l'Université de Cornel. Askal est devenu un logiciel à part entière, maintenu par l'université du Nebraska à Omaha [2]. Ce logiciel est utilisé, entres autres par le SCD de l'université d'Angers (voir figure 3.10 [3]). Même si le

1. Voir section 4.1.1
2. `http://library.unomaha.edu/askal` consulté le 22/12/2011
3. `http://bu.univ-angers.fr` consulté le 22/12/2011

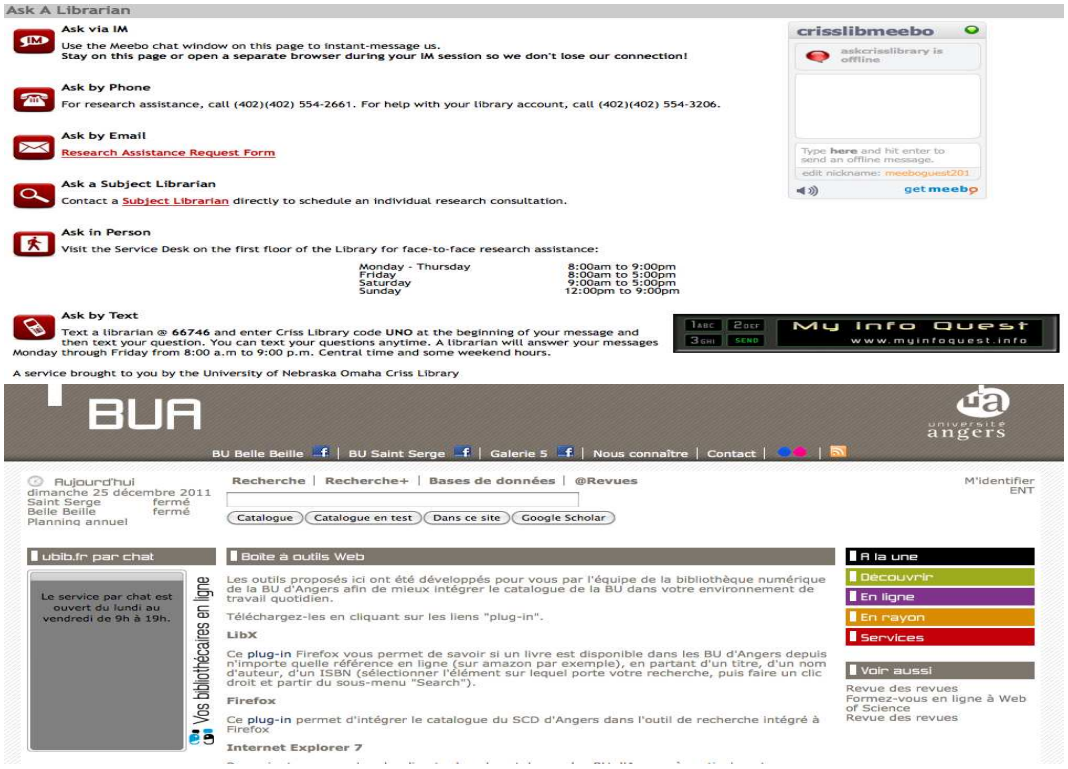


Figure 3.10: Modèle original de ASKAL et son adaptation à la BU d'Angers

projet « *Ask a Librian* » est initialement prévu pour des questions ouvertes, l'objectif principal est la constitution de bibliographies thématiques. Ainsi, des bibliographies « pointues » sont générées sur des sujets de connaissance bien définis et cadrés. Il s'agit de l'ancêtre de l'actuelle « Rue des facs [1] » ouvert début 2009 avec la charte du service d'information à la demande SI@DE [2]. Les documentalistes effectuent une curation sur les demandes et les transforment en requêtes compatibles avec les moteurs d'interrogation des bases de connaissances spécialisées. La recherche étant mutualisée sur des dizaines de SCD de grands établissements et d'universités d'Île-de-France, « Rue des Facs » redistribue les questions en fonction des spécialités de chaque établissement(Huyghe, 2010a). Des bibliographies sont générées par des spécialistes pour répondre aux besoins informationnels. Dans le cadre du projet « Le guichet du Savoir », proposé par la Bibliothèque municipale de Lyon, Calenge et di Pietro (2005) Ce service offre également des méthodes conviviales d'accès à l'information comme un nuage de mots clés (tag

1. `http://www.ruedesfacs.fr/` consulté le 22/12/11
2. SI@DE (Services d'information @ la demande) : `http://www.bnf.fr/fr/collections_et_services/poser_une_question_a_bibliothecaire/s.charte_siade.html` consulté le 22/12/11

cloud), « mur » Facebook ou flux RSS. En 2005, après un an d'existence, un premier bilan a été dressé. 4800 questions avaient été traitées et la page questions/réponses a été consultée 452 000 fois. Au 26 décembre 2011, 39 976 questions ont été traitées, ce qui montre une augmentation du traitement annuel des questions, si ce n'est de la demande. L'ENSSIB propose depuis 2009 le projet « Q ? R ! », pour « Question ? Réponse ! ». Cette interface est inspirée du projet « question point » de l'OCLC. Elle permet aux usagers de poser des questions à une équipe dédiée de spécialistes. Une modératrice réceptionne les questions, effectue des corrections orthographiques ou grammaticales et un classement des questions. La répartition effectuée, les analystes de l'équipe sont interrogés en fonction de leur spécialité. Catherine Jackson qui gère « Q ? R ! » note comme avantage certain de ce système le fait de faire émerger les besoins de formation des analystes. Les utilisateurs ont également l'occasion – unique – d'accéder à des réponses issues de la collection « papier » des documents primaires de la bibliothèque, ces collections étant généralement délaissées au profit des recherches sur le web (Jackson, 2009). La dernière offre française que nous décrirons est issue d'une collaboration entre l'université numérique Paris Ile-de-France (UNPIdF), la mairie de Paris et les SCD des universités de la région parisienne. Il s'agit de « Rue des facs », service ouvert début 2009 avec la charte SI@DE (déjà mise en œuvre par SINDBAD). Les documentalistes effectuent une curation sur les demandes et les transforment en requêtes compatibles avec les moteurs d'interrogation des bases de connaissances spécialisées. La recherche étant fédérée sur des dizaines de SCD de grands établissements et d'universités d'Île-de-France, Rue des Facs redistribue les questions en fonction des spécialités de chaque établissement (Huyghe, 2010b). Des conseils de recherche et des références bibliographiques sont proposés par des spécialistes pour répondre aux besoins informationnels spécifiques (cf. Figure [16]. Capture d'une réponse sur « Rue des Facs ») Discussion autour des services de références virtuels (SRV) Face à la pléthore de sources et de méthodes d'accès direct aux sources numériques de documentation, est-il encore utile d'introduire une étape de médiation documentaire ? Cette question est légitime, surtout si l'ensemble des usagers est autonome relativement à des outils parfois complexes, aux méthodes d'interrogation pointues. Comme le rappelait C. Nguyen dans son article sur les services de renseignements virtuels, les « sites web des bibliothèques (. . .) donnent aux étudiants un accès immédiat aux collections » (Nguyen, 2006).

SRV comme une béquille documentaire ?

Cependant, tout le monde n'a pas bénéficié d'initiation à la recherche documentaire. Jean Bouyssou, de Rue des Facs, s'est précisément posé la question de la maîtrise du panel d'outils offerts par les SCD par les étudiants. Sont-ils à l'aise avec l'interface d'interrogation et l'analyse des réponses ? Il semble que dans les usages classiques des OPAC 65 % des étudiants ne consultent que les résultats de la première page de résultats. Les 33% restant poussent leur investigation documentaire jusqu'à la deuxième page de résultats (de Saxcé, 2010). Il s'agit donc d'une recherche que l'on pourrait qualifier de « surface ». Dans ce cas, une médiation de la part d'un professionnel de la recherche documentaire est un atout majeur. Le documentaliste va interroger les bonnes sources et apprécier les meilleures références, et pas simplement les premières. Cela est d'autant plus vrai, que dans le cadre de « Rue des Facs », les documentalistes sont souvent des spécialistes du sujet traité. La documentation électronique offerte sur les sites des SCD, notamment les abonnements aux revues scientifiques, souvent anglophone, a pour public cible les doctorants et les chercheurs (de Saxcé, 2006). Les étudiants de licence et MASTER, pourtant majoritaires en université sont peu intéressés par cette documentation [ibid]. De plus, ils ne possèdent pas encore tous la connaissance des bonnes pratiques documentaires. Les services de type « questions et réponses » pourraient être particulièrement adaptés à cette population.

La dimension pédagogique et humaine des SRV

Cependant, une question posée par Agosto est la dimension pédagogique, à savoir si le documentaliste doit offrir un accompagnement dans la méthodologie de recherche ou effectuer lui-même la recherche et en offrir le fruit (Agosto *et al.*, 2011). Elle oppose deux visions du rôle du documentaliste au sein des systèmes de renseignement virtuel liés à l'enseignement. La première, dite « libérale », défend le point de vue de l'absence de responsabilité éducative dans le contexte du SRV. Le documentaliste doit se concentrer sur la recherche d'information pertinente pour répondre à la demande. L'autre vision du système, qualifiée de « conservatrice », soutient la thèse que même dans un système de SRV, la dimension éducative reste prioritaire. Sur cette question, dans un cadre universitaire, nous pensons comme Fritch et Mandernack (2001) et comme Galvin (2005) qu'au cours d'une session de référence virtuelle, le documentaliste peut profiter du

Je cherche des informations concernant la politique monétaire de Louis IX. Je crois qu'il a beaucoup fait pour une politique monétaire stable en désignant une monnaie officielle.

Merci d'avance

Question

Réponse :

Bonjour,

Vous recherchez des documents sur la politique monétaire de Louis IX.

Je vous conseille dans un premier temps d'interroger le Sudoc. Ce catalogue collectif vous permet d'effectuer des recherches bibliographiques sur les collections des bibliothèques universitaires françaises et autres établissements de l'enseignement supérieur, ainsi que sur les collections de périodiques d'environ 2400 autres centres documentaires. Il permet également de savoir quelles bibliothèques détiennent ces documents.

Pédagogie documentaire

Vous pourrez utiliser la recherche par mots du sujet avec les termes suivants :
- France louis IX finances
et pour élargir votre recherche :
- France moyen-âge finances publiques

Voici les références les plus pertinentes et par ordre alphabétique des auteurs :

- Causse, Bernard. *Eglise, finance et royauté : la floraison des décimes dans la France du Moyen Age*. Lille : ANRT, 1988

Références

- Contamine, Philippe. *Commerce, finances et société, XIe-XVIe siècles : recueil de travaux d'histoire médiévale offert à M. le Prof. Henri Dubois*. Paris : Presses de l'Université de Paris-Sorbonne, DL 1993

Figure 3.11: Capture d'une réponse sur « Rue des Facs ».

contact pour essayer de transmettre des compétences en recherche documentaire. Ainsi le choix entre offre de service et pédagogie n'a pas forcément lieu d'être. C'est ce que montre l'exemple tiré de rue des Facs (cf. 3.11. Capture d'une réponse sur « Rue des Facs »). Cette opinion est également partagée par Claire Nguyen qui déclare que dans ce cadre les médiations peuvent être « collectionner, sélectionner, (ré)orienter, et proposer les documents ; (...) informer et former » (Nguyen, 2012). Nous pensons également que la relation personnelle établie dans le cas d'une prescription au travers d'un SRV permet d'adapter la réponse de mieux profiler la sélection de références en fonction du niveau du demandant [ibid.]. On ne répondra pas de la même façon à un doctorant qu'à un étudiant en première année de licence.

Légitimité des résultats offerts par les SRV académiques

L'autre question posée par les systèmes de références virtuels est celle du rôle du documentaliste comme prescripteur. Nous pouvons nous interroger sur le bien-fondé du positionnement du documentaliste comme sélectionneur, évaluateur et prescripteur de documentation. Cette question, dans le cadre de bibliothèques spécialisées, avec des documentalistes formés dans le domaine du champ disciplinaire de l'établissement, rejoint la précédente problématique. La légitimité existe, mais la question pédagogique demeure, il faut apprendre à sélectionner les sources, un accompagnement difficile à effectuer à distance, surtout lors d'une session asynchrone (Tyckoson, 2003). La réponse à cette dernière objection peut trouver réponse dans les méthodes de navigation partagée (*co-browsing*), au cours de laquelle le documentaliste pourra effectuer une présélection des documents et expliquer ses choix (Nguyen, 2012) .

Conclusion sur les SRV

Nous avons également discuté sur l'éthique et la pédagogie associées aux systèmes de références virtuels. Si les vecteurs de communication sont multiples, le principe reste le même. Il s'agit de poser une question précise à un service de documentation, qui tentera d'y répondre dans un temps imparti, avec le plus souvent un conseil en méthodologie de recherche associé au contexte de recherche. Même si l'on peut discuter de l'intérêt ou de la validité pédagogique d'un tel service, il n'en est pas moins que les statistiques d'utilisation indiquent l'engouement du public. Par ailleurs, ces services en milieu universitaire semblent être un bon accompagnement méthodologique pour les étudiants de premiers cycles. De plus, un chercheur peut également apprécier l'aide d'une bibliographie « cousue main » par un documentaliste spécialiste d'un domaine de recherche dans lequel il s'aventure au cours d'une recherche pluridisciplinaire.

3.3.2 Introduction aux systèmes de recommandation

Traditionnellement, les individus avaient l'habitude de se recommander des produits ou services par le « bouche à oreille ». Aujourd'hui, l'offre (que ce soit d'informations ou de produits) augmente de jour en jour. Elle est proposée principalement à travers le vecteur d'Internet. Au-delà d'un seuil, plus d'informations conduisent à dégrader la qualité de l'information (Chen *et al.*, 2009). Le développement de systèmes automatisés

de recommandations (*Recommend System* ou RS) était donc un phénomène inéluctable dans l'optique de trouver des informations de qualité provenant de sources hétérogènes. Dès 2000, Burke faisait remarquer que de nombreux sites commerciaux comme Amazon ou encore e-Bay avaient compris l'intérêt (commercial) de contextualiser des offres d'hyperliens périphériques à l'hypertexte consulté par l'usager (Burke, 2000). Les moteurs de recherche commerciaux ont même créé des produits dérivés comme le « *Google AdSense* » afin d'optimiser leurs profits publicitaires en exploitant le RS. Le principe est simplement de proposer à des annonceurs privés de fournir des hyperliens vers leur site en marge des recherches des usagers. Le gros avantage de ce type de publicité est qu'elle est forcément ciblée autour des centres d'intérêt de l'usager. Les utilisateurs du portail messagerie en ligne Gmail auront forcément remarqué que les messages publicitaires en marge de leur outil est toujours en relation direct avec le contenu de leur courriel ouvert. La FAQ[1] officielle de Google :
« Notre objectif est de proposer aux utilisateurs de Gmail des annonces utiles qui correspondent à leurs centres d'intérêt (...) Le système que nous avons développé pour les annonces est similaire : en utilisant certains des critères qui permettent d'identifier les messages potentiellement importants à vos yeux, Gmail est en mesure de déterminer quelles sont les annonces susceptibles de présenter un intérêt pour vous. Par exemple, si vous avez récemment reçu un grand nombre de messages sur la photographie ou les appareils photo, il se peut que les offres d'un magasin d'appareils photo proche de chez vous et qui vous intéressent En revanche, si vous avez signalé ces messages comme étant du spam, vous ne souhaitez probablement pas les voir s'afficher[2]. »
Certains webmestres offrent contre rémunération des encarts publicitaires vides qui sont dynamiquement remplis par Google en fonction du contenu global de la page. Le but d'un système de recommandations est de réduire la surcharge d'informations(Herlocker *et al.*, 1999) en sélectionnant un sous-ensemble des éléments d'un ensemble universel basé sur les préférences des utilisateurs. Dans le domaine scientifique, nous assistons à une croissance rapide et continue des contenus des bibliothèques numériques. Cette constatation amène les utilisateurs à rechercher des outils et des services qui ne sont pas seulement adaptés à leurs besoins spécifiques de recherche, mais également à la

1. Foire aux questions, sorte de mode d'emploi sous forme de questions/réponses
2. Accédé en ligne le 17 Septembre 2011 sur la foire aux questions officielle de Google à l'URL suivante : `http://mail.google.com/support/bin/answer.py?hl=fr&ctx=mail&answer=6603`

Figure 3.12: Concept de QBE sur le site commercial Amazon.

veille technologique de leur domaine (Kapoor *et al.*, 2007).

3.3.3 Le concept de requête par l'exemple (QBE)

Pour expliciter le modèle conceptuel populaire de *Query By Exemple* (Zloof, 1977), prenons une démonstration générique avec un des portails commerciaux de type « deep web » ou Internet profond [1]. Des portails, comme celui du libraire Amazon [2], peuvent se prévaloir du concept QBE pour un domaine de connaissances. La surcharge d'information déroute l'utilisateur du système d'information (ici le consommateur). Les résultats de l'étude de Chen *et al.* (2009) indiquent que la surabondance d'informations engendre une perception de surcharge d'informations ce qui conduit les consommateurs à ne pas acheter. Accablées par ce manque à gagner, les enseignes de vente en ligne ont dû contre attaquer et mettre au point des stratégies pour accompagner et assister le client dans sa quête de produit. Le site Amazon construit ses recommandations uniquement à l'aide

1. Cette notion est explicitée dans la section 2.1.1
2. urlwww.amazon.com, accédé le 1er août 2012

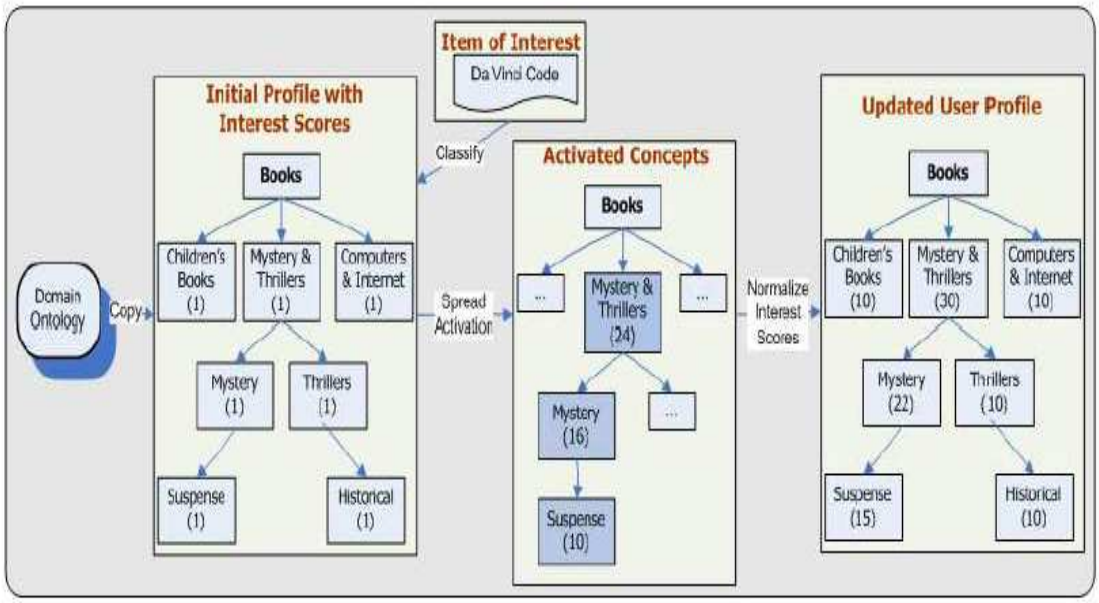


Figure 3.13: Exemple d'ontologie de la RS sur un site marchant Sieg *et al.* (2010)

des informations apprises sur ses propres clients par l'observation de leurs usages. Ainsi, sur le site d'Amazon, en sélectionnant « Livres » puis « Informatique et Internet » des requêtes sont automatiquement générées pour :

1. Parcourir une ontologie du domaine Amazon ;
2. Sélectionner le type « livre » de support de diffusion ;
3. Sélectionner la catégorie de connaissance « informatique et Internet » ;
4. Proposer les sous catégories de connaissance « Programmation, Internet, Systèmes d'exploitation, Bases de données, Multimédia et graphisme » ;
5. Mettre en exergue les termes clés les plus représentatifs de la catégorie ;
6. Calculer les catégories les plus consultées, et proposer les livres les plus populaires de ces catégories.
7. Proposer un hyperlien pour l'affichage des nouveautés de la catégorie

Ces différentes requêtes sont automatiquement générées par une navigation initiée par la proposition d'un contexte sémantique formel strict basé sur une ontologie de domaine

de la vente en ligne (cf. illustration 3.13), incluant des taxonomies des sujets proposés et thésaurisant des vocabulaires contrôlés pour chacun. Ensuite, l'utilisateur choisit un livre parmi ceux proposés. Ce choix va servir d'exemple au système et permettre de recréer une requête plus fine en partant du postulat que l'usager est intéressé par les « objets » dont les caractéristiques, dans la base relationnelle, sont proches de celles de l'objet déjà sélectionné. Cette navigation s'effectue sans que l'utilisateur n'ait eu à user d'opérateurs booléens, de mots clés et encore moins d'un quelconque langage d'interrogation de base de connaissances.

Cet exemple trivial [1] nous amène au travail plus scientifique l'équipe de Petroupolos dans le projet Clide (Petropoulos *et al.*, 2007a,b). Petropoulos adopte un contexte d'interaction visuelle dans l'optique de permettre aux utilisateurs d'amener le système à formuler des requêtes comme suite à une « navigation » graphique. Dans ce contexte, l'objet est l'interrogation d'une base de connaissances relative à l'état d'un système d'informations (ordinateurs et matériels actifs). Après le choix par navigation, l'utilisateur se voit également proposer des éléments approximativement similaires suivant leur état ou les caractéristiques techniques. Ainsi, dans l'exemple exposé dans la figure 3.14, après une sélection des tables *Com1* et *Net1* (respectivement *Computers* et *NetInterface*), l'utilisateur sélectionne le type de processeur (CPU) Pentium4 et la vitesse de transmission 54 MB. Puis l'usager coche les champs la mémoire vive (RAM), le prix et le type d'interface. Une requête est immédiatement générée pour afficher le prix et la quantité de mémoire des machines dont le processeur est un Pentium 4 et dont le débit maximal de l'interface réseau est à 54 MB. Comme indiqué en haut à droit par un indicateur vert, la requête aboutit. Selon les auteurs, la principale motivation de l'architecture Clide est de déterminer quelles requêtes donneraient des résultats par rapport à celles qui produisent des résultats vides ou une erreur. L'originalité de ce travail est de montrer les dessous du mécanisme en affichant les requêtes générées et les tables impactées. Malgré l'intérêt réel d'assister l'usager dans sa demande d'informations, il est indéniable que la représentation des tables de la base requiert une connaissance du SQL, à tout le moins de la modélisation Merise.

1. Cet exemple est une adaptation francophone du concept proposé par Sieg *et al.* (2010).

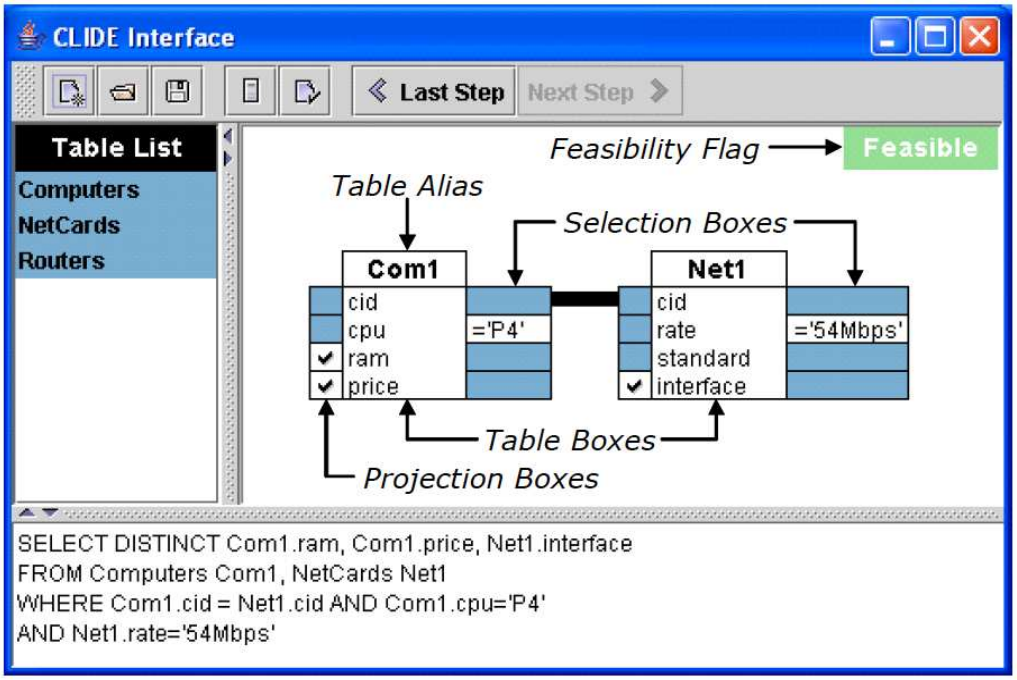


Figure 3.14: Exemple d'utilisation du système Clide.

3.3.4 Services rendus par les systèmes de recommandation

Les catalogues de bibliothèque n'exploitent pas encore systématiquement les fonctionnalités de recommandation. Pourtant, les utilisateurs semblent réclamer ce type de services, y compris pour la recherche de documents à travers un catalogue de bibliothèque. Cette étude se focalise sur 2 types de services de recommandation : « Les lecteurs de cet ouvrage ont aussi emprunté tel autre ouvrage » et « Plus de résultats comme celui-ci ».

3.3.5 Méthodologies des systèmes de recommandation

Dans cette partie nous introduirons les systèmes de recommandation d'un point de vue méthodologique. Les points abordés seront principalement la recommandation sociale , celle basée sur le contenu et les systèmes mixtes.

Votes	Utilisateur A	Utilisateur B	Utilisateur C	Utilisateur D	Utilisateur E
Objet 1	9	10	7	10	9
Objet 2	7	6	2	1	1
Objet 3	5	1	5	5	4
Objet 4	3	5	3	2	2
Objet 5	1	3	5	7	6

Tableau 3.5: Exemple d'échantillon de votes

Filtrage collaboratif

Une fois un document proposé par le système à l'usager, ce dernier se voit proposer la possibilité de lui attribuer une valeur. Cette note peut donner une appréciation intrinsèque pour le document, ou juger de l'adéquation de ce document avec le contexte de recherche. Cette note sera conservée dans le système pour être réutilisée. Selon la méthode de filtrage collaboratif dite « *Memory based* », ou heuristique, les notes peuvent aider à prédire l'appréciation d'un usager α sur un objet en se basant sur celle d'un autre utilisateur β ayant voté régulièrement de manière similaire. Pour déterminer quel sera l'utilisateur β le plus similaire à α, la corrélation de Pearson peut être utilisée (Resnick *et al.*, 1994). Cette méthode est également nommée « *Word of Mouth* » ou bouche à oreille (Shardanand et Maes, 1995) ou « *People-to-people correlation* » (Schafer *et al.*, 1999).

$$r = \frac{\sum (\alpha - |\alpha|)(\beta - |\beta|)}{\sqrt{\sum (\alpha - |\alpha|)^2 \sum (\beta - |\beta|)^2}} \tag{3.1}$$

Exemple de calcul de proximité entre usagers ayant voté pour un ensemble d'objets :

Le tableau ci-dessus donne les votes des utilisateurs pour les objets. Les corrélations calculées deux à deux donnent les résultats suivants :

Cela signifie que dans l'exemple chaque utilisateur pourra bénéficier des appréciations d'au moins un autre usager au profil similaire au sien (corrélation tendant vers 1). Une fois la base de vote utilisateur abondamment pourvue, elle peut être utilisée pour offrir une méthode de prédiction plus fine dite « *Model based* » basée sur des profils d'usagers (Breese *et al.*, 1998). Dans cette seconde méthode, les profils types sont établis à partir

Corrélation	Utilisateur A	Utilisateur B	Utilisateur C	Utilisateur D	Utilisateur E
Utilisateur A	X	0,699	0,243	0,215	0,246
Utilisateur B	0,699	X	0,265	0,341	0,413
Utilisateur C	0,243	0,265	X	0,977	0,669
Utilisateur D	0,215	0,341	0,977	X	0,996
Utilisateur E	0,246	0,413	0,669	0,996	X

Tableau 3.6: Proximité des usagers basée sur la corrélation de Pearson

de regroupement de ceux qui ont effectué des notations similaires. Ce sont les profils types ou modèles qui seront utilisés pour prodiguer des recommandations.

Avantages et inconvénients du filtrage collaboratif

Le tout premier avantage de la recommandation basée sur le filtrage collaboratif est que la connaissance du domaine de connaissance n'est pas un pré requis à la recherche d'information (Burke, 2002). Ce système permet également d'élargir la recommandation à des sujets transverses au domaine initial de connaissance en utilisant les autres centres d'intérêt des profils similaires. Cette sérendipité provoquée est appelée par Burke « *cross-genre niches* » (Burke, 2002). Selon Poirier *et al.* (2010), grâce à son indépendance vis-à-vis de la représentation des données, cette technique peut s'appliquer dans les contextes où l'analyse du contenu est difficile à automatiser. Nous rajoutons que pour des documents de type image, audio et vidéo les métadonnées ne sont pas systématiquement renseignées. Dans ce cadre, en dehors du filtrage collaboratif (ou d'un important travail de *crowdsourcing* descriptif préalable), il n'y aurait pas de méthode alternative de recommandation. Le dernier point positif est que la qualité de la recommandation proposée par filtrage collaboratif croit avec l'utilisation du système.

Claypool et al ont pointé un certain nombre de problèmes issus des méthodes initiales de recommandation (Claypool *et al.*, 1999). Par exemple, à l'état initial, le système de recommandation basé sur le filtrage collaboratif est inutilisable pour cause de « *cold start* ». Ce problème de démarrage à froid s'exprime de la manière suivante : sans note, pas de recommandation possible. Cette difficulté est reproduite lors de l'ajout d'items ou d'usagers. Avec un nombre trop faible d'évaluations pour un corpus vaste,

les données sont trop éparses pour établir des corrélations suffisantes. Ce phénomène est appelé « *sparsity* », ou éparpillement (Claypool *et al.*, 1999). Il arrive, dans le cadre d'une tentative de classification des individus, qu'un élément soit à la frontière de plusieurs groupes. Par exemple, dans un système de mise à disposition et de notation de littérature scientifique, il peut arriver qu'un usager α soit autant intéressé par la science $S1$ que par la science $S2$. Malheureusement, les recommandations de sur les documents relatifs à ces deux sciences sont notés par des individus appartenant à deux groupes distincts. Comme α vote de manière atypique, il ne se verra intégré dans aucun des deux groupes. Ce phénomène connu sous le nom de « *grey sheep* », ou mouton gris (Burke, 2002, Claypool *et al.*, 1999). Si d'autres personnes adoptent son comportement, ils formeront un groupe à part qui produira de la recommandation pour la nouvelle « *cross-genre niche* ». Il est également indéniable que le principe de popularité sera privilégié par le filtrage collaboratif. Plus un objet sera noté positivement, plus il sera recommandé et donc sera de nouveau évalué. Ce principe de notoriété auto engendrée sera peut-être plus dû à l'ancienneté qu'à la qualité réellement perçue par les usagers. Ce problème peut être contrebalancé ou au contraire intensifié par une faille du système de recommandation sociale : la fraude au vote avec des identités multiples. Il peut être tentant de modifier les recommandations dans une optique marchande en votant sous plusieurs identités. Cette technique est appelée « *shilling* » et fait l'objet de nombreuses études (Lam et Riedl, 2004, Williams *et al.*, 2006).

L'indexation et la catégorisation

L'autre méthode traditionnelle de filtrage est basée sur la description et l'analyse des contenus proposés par le système. Ce procédé est principalement basé sur des techniques d'analyse textuelle, mais peut être étendu à des contenus divers contenant des métadonnées. Les photos illustrent le propos, composées d'un contenu binaire, elles offrent la possibilité de contenir des métadonnées auto générées comme les coordonnées géographiques, l'exposition, l'orientation ou la date grâce au format EXIF (*Exchangeable Image File Format*). La technique de recommandation sur la base du contenu se fonde sur la relation entre le profil de l'usager et les métadonnées associées aux objets stockés dans la base de connaissances (Boutell et Luo, 2004, Lee *et al.*, 2006) . L'utilisateur peut entrer volontairement ses préférences lors de son inscription au service, elles seront

dites « fournies ». L'autre possibilité est de calculer les préférences par l'observation de son comportement Adomavicius et Tuzhilin (2005), dans ce cas elles seront « calculées » et vectorisées. Les préférences de l'utilisateur sont représentées sous forme d'un vecteur contenant les termes les plus représentatifs des goûts de l'usager. Ces termes clés peuvent avoir une valeur déterminée statistiquement en fonction de leur fréquence dans les documents consultés et/ou notés par l'usager au sein du corpus (Balabanović et Shoham, 1997). Par exemple, il est possible d'utiliser l'algorithme $tf.idf$ pour pondérer termes clés issus de textes (Salton et Waldstein, 1978).

$$tf_{(m,d)} = \frac{n}{card(d)} \tag{3.2}$$

Exemple de calcul de *term frequency*

Considérons un document d contenant 100 mots dans lequel le terme m apparaît n fois avec $n = 3$. La fréquence du terme (tf) pour m au sein du document d est alors le quotient entre le nombre d'occurrences de n du mot m dans le document d et le nombre total de mots dans d. Ce qui appliqué à l'exemple donne $\frac{3}{100}$.

L'inverse de la fréquence de documents (Jones, 1972) est calculée ainsi par le logarithme du quotient entre le cardinal de l'ensemble du corpus C et le cardinal du sous corpus C' des documents de C qui contiennent le terme m. Nous ajoutons 1 au dénominateur pour généraliser la fonction au cas de l'absence du terme dans le corpus.

$$idf_m = log(\frac{card(C)}{1 + C'_{m,C}}) \tag{3.3}$$

Exemple de calcul de *inverse definition frequency*

Maintenant, supposons que nous avons 10 millions de documents dans le corpus C et que le terme m apparaît dans un millier de ceux-ci. Appliqué à notre exemple le résultat de *idf* est $log(10000000/1000)$ soit 4.

Finalement, le poids pondéré d'un terme dans un document par rapport à un corpus s'obtient en multipliant les deux mesures tf et *idf*

$$tf.idf_{m(C',C)} = \frac{n}{card(d)}.log(\frac{card(C)}{1 + C'_{m,C}}) \tag{3.4}$$

Exemple de calcul de *term frequency . inverse definition frequency*

La valeur $tf.idf$ dans notre exemple précédent est le produit de ces quantités : 0,03 x 4 = 0,12. Ainsi le terme m sera statistiquement pondéré avec un coefficient de 0,12 dans le document d du corpus C.

Cet algorithme basique est rarement utilisé seul, remplacé par des générations plus récentes et sophistiquées de combinaisons, comme Terrier (Ounis *et al.*, 2005), avec notamment okapi BM25, mais reste le fondement de la pondération de termes représentatifs de documents dans des corpus textuels. Les méthodes basées sur la vectorisation de requêtes montrent des résultats prometteurs. Berry et al suggèrent la récupération de la requête sous forme matricielle par l'algorithme populaire LSI ou indexation sémantique latente. Dans essence, l'algorithme crée un espace vectoriel de dimensions réduites qui offre une représentation à n dimensions d'un ensemble de documents (Dumais *et al.*, 1988). Quand une requête est entrée, sa représentation numérique est comparée avec les cosinus d'autres documents de la base, et l'algorithme retourne les documents pour lesquels la distance est la plus faible. Cette méthode peut être adaptée pour recommander des documents en fonction des besoins des usagers. Dans le cas de données non textuelles, il est cependant possible de mettre en exergue et d'évaluer statistiquement les centres d'intérêt de l'usager. Il est ainsi envisageable de faire des évaluations statistiques sur les coordonnées géographiques des photos préférées ainsi que sur leur orientation (portrait ou paysage) ou encore type d'équipement utilisé. Ces données seront à croiser avec les préférences utilisateurs, qu'elles soient renseignées par l'utilisateur ou calculées sur les statistiques d'utilisation de l'usager.

Avantages et inconvénients du filtrage sur le contenu

Les avantages du filtrage sur le contenu sont similaires à ceux observés par le filtrage collaboratif (Burke, 2000). Ainsi, la connaissance du domaine n'est pas obligatoire pour l'usager, car les recommandations seront issues des données du corpus. La finesse des recommandations système évoluera également avec la taille du corpus.

Cependant, le système basé sur les seules données du corpus ne pourra pas proposer de « sérendipité » faute de corrélation avec les usagers. De plus, comme le fait remarquer Poirier, chaque utilisateur est absolument indépendant des autres. Ainsi, un usager qui aura correctement rempli son profil avec ses thématiques de prédilection recevra des propositions même s'il est le seul inscrit (Poirier *et al.*, 2010).

L'inconvénient majeur d'un moteur de recommandation orienté données sera, dans un premier temps, comme pour le type collaboratif le problème posé par le nouvel utilisateur qui n'a pas encore de profil établi et donc pas de données de référence « observées ». Ensuite, il est évidemment plus complexe d'indexer des données non textuelles. De plus, l'usager sera « sclérosé » dans un contexte de recherche, celui qu'il a déjà positionné comme étant son centre d'intérêt. Ce problème est identifié comme étant l' « *overspecialization* », ce qui annihile toute possibilité de sérendipité par proposition de sujets connexes.

Les méthodes hybrides de recommandation

De manière triviale, l'hybridation de systèmes de recommandation résulte de la combinaison de méthode de filtrage collaboratif avec celle basée sur le contenu. Cette vision de l'hybridation a été affinée par Burke puis par Adomavicius et Tuzhilin (Adomavicius et Tuzhilin, 2005, Burke, 2002).

Burke recense les sept techniques d'hybridation suivantes (Burke, 2002) :

1. *Weighted* / Pondération : La valeur de recommandation d'un item est issue de la somme des méthodes disponibles. Par exemple *P-Tango* Claypool *et al.* (1999) donne une valeur égale au filtrage collaboratif et à aux prédictions basées sur le contenu. Cette valeur est ensuite pondérée par une confirmation des usagers.
2. *Switching* / Bascule : Le système choisit d'appliquer soit une méthode orientée données, soit un filtrage social selon le contexte de recherche de l'utilisateur.
3. *Mixed* / Mixte : Cette technologie permet de proposer des recommandations provenant des méthodes traditionnelles dans l'optique de limiter les inconvénients de chaque méthode classique.

4. *Features combination* / Combinaison : Cette méthode offre la possibilité d'enrichir les données intégrées a priori dans le système avec les votes des utilisateurs, qui enrichissent la base a posteriori. Le calcul de recommandation se fait sur l'ensemble des données.
5. *Cascade* / Cascade : Ce procédé consiste à une double analyse des profils utilisateurs. La première passe sert à faire émerger des candidats, Le deuxième à affiner la sélection d'utilisateurs.
6. *Features augmentation* / augmentation : Il s'agit d'une technique similaire à la précédente pour le premier passage. Si le nombre de candidats est trop élevé au premier passage, alors un deuxième fera une discrimination supplémentaire en intégrant les données des objets recommandés.
7. *Meta level* / niveau modèle : Comme pour les deux dernières méthodes, il s'agit de passer filtrer deux fois les usagers pour déterminer des similarités. La différence est que le premier passage permet de générer un modèle ou profil type d'utilisateur.

Adomavicius et Tuzhilin proposent une classification des méthodes hybrides de recommandation reposant sur quatre axes (Adomavicius et Tuzhilin, 2005) :

1. *Combining separate recommenders* / Combinaison des résultats séparés : La méthode collaborative et la méthode basée sur le contenu sont appliquées séparément, puis leurs prédictions sont combinées.
2. *Adding content-based characteristics to collaborative models* / Ajout de résultats issus du contenu aux prédictions collaboratives : Ce système utilise l'approche collaborative traditionnelle entre individus « *People-to-people correlation* », à laquelle il ajoute des recommandations basées sur la classification du contenu et des goûts renseignés par les usagers.
3. *Adding collaborative characteristics to content-based models* / Ajout de prédictions collaboratives aux groupes d'intérêt issus du contenu : Le principe de ce modèle n'est pas d'inverser par rapport au précédent, mais d'incorporer quelques caractéristiques de la méthode collaborative par profil de groupe « *Model based* » dans l'approche à base de contenu.

4. *Single unifying recommendation model* / Modèle de recommandation unifié : Construction d'un modèle général qui incorpore les caractéristiques des deux modèles au sein d'un même algorithme.

Conclusion sur les modèles de recommandation historiques

Les deux premiers types de modèles de recommandation se chevauchent sur un axe historique dans les années 90. Cette première partie a présenté les méthodes et algorithmes associés à la base des systèmes de recommandation, à savoir les systèmes basés sur le filtrage collaboratif et ceux issus d'un traitement par catégorisation du contenu. Nous avons exposé que les systèmes de recommandation par filtrage collaboratif sont issus d'un traitement statistique sur l'opinion exprimée des usagers. Il est apparu que les méthodes basées sur les données sont adaptées des règles de traitement automatique du langage, notamment de l'indexation automatisée et de la pondération de termes représentatifs. Pour pallier aux faiblesses inhérentes à ces modèles initiaux, dès la fin des années 90, des méthodes hybrides sont apparues.

Les folksonomies

Les systèmes d'annotation sociale permettent aux utilisateurs d'annoter des ressources avec des étiquettes personnalisées. Ces étiquettes ou *tags* servent à naviguer des espaces d'informations vastes et complexes, sans la nécessité de s'appuyer sur des hiérarchies prédéfinies comme les taxonomies. Ces systèmes permettent aux utilisateurs d'organiser et de partager leurs ressources propres, ainsi que d'en découvrir de nouvelles annotées par d'autres utilisateurs. Des recommandeurs de *tags* dans de tels systèmes permettent d'aider les utilisateurs à trouver des balises appropriées pour les ressources qu'ils déposent, référencent ou consultent. Cela contribue à la consolidation de l'ensemble du système d'annotation et bénéficie à tous les utilisateurs et à toutes les ressources (Gemmell *et al.*, 2010). Ce système participatif par filtrages collaboratifs permet de mettre en place à moindre coût des systèmes de recommandation efficace sur les espaces de partage de signets génériques comme *del.icio.us* (cf. Figure 3.15) ou scientifiques avec *citeulike*. Dans les deux cas, les utilisateurs proposent des liens sur vers des documents présentant un intérêt à leurs yeux et ajoutent des mots clés. Un même document est souvent proposé et annoté par plusieurs utilisateurs.

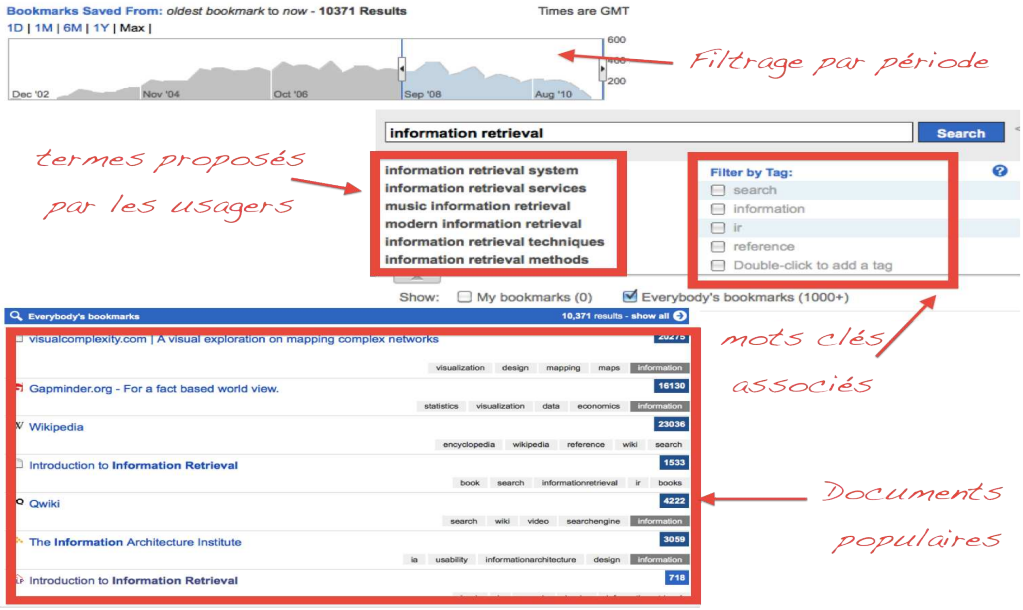


Figure 3.15: Système de recommandation folksonomique : del.icio.us

Les statistiques globales d'utilisation

En appliquant des techniques et méthodes du domaine de l'aide à la décision, une nouvelle approche pour pallier l'absence de l'utilisateur du système d'interaction dans les systèmes de recommandations existantes est proposée.

3.3.6 Conclusion

Les outils de recherche existants sont génériques, pragmatiques et offrent un accès à l'information acceptable pour des sujets basiques. Les moteurs (et méta-moteurs) de recherche nous relient à des milliards de documents qui peut être consultés rapidement grâce des mots clé. La recherche par mot clé plein constitue le point de départ pour la majorité des utilisateurs. Cependant, si cette stratégie fonctionne bien pour une minorité de requêtes, l'utilisateur type est souvent confronté soit avec une liste de résultats vide ou avec une liste contenant des milliers, voire des millions, de réponses possibles (Eissen et Stein, 2002). Il devient donc évident que la connaissance de l'usage des outils et méthodologies basiques de recherche d'information forment un pré-requis nécessaire, mais non suffisant à l'activité de collecte d'informations pertinentes, particulièrement pour le domaine de la science.

Chapitre 4

Qualité de l'information

Le langage secret de la statistique, si attractif dans une société qui vit beaucoup de faits et de chiffres, peut être employé pour faire du sensationnel, pour gonfler les résultats ou pour simplifier à l'extrême..

Y-F. Le Coadic

Introduction

Pour l'usage commun, une recherche d'informations est synonyme de requête passée sur le moteur commercial Google. Quelques mots clés sont proposés en entrée pour un résultat souvent supérieur au million de réponses. L'utilisateur se heurte alors au dilemme :

- Choisir au hasard quelques réponses parmi celles proposées dans la première page ;
- effectuer une fouille poussée par analyse systématique des centaines de pages de réponses jusqu'à avoir trouvé une réponse pertinente. Nous allons, pour traiter de la pertinence de l'information, définir quelques termes.

Résultat *ad hoc* ou *ad hoc retrieval* (Manning *et al.*, 2008)

. La locution latine *ad hoc* signifie « pour cela », employé ici comme adjectif peut être compris comme qui « en réponse à un besoin spécifique ».

Dans notre contexte de RI, trouver une réponse *ad hoc* à notre besoin d'information [1] signifie que la recherche produit un document qui soit en adéquation complète avec notre requête dans le contexte précis et arbitraire qui est le notre.

Pertinence

Le terme de pertinence, en français peut être traduit dans la littérature anglaise « *relevance* », ou « *aboutness* » qui désigne l'à propos ou l'adéquation. Cooper (1971) a fourni une définition de la pertinence logique comme une base formelle, donc un critère objectif, pour l'évaluation des systèmes de recherche. La pertinence de son point de vue est qualitative, donc intrinsèque.

Wilson (1973), bien que s'appuyant sur les travaux de Cooper pour le côté qualitatif d'un document, contextualise l'intérêt d'un document. Pour lui, la pertinence d'un document dépend en premier lieu du problème particulier étudié.

4.1 RI et qualité de l'information

Pour aller plus loin dans le domaine de la contextualisation de l'information pertinente, Saracevic (1975) déclarait qu'une information est pertinente (ou non), que par rapport à quelque chose, ici un sujet de recherche et dans un contexte bien particulier.

C'est dans cette optique que Pia Borlund (2003) écrivait cette phrase, qui de notre point de vue synthétise la notion de contexte et de l'individualité de la pertinence : « *Pertinence represents the intellectual relation between the intrinsic human information need and the information objects as currently interpreted or perceived by the cognitive state of an assessor or user.* [2] ». Elle déclare également que même si un document est vraisemblablement pertinent relativement à sujet donné, cette pertinence est tout de même interprétée par des individus différents dans des contextes différents. Un travail,

1. Nous étudions plus longuement le besoin d'information dans le chapitre5.1.

2. Proposition de traduction : La pertinence représente la relation intellectuelle entre le besoin d'informations humaine intrinsèque et les objets d'information tel qu'il est interprété ou perçu par l'état cognitif de l'utilisateur.

quelle que soit l'excellence de sa facture, ne sera, du point de vue du lecteur, pertinent que dans un cadre bien particulier.

Nous nous rangeons à l'avis de Borlund. Nous ne mesurerons donc la pertinence d'un document que dans un contexte et de notre point de vue, par uniquement par rapport au sujet étudié. Nous gardons cependant à l'esprit qu'il existe des facteurs indéniables de qualité intrinsèque, qui peuvent influer sur le choix d'un document ou son rejet (voir partie 1.2). D'un point de vue pratique, pour mesurer la pertinence les critères principaux sont le rappel et la précision (voir 4.2 et 4.3). Comme indiqué dans les paragraphes suivants 5, du point de vue d'un demandeur d'information, la pertinence peut être considérée comme une décision de sélection dans le processus de recherche d'information. Pour mener à bien une recherche qualitativement satisfaisante et pertinente, un utilisateur aimera connaître la « qualité » des documents référencés dans le SRI. Cet usager aimera également connaître les statistiques clés sur les résultats retournés par système pour une requête :

- L'efficacité
  Une recherche est efficace si l'utilisateur perçoit les documents retournés comme des informations de valeur par rapport à son besoin d'informations personnel.
- Précision et exactitude
  La précision répond à une question simple : Quel pourcentage des résultats renvoyés est considéré par l'utilisateur comme exploitable ? Les documents exploitables sont ceux dont le contenu est exact et démontré, mais également pertinent dans le cadre défini par le besoin d'informations.
- Rappel
  Quelle fraction des documents pertinents dans la collection parcourue ont été retournés par le système ?
- La première option, le plus souvent, choisie n'est efficace que si les mots clés choisis pour la requête sont suffisamment représentatifs du domaine du champ sémantique du concept recherché. De plus, le thème de recherche ne doit pas présenter d'ambiguïté pour ne pas renvoyer trop de réponses sans intérêt.
- La deuxième option est sujette au hasard Il est tout à fait possible de trouver une ou plusieurs réponses pertinentes rapidement, mais l'utilisateur est le plus souvent « submergé » sous la masse d'informations [1] non pertinentes et cette méthode plus

1. La littérature anglo saxonne utilise le terme d'*information overload(Maes, 1994)*

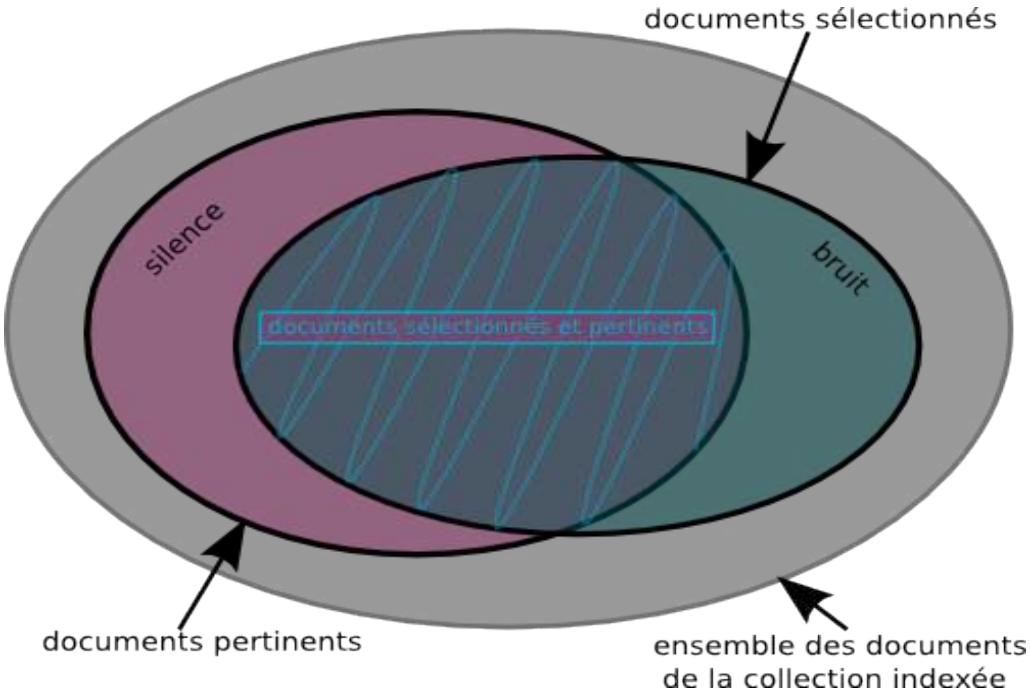


Figure 4.1: Résultat de recherche d'informations documentaire

précise est chronophage. Pierre Lévy (1998) utilisait une expression très imagée de « déluge informationnel » pour qualifier la surcharge informationnelle.

Nous proposons de représenter la recherche et résultat d'une requête, dans le cadre de la RI, de la manière suivante :

Soit un ensemble N de documents (le corpus), A est un sous-ensemble de N qui comporte les réponses pertinentes retournées par la requête x d'un usager, B est le sous-ensemble de N constitué des réponses obtenues. On note C l'ensemble théorique de toutes les réponses pertinentes du corpus.

$$\forall x : \exists (A , B , N) \, tq. \, A \subset B \subset N \tag{4.1}$$

4.1.1 Bruit et taux de rappel

De manière triviale la RI est confrontée à un problème de pertinence qui se mesure sur deux axes.

Plus le rappel tend vers 1, plus le silence est faible pour une recherche fixée.

$$rappel = \frac{Card(A)}{Card(C)} \leqslant 1 \tag{4.2}$$

Le premier axe de cette problématique est lié au volume de données, le second à la visibilité de ces dernières. Le diagramme de VENN (fig. 4.1 page 104) illustre ce propos en montrant que pour une recherche, le résultat (en bleu) sera la jonction des documents pertinents (en orange) trouvés et de « faux positifs » détectés par le Système de RI (SRI) comme pertinents mais ne l'étant pas (en gris). Les « faux positifs » sont appelés bruit et polluent la recherche, noyant les documents pertinents dans un flot d'information compact.

4.1.2 Silence et taux de précision

Le silence est évoqué lorsque des réponses pertinentes ne sont pas proposées par le système d'interrogation de la base de données, alors qu'elles existent. Cela peut arriver notamment avec les catalogues de bibliothèque. Les causes du silence peuvent avoir des causes de divers ordres :

- L'utilisateur formule une requête comprenant trop de termes coordonnés par l'opérateur ET ;
- L'indexation de la base est insuffisante ;
- L'indexation de la base suit un langage rigide et compliqué que l'utilisateur ne connaît pas (exemple : indexation et recherche à partir seulement d'un thésaurus).

Nous employons le terme de bruit lorsque des réponses non pertinentes sont proposées par le système d'interrogation de la base de données. Ces réponses sont mêlées à des réponses pertinentes mais ces dernières risquent de ne pas être vues par l'utilisateur. Cela peut arriver également avec les catalogues de bibliothèque. Soit un ensemble N de documents, A est un sous-ensemble de N qui comporte les réponses pertinentes pour la recherche d'un usager, B est le sous-ensemble de N constitué des réponses obtenues Plus A est inférieur à B, plus le bruit est grand et le taux de précision est faible.

$$précision = \frac{Card(A)}{Card(B)} \leqslant 1 \tag{4.3}$$

Mesure de l'efficacité d'un système d'indexation et de recherche établie à partir du ratio entre le nombre de documents pertinents trouvés lors d'une recherche documentaire et

le nombre total de documents trouvés en réponse à la question. C'est un indicateur de mesure du bruit.

Les causes du bruit peuvent être multiples :

- l'utilisateur n'utilise pas assez de termes dans sa requête ;
- l'utilisateur formule une requête comprenant trop de termes coordonnés par un OU logique (inclusif) d'algèbre booléenne, signifiant « vérifiant l'une ou l'autre option ou les deux ».

Cet indice mesure l'efficacité d'un système d'indexation et de recherche. Il est établi à partir du ratio entre le nombre de documents pertinents trouvés lors d'une recherche documentaire et le nombre total de documents pertinents existant dans le système. C'est un indicateur de mesure du silence. Un SRI idéal serait à même d'échapper au bruit et d'indexer l'intégralité de l'Internet pour éviter le silence.

De l'autre côté, à cause de la notion de visibilité (voir la partie 2.1.1 page 37) sur Internet, un grand nombre d'informations pertinentes échappent au SRI et ne sont donc pas intégrées au résultat.

4.1.3 F-measure

introduction

Pour obtenir l'indice de qualité globale d'un résultat de recherche, on introduit la notion de F-measure (ou F-score). La F-Mesure se calcule en effectuant une combinaison de la précision et du rappel (Van Rijsbergen et Lalmas, 1996).

Méthode simple de calcul de la F-measure

La première méthode est celle dite « par moyenne harmonique », qui consiste à utiliser l'harmonique arithmétique entre la précision et le rappel pour calculer le F-score.

$$F_1 = \frac{2.precision.rappel}{precision + rappel} \tag{4.4}$$

Pour illustrer ce propos, posons :
Une requête effectuée sur un moteur de recherche retourne 100 résultats, dont 60 sont jugés pertinents. Nous savons que 200 documents pertinents n'ont pas été retournés. La

précision p sera alors de $\frac{6}{10}$ soit $0,6$et le rappel r de $\frac{6}{26}$, soit $0,24$.
L'harmonique de la précision et du rappel donnera une F-measure de :

$$F_1 = \frac{2.p.r}{p+r} = \frac{2.\frac{6}{10}.\frac{6}{26}}{\frac{6}{10}+\frac{6}{26}} \simeq 0,4 \tag{4.5}$$

Méthode avancée de calcul de la F-measure

On note usuellement la F-measure lissée (ou pondérée) F_β, avec β un réel positif, on la définit de la manière suivante :

$$F_\beta = \frac{(1+\beta^2)(précision.rappel)}{\beta^2.précision + rappel} = \frac{(1+\beta^2)(2u+b+s)}{\beta^2 s + (1+\beta^2)u + b} \tag{4.6}$$

note : Le choix du β dans la formule correspond à l'importance donnée à la précision par rapport au rappel, ce qui donne lieu à trois cas de figure.

1. Concrètement, on utilisera une valeur de β élevée si l'on désire privilégier le rappel sur la précision.
2. Inversement, une valeur proche ou égale à 0 mettra en avant la précision.
3. Une valeur dite *équilibrée* ou *neutre* sera voisine de 1 et ne mettra en avant aucune des deux.

Le résultat de la F-measure sera très variable pour une même recherche en fonction du β. Il s'agit donc d'un indice subjectif, sujet à de grandes variations en fonction de l'objectif.
Exemple :
Dans le cadre d'une recherche en biologie sur les membranes cellulaires, le moteur *Google Scholar* donne, en français, pour le terme « membrane », en généralisant la proportion des résultats donnés pour la première page le résultat suivant :

- 20 % de documents n'ayant pas de rapport avec le thème recherché. Il s'agit donc du *bruit*. La totalité des résultats étant de 45 400, le bruit estimé est d'environ 9 000.

- Pour l'exemple, nous supposons que l'indexation documentaire de Google Scholar n'est pas exhaustive. Les résultats non retournés forment le *silence*. On pose leur nombre à 50 000 de manière arbitraire.
- Dans cet exemple, le nombre documents pertinents sélectionnés par la recherche, par rapport au nombre total d'articles disponibles pertinents donne le rappel et sera de :

$$\frac{Card(A)}{Card(C)} = \frac{0,8 \times 45400}{0,8 \times 45400 + 50000} = 0,42 \tag{4.7}$$

  soit un résultat de 42 %, ce calcul a été effectué avec la formule 4.2
- La *précision* sera calculée sur la base du pourcentage de résultat retourné par le moteur par rapport au nombre total de résultats retourné. Reprenons la formule 4.3

$$\textit{précision} = \frac{Card(A)}{Card(B)} = \frac{0,8 \times 45400}{0,8 \times 45400 + 0,2 \times 45400} = 0,8 \tag{4.8}$$

  soit une précision de 80%.
- La F-measure sera calculée pour 3 valeurs de β différentes selon les priorités exprimées pour la recherche à l'aide de la formule 4.4.

  1. $\beta = 0,3$ exprime une forte précision
     $F_\beta = \frac{(1+\beta^2)(precision.rappel)}{\beta^2.precision+rappel} = \frac{(1+0,3^2)(0,8\times0,42)}{0,3^2\times0,8+0,42} = 0,74$
  2. $\beta = 1$ exprime l'équilibre entre neutralité et le rappel
     $\frac{(1+1^2)(0,8\times0,42)}{1^2\times0,8+0,42} = 0,55$
  3. $\beta = 5$ exprime un taux de rappel élevé.
     $\frac{(1+5^2)(0,8\times0,42)}{5^2\times0,8+0,42} = 0,43$

  En conclusion, lors de l'usage d'une F-mesure, la qualité du système va fortement dépendre de la prise en compte des objectifs, à savoir l'exhaustivité ou la finesse du résultat. Nous pourrons ultérieurement utiliser ces indices qualitatifs pour valider un outil de recherche d'informations dans le contexte d'un corpus pré-indexé manuellement.

4.2 Le PageRank

Le PageRank est un indice de popularité pour une page web, calculé selon un algorithme très sophistiqué, utilisé par Google. Initialement, l'algorithme « *Method for Node Ranking in a Linked Database* » fut développé par l'université de Stanford en 1997 et déposé par Page (2001) en 1998 pour le compte de Standford [1] (renouvelé en 2001) qui en a concédé la licence à Google jusqu'en 2011. Cependant, le nom « *PageRank* »– littéralement rang de la page [2] – est une marque déposée par la société Google. Le PageRank fait partie des critères utilisés pour déterminer le positionnement d'une page dans l'affichage des résultats des requêtes sur Google. Il est présenté comme un gage de qualité basé sur la popularité. Examinons les bases de son fonctionnement.

Selon le collectif italien Ippolita (2008), le fonctionnement du PageRank repose sur la popularité d'une page web. Cette popularité est calculée à partir du nombre de sites qui ont au moins un lien pointant vers elle. À égalité de liens, deux pages web auront des PageRank différents selon l'importance de celles à qui elles sont reliées.

Partons du principe que le web est comparable à un graphe (Eisermann, 2009). Chaque document hypertexte est un nœud de ce graphe et les hyperliens forment des arcs valués entre ces nœuds. La valeur PageRank notée PR d'une page u dépend de la somme des valeurs des PageRank de chaque page v de l'ensemble B_u. Ce dernier contient toutes les pages ayant au moins un lien vers la page u, divisé par le nombre $L(v)$ de liens vers la page u depuis B_u (soit le cardinal de B_u).

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \tag{4.9}$$

Exemple : Nous souhaitons calculer le PageRank d'une page A, dans un micro internet composé de quatre pages web : A, B, C et D. La première approximation du PageRank serait répartie également entre ces quatre documents. Au départ, chaque document commence par une estimation de 0,25 PageRank (1/4).
Posons que :

– la page B ayant un lien vers les pages C et A ;

1. Le brevet est accessible en ligne à l'url suivante :`http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u="netahtml"PTO"srchnum.htm&r=1&f=G&l=50&s1=6,285,999.PN.&OS=PN/6,285,999&RS=PN/6,285,999`

2. ... mais aussi un jeu de mot avec le nom du chercheur Lawrence Page

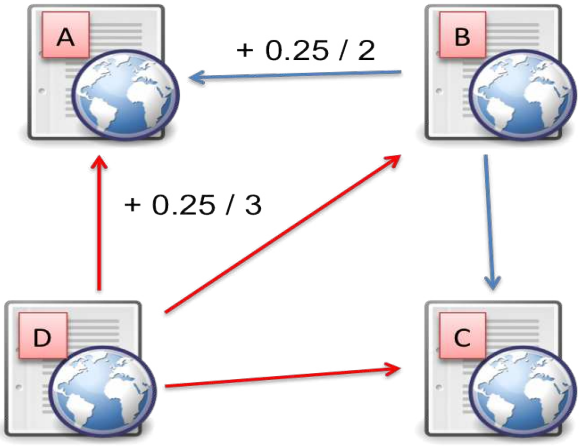


Figure 4.2: Illustration du PageRank

– la page D a des liens vers les trois pages.

Le crédit de lien est divisé entre tous les liens sortants d'une page.

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} \tag{4.10}$$

Ce qui, dans notre cas, donne :

$$PR(A) = \frac{1/4}{2} + \frac{1/4}{1} + \frac{1/4}{3} \simeq 0,125 + 0,25 + 0,083 \simeq 0,458 \tag{4.11}$$

Ainsi, la page B donne un vote de 0,125 à la page A et un vote de 0,125 à la page C. Seul un tiers du PageRank D est compté pour un PageRank (environ 0.083 voir illustration 4.2). En d'autres termes, le PageRank conféré par un lien sortant est égal au PageRank du document divisé par le nombre normalisé de liens sortants (L). Les liens vers des URL spécifiques ne comptent qu'une fois par document.

Pour l'usage grand public, le PageRank est lissé de manière logarithmique. Les internautes peuvent ainsi connaître une approximation grossière du PageRank (sur une échelle de 0 à 10) d'une page en consultant la Google Toolbar.

Plusieurs problèmes sont soulevés par le PageRank :

– La fraude aux liens de complaisance qui gonflent artificiellement le PageRank.
– La quantité de données. Calculer des relations dans une matrice avec des milliards de sites à traiter prend un temps qui rend le résultat approximatif dès la fin du calcul.

L'algorithme a été modifié pour prendre plus en compte le contenu et la sémantique des sites web pour le classement. Espérons que Google Panda apportera une solution à ces problèmes.

4.3 Conclusion

Les facteurs qui tendent à la réussite d'une recherche d'information sont multiples. Le méthodologie de recherche d'information est une clé importante pour une collecte efficace qui trouve son équilibre entre bruit et silence. Cependant la qualité des interfaces de recherches, des bases de connaissances ainsi que de la méthode d'indexation font également la différence entre un système bruyant ou silencieux et un système efficace. S'il faut une méthodologie pour accéder à l'information, savoir trier l'information est également un tâche majeure, même au sein de la littérature scientifique. Une bonne connaissance de la scientométrie est un plus indéniable pour choisir les lectures qui permettront d'étayer une réflexion scientifique solide. Du point de vue de l'usager d'un système d'information, la notion de pertinence peut être interprétée comme une quête de sens. Les déductions tirées de cet examen qualitatif vont créer un lien psychocognitif entre les utilisateurs de SRI et l'information. Certains de ces liens deviennent alors des frontières qui déterminent ce que sera, et ce qui ne sera pas, incorporé dans la structure cognitive (Bartlett et Toms, 2004). Quels sont les enjeux qui font d'une recherche d'informations une réussite ou un échec ? Posons-nous la question de l'intégration de l'information et du rapport qu'a le processus de recherche avec les acquis d'un usager.

Chapitre 5

Les écoles de pensée en RI : Processus et Cognition

Mais ces hommes sages, qui avaient jeté un regard savant sur la nature de la cognition humaine, en restèrent là ; et convaincus par tant d'essais qu'il n'en pouvait résulter rien d'absolument certain, ils bornèrent là leur recherche et s'arrêtèrent dans le doute..

Charles François Dominique Villers

Philosophie de Kant ou Principes fondamentaux de la philosophie transcendentale, p. 59

Ce que les hommes veulent en fait, ce n'est pas la connaissance, c'est la certitude.

Bertrand Russel

Introduction

Comme le montre ce début de ce manuscrit, le processus de formulation de la requête informationnelle n'est pas transparent pour la plupart des utilisateurs du système d'information.

L'utilisateur du service de recherche d'information (SRI) documentaire « a parfois des difficultés à exposer sa question (...), il faut l'aider à reformuler une problématique afin qu'elle puisse trouver un écho dans l'arbre des collections (Denecker *et al.*, 2000, p. 17-18) ».

Définition de collection

> Le terme de collection s'entend comme le regroupement volontaire de documents, d'objets, d'informations de provenances diverses, rassemblés en raison de la similitude d'un ou de plusieurs de leurs caractères. ADBS (2012).

Mais pour qu'un usager puisse convenablement formuler sa requête, faut donc qu'il prenne conscience de son besoin d'informations.

5.1 Le besoin d'information

Le besoin d'information selon Le Coadic (1997)

> Du site Internet à la bibliothèque en passant par le centre de documentation et le musée, du livre à la revue en passant par le journal, la radio, la télévision, le cinéma, de la banque d'information à la bibliographie en passant par la revue de sommaires, les systèmes, les services, les produits d'information sont destinés à répondre aux besoins d'information d'usagers multiples et variés qui feront de l'information qu'ils auront obtenue des usages multiformes.

Le Coadic (2008) reprenant Baudrillard (1973) posait le besoin d'information comme « problème cognitif à résoudre ». Dans un contexte donné, le besoin d'information est le constat pour un individu d'un « état de connaissance insuffisant ou inadéquat » pour atteindre un objectif (cf. schéma ci-après).

contexte –> problème –> besoin

Le besoin d'information en contexte

Les sciences cognitives ont connu un fort essor dans les années 90, c'est durant cette période que les sciences documentaires ont décidé d'en tirer partie pour comprendre les mécanismes cognitifs liés à la recherche d'information, en y associant parfois des éléments de psychologie. En effet, « les sciences cognitives apportent de précieuses indications sur les mécanismes cognitifs et sur la manière dont un individu traite l'information et utilise la bibliothèque (Denecker *et al.*, 2000, p. 21) ». Plusieurs méthodes de recherche d'information ont émergés de ces réflexions. Examinons ce cheminement.

Le problème du besoin d'information étant posé, il reste à le résoudre. Dans cette optique, Brookes (1980) propose son « équation fondamentale de la science de l'information » pour expliciter la transition de l'état de connaissance parcellaire initial C d'un individu λ , vers un état $C^{'}$ de connaissance augmenté par capitalisation du différentiel de connaissance δ extrait d'une information i.

$$C^{'} = C + \delta_{C_i} \tag{5.1}$$

Le besoin d'information, aussi appelé anomalie de connaissance par le Coadic, que nous appelons « différentiel de connaissance[1] » δ_C entre l'état initial C et l'état de satisfaction informationnelle $C^{'}$ par rapport au problème est résolu par l'apport de l'information i.

De leur coté, Tricot et Rouet (2004) définissent le besoin d'informations comme un besoin de réduction d'incertitude. Il est difficile de considérer qu'une seule information peut combler le différentiel de connaissance. Le Coadic (2008) poursuit donc son analyse de la vision de Brookes (1980) en y ajoutant que la transition d'état se fait plus comme une fonction récursive par la somme des informations $i_{1,n}$ extraites de divers documents. L'équation fondamentale de la science de l'information pourrait alors trouver une

1. Différentiel comme combinaison des accroissements infinitésimaux des états de connaissance, la transition d'état ne se fait pas sur une seule information, mais comme somme de petits éléments.

explicitation fonctionnelle ainsi :

$$C' = C + \sum \delta_{C_{i_{(1,n)}}} \tag{5.2}$$

Le Coadic (2008) pose le problème historique du manque d'intérêt de la bibliothéconomie en général pour les besoins d'information de l'usager préférant centrer sa réflexion sur le document. Il était considéré comme acquis qu'il est de la responsabilité de l'usager du système documentaire de diagnostiquer et d'identifier ses besoins. En particulier, sans une connaissance détaillée de la collection, et de l'environnement de numérique, la plupart des utilisateurs trouvent qu'il est difficile de formuler des requêtes en adéquation avec leurs besoins(Salton et Buckley, 1997). Connaître le mode de classification d'une collection n'est cependant pas suffisant pour en comprendre la complexité et donc l'utiliser efficacement car « le système de classification qui organise la collection en un ensemble intellectuel cohérent reste souvent impénétrable au public (Denecker *et al.*, 2000, p. 18) ». En prenant le point de vue de l'usager, on ne peut que se ranger à l'avis de Denecker *et al.* et « on finit par se douter que sa manière d'appréhender la bibliothèque n'est tout à fait conforme aux projections des bibliothécaires ». (Denecker *et al.*, 2000, p. 18) Nous présentons que l'opération initiale doit être intuitivement pensée comme un brouillon de formulation. Les éléments initialement récupérés pourront alors être confrontés à un filtre de pertinence basé sur l'expérience de l'utilisateur. Ainsi, de nouvelles formulations de requêtes améliorées pourraient être construites dans l'espoir de récupérer d'autres objets utiles lors des opérations de recherches ultérieures. Au fur et à mesure, l'utilisateur enrichit sa connaissance et sa compréhension du domaine qu'il étudie. Dans cette partie, nous allons tenter de démontrer la prépondérance d'une méthodologie de recherche d'information efficace. Dégager de l'information pertinente est une activité dont l'aspect (psycho) cognitif ne doit pas être négligé, sous peine d'aller vers de graves déconvenues.

Comme le disaient Salton et McGill (1986) dans *Introduction to modern information retrieval* :

« *Information retrieval also takes on aspects of behavioral science, since retrieval*

systems are designed to aid human activites[1]. »
Les difficultés de la recherche d'information ne peuvent être résolues simplement par la création d'interfaces qui permettraient à l'utilisateur de poser une question au système d'informations, à charge pour celui-ci de produire les documents pertinents. Salton rappelait ici qu'une recherche d'information ne se réduit pas à l'usage d'un système d'informations, c'est avant tout une activité humaine exigeant une méthodologie. La recherche d'information est une activité cognitive complexe. Elle fait appel à de nombreux savoirs et se compose en plusieurs tâches. De nombreux chercheurs se sont essayés à modéliser cette recherche d'information afin de produire des outils techniques d'aide à la RI, ou d'améliorer l'apprentissage de la RI.

5.2 Information et connaissance

5.2.1 Savoir ou savoir-faire

Le besoin d'information est un déficit d'informations qui peut avoir deux causes le besoin de savoir ou celui de savoir-faire. Le savoir regroupe les connaissances déclaratives (théories, lois, classifications...) alors que les savoirs faire sont les mises en situation des connaissances. Robert Brien synthétisait ainsi les rapports entre savoir et savoir faire : « Les connaissances déclaratives sont utilisées pour se représenter le réel et les connaissances procédurales pour interagir sur lui (Brien, 1994, p. 76) » Cette distinction semble maintenant admise : « La distinction entre mémoire déclarative (...) et mémoire procédurale est maintenant largement acceptée (Weil-Barais *et al.*, 2011) ».

5.2.2 Connaissances générales et spécifiques

Une autre manière de proposer une typologie des connaissance et celle qui oppose les connaissances générales et spécifiques (Denecker *et al.*, 2000, p. 27) . Si les connaissances générales sont interdisciplinaires, celles plus spécifiques traitent de sujets plus pointus. Tardif à ce propos pose comme définition : « Les connaissances spécifiques sont étroitement liées à des contenus disciplinaires ou à des champs de connaissance

1. Traduction proposée : La recherche de l'information couvre également des aspects des sciences du comportement, puisque les systèmes de recherche sont conçus pour aider les activités humaines.

particuliers (Tardif, 1997) ». En cas d'expérience incomplète d'un champ de connaissance, l'usager dans son besoin de connaissance ne conceptualisera pas certaines parties du texte des documents issu du résultat de sa recherche. Il sera obligé de combler les zones d'ombres par extrapolation avec ses connaissances génériques. Cette opération de déduction logique à partir du contexte et de ses connaissances s'appelle l'inférence. Inférer est une phase essentielle de l'acquisition de connaissance. Pour résoudre un besoin d'information, il faut souvent savoir « lire entre les lignes » et déduire la réponse. Cette étape active de cognition de par sa nature dynamique favorise le passage en mémoire de savoir. Ce peut être un atout cognitif en cas d'interprétation exacte de l'information. Cependant, Richard présente le danger de faire passer durablement des erreurs, ou des approximations en mémoire à long terme. Ces défaillances risquent de handicaper l'usager dans son savoir-faire par exploitation de savoirs incorrects ou non généralisables (Richard, 1990).

5.2.3 Connaissance et classification

Nguyen-Xuan positionnait le concept au centre du modèle cognitif grâce auquel l'usager perçoit et organise le monde [1] (Nguyen-Xuan, 1995). Ainsi, il serait possible de quantifier l'expertise d'un individu [2] à l'aune de la densité de son réseau sémantique sur un sujet donné (Tardif, 1998).

Le savoir, ou connaissance propositionnelle, est organisé à la manière d'un réseau sémantique, appelé aussi arbre conceptuel. Il se compose de nœuds, les concepts, et d'arcs qui représentent les relations entre les concepts [3].

Définition de concept en sciences cognitives :

> Le terme de concept s'entend comme support des connaissances déclaratives ou selon (Monday, 1996) comme représentation mentale abstraite d'un objet ou d'une idée.

1. Nous consacrerons le chapitre 10 au choix d'un modèle organisationnel de la connaissance.
2. Il est entendu la quantification ce que l'individu connaît, pas ce qu'il est capable de faire de cette connaissance. C'est toute la question du savoir et du savoir faire. Ce qui est sûr, c'est qu'il n'y a pas de savoir-faire sans savoir.
3. Nous proposerons dans la partie 11.5 des méthodes graphiques pour représenter cette architecture de connaissance.

	Étape	*Description*
1	Constitution d'un objectif	Représentation cognitive du but à atteindre
2	Sélection de documents	Évaluation sur critères de pertinence
3	Extraction de l'information	Traitement de l'information identification de termes en rapport avec le thème
4	Intégration de l'information	Intégration des documents à la bibliographie et assimilation des concepts proposés
5	Recyclage de l'information	Évaluation sur critères de pertinence

Tableau 5.1: Processus de gestion cognitive du modèle de Guthrie

Il existe une multitude de manières de concevoir la recherche d'information. Nous allons dans cette section analyser ce qu'est une recherche d'information sous différents axes.

5.3 Modèle de Guthrie

Guthrie (1988) a le premier établi un modèle de l'activité cognitive lors de la recherche d'information. Bien qu'ayant été révisée depuis, cette visualisation du processus de recherche a servi de base pour beaucoup de chercheurs. Cinq étapes ont été définies qui mènent de la formation d'un but au traitement de l'information. Tout d'abord, la personne qui commence une activité de recherche forme un but et ce faisant il élabore une représentation de l'objectif à atteindre. Puis elle doit chercher la source d'information qui est censée lui apporter les réponses attendues et en extraire l'information. Enfin, l'information est intégrée, c'est-à-dire qu'elle est reliée aux autres savoirs que possède le sujet.

Ces étapes constituent un cycle car la personne peut reprendre les quatre premières étapes si l'information n :a pas été trouvée ou est insuffisante. Il fait l'hypothèse que la recherche d'information comporte cinq étapes ou constituants cognitifs (voir aussi le tableau 5.1) :

1. la constitution d'un objectif de recherche ;

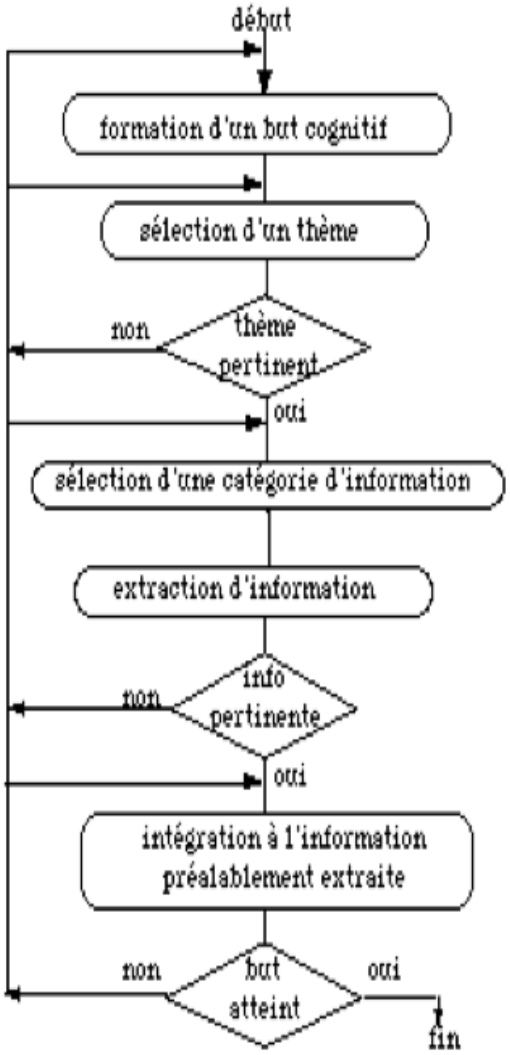


Figure 5.1: Diagramme « État-Transition » du modèle de Guthrie proposé par par Tricot

2. la sélection d'une catégorie : le sujet choisit, parmi les sources d'informations disponibles, celle qui paraît la plus pertinente à partir des caractéristiques structurales du document (par exemple, un paragraphe ou une colonne dans un tableau) ;
3. l'extraction de l'information : dans la catégorie choisie, le sujet traite l'information de contenu (par exemple, identification d'un mot ou d'une valeur numérique) ;
4. l'intégration : La collection est augmentée des informations acquises ;
5. le recyclage : Si l'objectif de recherche formalisé à la première étape n'est pas atteint, alors un branchement vers l'étape deux est nécessaire jusqu'à ce que le but soit atteint.

La thèse de Guthrie a été la base de recherche de nombreux modèles de représentation de l'activité cognitive de recherche d'information.

5.4 Le modèle « évaluation, sélection, traitement » (EST)

Le modèle EST a été conçu par les chercheurs français Jean-François Rouet, Jérôme Dinet et André Tricot. Jean-François Rouet est chercheur en psychologie et directeur de recherche au CNRS. Initialement, ce modèle se déclinait en trois étapes, ou « modules », pour reprendre le terme utilisé par Jean-François Rouet (Rouet *et al.* (2004)) : Evaluation, Sélection, Traitement. Suite à une recherche dans un thème donné, le processus se décompose linéairement sur l'axe temporel en 3 étapes

1. Sélectionner les documents qui semblent les plus en adéquation avec l'objectif de recherche.
2. Traiter les documents, c'est dire les lire, les analyser pour les comprendre et en évaluer le contenu, toujours par rapport au contexte de recherche.
3. Intégrer les documents sélectionnés au corpus en cours de réalisation afin de créer une bibliographie relative au sujet traité.

Conceptuellement, le modèle EST distingue trois aspects de la gestion des activités de RI (voir le tableau récapitulatif à double entrée 5.2) :

1. la planification de l'activité ;

	Planification	*Contrôle*	*Régulation*
E	Construire une représentation de la tâche et de la solution	Vérifier si la solution correspond à la tâche initialement définie	Modifier la solution et/ou la représentation de la tâche
S	Identifier les catégories disponibles	Vérifier si la catégorie correspond à la tâche initialement définie	Modifier les critères de pertinence
T	Évaluer les paramètres et choisir une stratégie	Vérifier si le contenu traité correspond à la tâche	Interrompre l'activité, relire, corroborer

Tableau 5.2: Processus de gestion cognitive et processus de base du modèle EST
Dinet et Rouet (2002)

2. le contrôle du résultat en cours de l'activité ;
3. la régulation ou modification dynamique du déroulement de l'activité.

5.5 Le modèle « *Infomation Search Process* »

Carol Collier Kuhlthau est professeur émérite et chercheur à l'université d'état du New Jersey au sein de l'école Rutgers (School of Communication, Information, and Library Studies). Elle met en évidence les étapes de du processus de recherche d'information tout en y associant des sentiments, pensées, actions, et tâches. Dans son livre *Seeking Meaning*, Kuhlthau (2003) énonce le principe suivant :

« Le processus de recherche d'information est enclenché par un état d'incertitude dû à un manque de compréhension, à un sens inexpliqué, à une structure incomplète. Il s'agit d'un état de nature cognitive qui provoque généralement des symptômes affectifs comme l'anxiété et le manque de confiance. Il faut s'attendre à la présence d'incertitude et d'anxiété dans les premières étapes du processus de recherche d'information. Les symptômes affectifs d'incertitude, de confusion et de frustration sont accompagnés de pensées vagues et floues à propos d'un sujet ou d'une question. Au fur et à mesure que les états de connaissance se précisent en des pensées plus claires, on note une évolution

	Tâche	*Ressenti*	*Description*
1	Reconnaître un besoin d'information.	incertitude et appréhension	faire le lien avec son expérience et connaissances
2	Identifier un thème	optimisme	identifier et de sélectionner le thème général
3	Exploration	confusion, incertitude et doute	première recherche pour faire émerger les concepts du domaine
4	Formulation de la requête	l'incertitude diminue, la confiance augmente	mise perspective les points clés récoltés
5	Étape de collection	Le sentiment de confiance continue à augmenter	collecte des documents ciblés
6	Présentation des résultats	soulagement et satisfaction.	Filtrage, classement et présentation des résultats

Tableau 5.3: Processus de base du modèle de ISP de Kuhlthau et sentiments associés

parallèle des émotions vers plus de confiance [1] ».

Kuhlthau semble persuadée du parallèle entre l'expérience du sujet (chez nous le jeune chercheur), son état émotionnel et l'évolution de la qualité la recherche d'information. Son approche est pédagogique, son contexte de recherche se situe donc dans le cadre de la formation de chercheurs-apprenants. Cependant, lorsqu'un chercheur, même expérimenté, s'aventure sur un domaine qu'il ne maîtrise pas pour une recherche transverse, nous supposons que le processus doit être identique [2].

Pour un processus complet de recherche, Kuhlthau (1991) distingue six étapes que nous avons résumées dans le tableau 5.3 et explicitées ci-après :

1. Au commencement, une personne prend connaissance d'un manque, que ce soit en matière de connaissance ou de compréhension. La personne doit faire face à des sentiments d'incertitude et d' appréhension. La tâche est simplement de reconnaître un besoin d'informations. De manière « diffuse » le chercheur voit

1. Traduction de Paulette Bernhard, professeur à l'école de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal.

2. Cette dernière supposition n'est pas tirée d'un écrit de Kuhlthau, mais procède d'une analogie que nous pensons logique.

naître des réflexions sur la manière d'envisager le problème, de comprendre la tâche. Pour faire face à cette problématique, il doit essayer d'en faire le lien avec ses expériences et connaissances antérieures. Ces actions impliquent fréquemment de discuter de sujets et approches possibles.

2. Au cours de la « sélection » , la tâche est d'identifier et de sélectionner le thème général sur lequel effectuer sa recherche et de l'approche à adopter. Le sentiment d'incertitude cède alors fréquemment le pas à l'optimisme. La personne se sent alors prête à commencer la recherche. Les réflexions sont alors centrées sur le poids des éléments à chercher par critères d'intérêt personnel, des informations disponibles, et du temps alloué. Certains peuvent effectuer une recherche préliminaire des informations disponibles, les balayer pour un aperçu de termes alternatifs. Lorsque, pour une raison quelconque, cette sélection est retardée ou reportée, les sentiments d'anxiété sont susceptibles de s'intensifier jusqu'à ce que le choix des termes soit fait.

3. L'étape d'« exploration » se caractérise par des sentiments de confusion. L'incertitude et le doute augmentent souvent chez le chercheur au cours de cette période. La tâche est d'étudier le thème général en vue d'élargir la compréhension personnelle du domaine. Ces réflexions permettent au chercheur d'être suffisamment informé sur le sujet pour se constituer un point de vue. À ce stade, une incapacité à formuler précisément une requête rend la communication entre l'utilisateur (s'il est débutant) et le système maladroite. Cette partie comporte une phase de la localisation d'informations sur le sujet général. La lecture d'informations générales permet de se renseigner sur le thème. De nouvelles informations viennent s'agréger à ce qui est déjà connu. Il semble plus pertinent pour Kuhlthau de laisser l'utilisateur recouper ses informations et les synthétiser pour laisser émerger des schémas de pensées. Cependant, les informations rencontrées correspondent rarement à un schéma unique. Les informations provenant de différentes sources semblent souvent contradictoires, voire incompatibles, du fait de l'évolution des recherches et des courants de pensée. Les utilisateurs peuvent trouver la situation plutôt décourageante et se sentir menacés. Ce qui provoque un sentiment d'inadéquation ainsi que la frustration. Cette stratégie a un coût, certains chercheurs sont peut-être enclins à abandonner la recherche à ce stade.

4. L'étape de « formulation » est le pivot du processus de recherche, lorsque les sentiments d'incertitude diminuent et que la confiance augmente. Grâce à l'ensemble des documents parcourus, une idée directrice du domaine de recherche prend forme. Cette étape de réflexion implique d'identifier et de sélectionner les éléments qui mettent en perspective les points clés autour desquels la collecte d'information va s'organiser. La formalisation des sujets va permettre au chercheur de s'approprier les concepts par une reformulation personnalisée.
5. L'étape de « collection » intervient lorsque l'utilisateur maitrise les fonctions du système d'information. À ce stade, il est à même de recueillir des informations liées au thème ciblé par une recherche exhaustive de toutes les ressources disponibles. Le sentiment de confiance continue à augmenter à mesure que disparaît l'incertitude dans le projet d'approfondissement.
6. L'étape de « présentation » est la dernière de la recherche. Il s'agit de l'écriture finale du travail à rendre. La recherche d'information est achevée. Il s'agit alors de mettre au propre les notes et les brouillons et se préparer à exploiter les résultats. Les doublons et le bruit sont éliminés. Les résultats sont donc filtrés et présentés par ordre inverse d'importance. Cette dernière étape s'accompagne de sentiments de soulagement et de satisfaction si la recherche s'est bien passée (ou de déception si elle n'a pas abouti).

5.6 Le modèle « TIMS » de Dillon

Dans son article de référence, Dillon (1996) énumère quatre compétences qu'il est important de maîtriser pour une recherche d'information efficace. Contrairement aux autres chercheurs, Dillon ne positionne pas son modèle sur un axe séquentiel strict avec des jalons. Ces compétences sont interdépendantes, mais pas bloquantes, pour mener à bien la recherche. Comme le montre le schéma original de Dillon (cf. Fig. 5.2), la découverte du type d'un document peut obliger l'utilisateur à revenir sur la planification globale de la tâche.

1. L'utilisateur du système d'information (SI) doit être capable de formuler son objectif pour construire une requête pertinente. Si cette étape est mal exécutée,

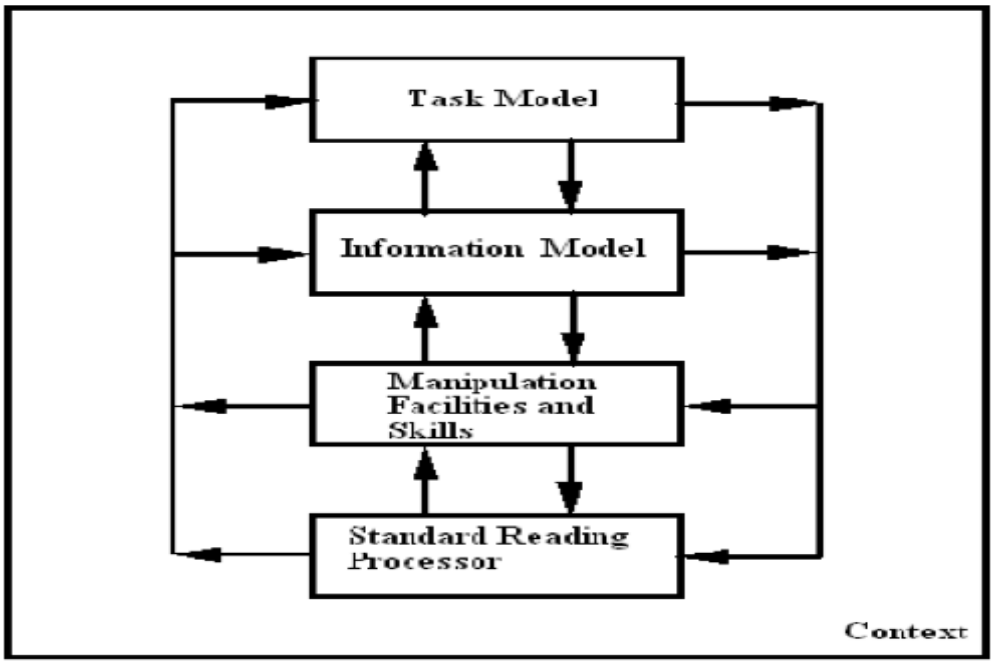


Figure 5.2: Interactions dans le modèle de TIMS, schéma original de Dillon

cela aura pour conséquence la formulation d'une requête imprécise avec pour résultat du bruit et du silence, mais également des réponses pertinentes.

2. L'utilisateur doit être à même de pré-visualiser le type de documents qu'il désire trouver ainsi que sa structure. Selon son niveau de compétence, cela peut prendre la forme d'un article de Wikipédia, d'un article de vulgarisation ou d'un article scientifique dans une langue étrangère. L'expérience permet de localiser plus facilement les points clés d'un document comme le titre, le résumé ou encore la bibliographie.

3. Pour trouver de l'information, l'utilisateur doit manipuler le support d'information. Dillon nomme cette capacité « *Manipulation Skills and Facilities (MSF)* », littéralement « aptitude et habileté à manipuler ». Nous n'aborderons pas le cas du livre, hors sujet dans notre cadre. Pour un document électronique, cela concernera l'habileté à naviguer dans un portail, un site et même dans une page affichée à l'écran. Si la prise en main peut prendre du temps et ralentir la recherche, elle ne peut pas la bloquer.

	Capacité	*Skill*	*Description*
T	Plannification de la tâche	*Task model*	Formation de l'objectif de la recherche. Sujet à réadaptation si besoin.
I	Prévisualisation structurelle	*Information model representation*	Représentation mentale de la structure du document
M	Manipulation	*Manipulation*	Habileté à manipuler le système d'information
S	Lecture	*Standard reading processor*	Lecture et compréhension

Tableau 5.4: Processus de base du modèle TIMS

4. Capacité à exercer une activité de lecture. Depuis compréhension de concepts complexes et/ou abstraits, jusqu'à la capacité de synthétiser l'information. La lecture peut prendre plus de temps et se focaliser sur de la vulgarisation.

A chaque étape de la recherche documentaire, un nouvel élément peut pousser l'utilisateur à revenir sur une étape antérieure pour affiner sa stratégie globale. Dillon donnait comme exemple :
« *For example, a scenario can be envisaged where, reading an academic article for comprehension, the task model interacts with the model to identify the best plan for achieving completion. This could involve several TM —> IM and IM—>TM interactions before deciding perhaps to serially read the text from start to finish*[1] ».
Ainsi, pour une recherche donnée, le type ainsi que les métadonnées des documents sélectionnés peuvent amener l'utilisateur à décider soit de lire le document, soit d'en sélectionner un autre, soit de relancer une recherche.

1. Proposition de traduction : Par exemple, un scenario peut être envisagé dans lequel, en lisant un article académique pour la compréhension, la planification de la tâche interagit avec le modèle pour identifier le meilleur moyen pour réussir ? Cela peut intégrer plusieurs interactions et manipulation avant de, peut-être, se décider, à lire le texte en entier.

5.7 Le modèle « *berrypicking* » de Bates et le phénomène de sérendipité

Marcia J. Bates a publié de nombreux ouvrages dans les domaines de la stratégie de recherche du système d'informations, la conception centrée sur l'utilisateur des systèmes de recherche de l'information. Ses recherches portent tant sur l'organisation des connaissances que les comportements de recherche. Marcia J. Bates est professeur émérite au *Department of Information Studies* de l'université de Californie.

Très intuitif, le modèle Bates (1993) est appelé « *berrypicking* », car il compare la recherche d'information à une cueillette de baies éparpillées dans les « buissons de connaissance ». Dans un processus de recherche documentaire, le chercheur consulte un premier document qui lui offrira de nouvelles pistes que ce soient des idées ou d'autres références. Les idées développées dans un premier document mèneront à d'autres axes de recherches ou points de vue, qui amèneront d'autres recherches. Le chercheur se constituera rapidement une arborescence documentaire, qui pourra même devenir un maillage, car les liens bibliographiques finissent par se recouper. Ce modèle permet de constituer rapidement une bibliographie autour d'un concept. Cependant, l'abondance d'informations et de documents peut également « noyer » le chercheur dans un flot d'idées contradictoires et le rendre indécis sur sa recherche au lieu de le conforter dans l'exploitation de son idée originale.

Cette méthode est cependant excellente pour développer la faculté de sérendipité [1] Selon Merton et Barber (2003), ce néologisme est construit à partir du vocable anglais « *serendipity* » lui-même issu de l'ancien nom propre « *Serendip* » qui désignait en perse ancien le Sri-Lanka. La légende contée par Walpole (1754) veut que trois jeunes princes découvrirent la solution d'une énigme par un enchaînement de déductions logiques issues de leur observation. L'histoire complète est narrée en français par van Andel et Jacquemin (2005) dans l'article *Sérendipité, ou de l'art de faire des trouvailles*, indispensable pour étoffer sa RI. Cependant dans leur livre van Andel et Bourcier (2008) distinguent bien chance et pugnacité. La sérendipité n'est donc pas le hasard. Il s'agit d'un usage continu de la sagacité et de la perspicacité du chercheur. Il s'agit

1. La sérendipité est le don ou la faculté de trouver quelque chose d'imprévu et d'utile en cherchant autre chose, ou encore, l'art de trouver ce qu'on ne cherche pas.

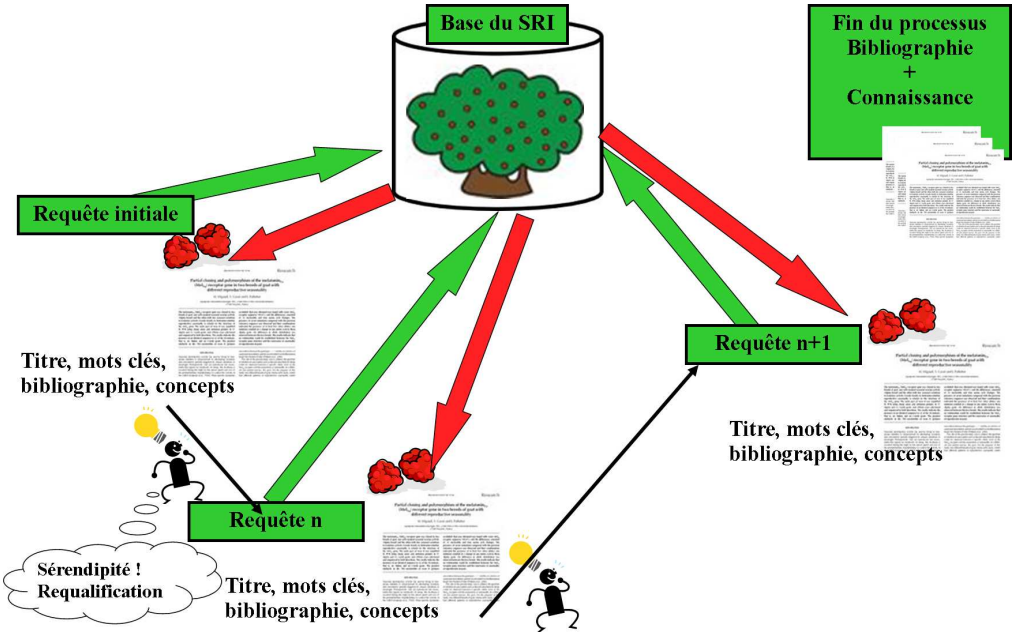


Figure 5.3: Berry Picking et sérendipité

de recouper des idées issues du texte ou du paratexte[1] ainsi que des bibliographies pour en extraire une « pépite » de connaissance. Il peut arriver également, en cherchant à comparer deux courants de pensée d'en faire émerger un troisième moins connu. Il faut cependant maîtriser la chaîne de recherche pour prétendre faire émerger quoi que ce soit d'un processus documentaire. Nous pouvons dire que dans le cadre du modèle de Bates la recherche documentaire avance par « saut » d'un document vers un autre. Le lien motivant l'analyse du document suivant n'est plus une évolution séquentielle classique comme un classement dans un moteur, mais la mise en exergue d'une référence bibliographique rencontrée à plusieurs reprises, un hyperlien, une métadonnée ou une idée novatrice.

5.8 Le modèle « *information-seeking* » de Marchonini

Gary Marchionini est professeur et doyen à l'École de l'information et de biblio-

1. Le paratexte d'un document scientifique a été inventorié par Genette et McIntosh (1988)

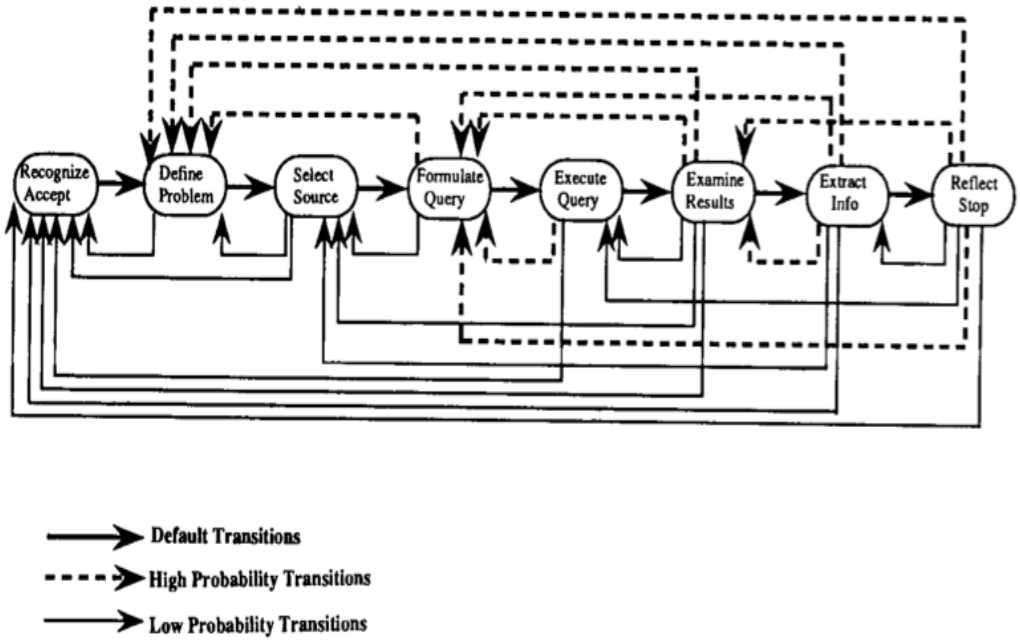


Figure 5.4: « Information seeking », schéma original de Marchionini

théconomie au sein de l'Université de Caroline du Nord. Dans le troisième chapitre « *Information-Seeking Perspective and Framework* » de son livre *Information Seeking in Electronic Environments* » Marchionini (1997) propose une autre vision du processus de recherche d'information que nous résumons dans le tableau 5.5 et la figure 5.4 et dont voici l'explication détaillée :

1. *Recognize and accept an information problem* / Reconnaître et accepter un besoin d'information.
   Cette première partie est simplement la prise de conscience d'une lacune. Cette partie peut être de la simple curiosité ou un besoin documentaire professionnel.
2. *Define and understand the problem* / Définir et comprendre la problématique de recherche.
   La définition de la problématique est une étape critique dans la recherche d'information. Ce sous-processus reste actif tant que la recherche d'information progresse. Dans le schéma 5.4 les flèches en pointillé évoquent la forte probabilité d'avoir à redéfinir la problématique au fur et à mesure de l'avancement (et donc de la compréhension) du sujet.

	Capacité	*Skill*	*Description*
1	Reconnaître et accepter un besoin d'information.	*Recognize and accept an information problem.*	Prise de conscience d'une lacune.
2	Définir et comprendre la problématique.	*Define and understand the problem.*	La définition du problème, processus en potentielle évolution.
3	Choix d'un système de recherche.	*Choose a search system.*	Choisir une base de connaissance ou un moteur de recherche.
4	Formulation de la requête	*Formulate a query.*	Composition syntaxique de la requête (usage un langage formel et d'opérateurs booléens)
5	Exécution de la recherche.	*Execute search.*	Clic ou combinaison de touches.
6	Examen des résultats.	*Examine results.*	Évaluation de la pertinence des résultats, mais aussi de celle de la requête.
7	Extraction de l'information.	*Extract Information.*	Extraction d'information, de citations et d'éléments bibliographiques.
8	Réflexion, itération, arrêt.	*Reflect, iterate, stop.*	Choix d'arrêter la recherche, de la continuer ou de réévaluer la requête ou la problématique.

Tableau 5.5: Processus de base du modèle de Marchionini

3. *Choose a search system* / Le choix d'un système de recherche.
   Le choix d'un système de recherche dépend de l'expérience du demandeur d'information. Cependant, le choix peut être bridé par une simple question d'accès aux bases de connaissances. Le type des réponses attendues lors de la définition de la problématique de recherche sera également un facteur de choix. En effet la qualité des réponses n'est pas la même sur Wikipédia et sur un portail scientifique. Pour cette étape, la connaissance du domaine est une variable cruciale dans le choix d'un système de recherche. Plus l'utilisateur est affuté dans son domaine, plus son système de prédilection lui sera familier et moins il sera enclin à utiliser un service générique.
4. *Formulate a query* / formulation de la requête.
   La formulation de requêtes implique la compréhension de la syntaxe de recherche du système choisi. Dans de nombreux cas, la formulation la première requête identifie un point d'entrée. Cette étape est suivie par la navigation et / ou des re-formulations de la requête. La formulation de requêtes implique une bonne connaissance de la sémantique du domaine et de son champ lexical, qu'il s'agisse d'un portail de recherche ou d'un moteur. De plus, l'utilisateur doit également maîtriser le langage de requête utilisé par l'interface pour formuler une requête précise s'il s'agit d'un moteur.
5. *Execute search* / L'exécution de la recherche.
   Dans le cas d'une interface de type formulaire, connectée à une base de données en ligne, l'exécution peut être déclenchée en cliquant sur un bouton, une pression de touche spéciale (entrée par exemple). Dans le cas d'un portail composé d'hypertextes, l'exécution peut être lancée en cliquant les liens disponibles fournis par l'auteur (ou le webmestre).
6. *Examine results* / L'examen des résultats.
   La réponse fournie par le système est un résultat intermédiaire et doit être examinée par le demandeur d'information comme un moyen d'évaluer les progrès accomplis vers l'objectif de la tâche de recherche d'information. Cette réponse n'est donc pas une fin en soi. Selon la quantité et la qualité des réponses obtenues par le demandeur d'information lors de cette étape, il est à même d'évaluer de ses progrès dans le processus global, mais également dans l'activité de RI. Un

équilibre entre bruit et silence avec une pertinence correcte permet de passer à l'étape suivante. Il est probable qu'il faille à cette étape relancer une requête (étape 4) ou redéfinir le problème (étape 2) comme le montre le schéma 5.4 page 130.

7. *Extract Information* / Extraire de l'information.
   Pour extraire des informations, un demandeur d'information doit faire preuve de compétences comme la lecture, l'analyse et la classification. Le chercheur doit être capable de citer de manière clairement identifiable (copie de texte) et de référencer la citation.

8. *Reflect, iterate, stop* / Réflexion, itération, arrêt.
   Une recherche d'information aboutit rarement en une seule requête. Plus souvent, l'ensemble initial récupéré sert de rétroaction pour les formulations de requêtes supplémentaires et les exécutions. Décider quand arrêter la recherche exige une évaluation de l'ensemble du processus de recherche d'information. Il arrive régulièrement à cette étape de devoir réexaminer les résultats (étape 6), peut-être de relancer une requête (étape 4), voire dans certains cas de redéfinir sa problématique de recherche (étape 2).

5.9 Analyse comparative et critique des méthodologies

Au vu des quelques éléments que nous venons d'examiner, ll est possible de classifier les modèles de recherche d'information en 3 catégories.

- Recherche organisée, planifiée, de manière séquentielle ;
- Recherche par « tâches » ;
- Recherche par « sauts ».

Selon la thèse de Kuhlthau, d'une manière générale lors d'une recherche d'information (RI), le sujet passe de l'incertitude à la sérénité, à la condition, bien sûr, que le travail progresse correctement et que l'information trouvée le satisfasse. Cette optique met en avant l'aspect émotionnel de la recherche en balisant la méthode de recherche pour une procédure « sereine ». Entre les étapes de « formulation » et de « collection » Kuhlthau

annonce que la maîtrise du système d'interrogation et du domaine de connaissance sont liées.

Pour nuancer ces propos il faut distinguer que l'expertise en recherche d'information peut être, de son point de vue, représentée par une mesure sur deux axes. En effet, si la connaissance du domaine de recherche est indispensable à une recherche efficace, la bonne maîtrise des outils de recherche est également d'une importance capitale. Dans une étude de 2004 portant sur une cinquantaine d'étudiants, les chercheurs Ihadjadene et Martins (2004) montrent l'équivalence entre expertise du domaine et familiarité dans l'usage du Web. Ainsi, dans cette étude, la moitié des participants sont des experts en psychologie et l'autre des étudiants en d'autres matières. Dans chaque groupe, la moitié maîtrise les techniques de recherche sur Internet, l'autre est peu à l'aise. Dans le cadre de recherches sur des concepts de psychologie, les auteurs comparent des novices et des experts soit du domaine, soit de l'usage du Web. Les étudiants qui ont la double compétence de la connaissance du domaine et de la maîtrise d'Internet présentent de meilleures performances. Selon les auteurs : « L'indice de productivité relatif à la consultation des pages Web n'est significativement bas que chez les participants peu familiers avec le Web (qu'ils soient experts ou novices dans le domaine) ; on observe le même phénomène en ce qui concerne le nombre de (re)formulations de la requête ainsi que dans le coût temporel des réponses correctes. (...) l'absence de familiarité avec Internet rend la recherche d'information plus difficile cognitivement. » Le modèle de base est souvent celui proposé par Guthrie. D'autres chercheurs, plutôt que de vouloir l'approfondir ou le corriger, ont proposé une approche totalement différente, au sens où le caractère séquentiel de la recherche d'information est abandonné au profit d'une description des composantes de l'activité de recherche. C'est le cas du modèle TIMS initié par Dillon. Ce modèle insiste sur une description des types d'habiletés nécessaires lors d'une RI.

5.10 Conclusion

Dans l'optique de réaliser un système optimisé de recherche documentaire, nous reprendrons les conclusions de l'étude de Rouse et Rouse (1984) relative à la littérature de l'information sur les comportements de recherche :
« *Because information needs change in time and depend on the particular information*

seeker, systems should be sufficiently flexible to allow the user to adapt the information seeking process to his own current needs. Examples of such flexibility include the design of interactive dialogues and aiding techniques that do not reflect rigid assumptions about the user's goals and style[1]. »

Nous retiendrons le caractère évolutif des besoins d'un individu en matière de recherche d'information. C'est pourquoi les systèmes doivent être suffisamment souples pour permettre à l'utilisateur d'adapter le processus de recherche de l'information à ses propres besoins. Une forte interaction entre l'interface et le demandeur d'informations doit être possible. Nous en déduisons qu'il faut offrir plusieurs méthodes de recherches pour accompagner le chercheur dans ses évolutions. À ce stade de la réflexion, nous pensons que l'offre de recherche que nous désirons proposer doit fortement s'imprégner des modèles de recherche décrits dans cette partie. En combinant les aspects les plus intéressants de ces méthodes et en y associant des technologies de présentation, nous ambitionnons de simplifier (en le raccourcissant) le processus de recherche d'information. Cependant, comme le processus cognitif associé à l'usage d'un SRI est vital pour l'usager novice nous en tiendrons compte lors de sa modélisation. Nous ferons nôtres les phrases suivantes de Baeza-Yates et Ribeiro-Neto (2008) sur l'état d'esprit de l'usager en début de recherche :

« *When users approach an information access system they often have only a fuzzy understanding of how they can achieve their goals. Thus the user interface should aid in the understanding and expression of information needs.*[2] »

Nous retiendrons donc que ce SRI se doit d'être un outil d'apprentissage du domaine tout en restant abordable au sens de Kuhlthau. En effet, en tenant compte des sentiments liés aux étapes de la recherche, nous éviterons de décourager le demandeur d'information.

1. Piste de traduction : Parce que l'information a besoin d'évoluer dans le temps et dépend de l'individu, les systèmes de recherches devraient être suffisamment flexibles pour permettre à l'usager d'adapter le processus de recherche à ses propres besoins temporaires. Des exemples d'une telle flexibilité incluraient des dialogues interactifs et des techniques d'aides qui ne n'imposeraient pas des présomptions rigides en terme d'objectifs et de catégorie d'utilisateur.

2. Proposition de traduction : Lorsqu'un utilisateur accède à un système d'accès à l'information il n'a souvent qu'une compréhension floue de la façon dont il peut atteindre son objectif. Ainsi, l'interface utilisateur devrait faciliter la compréhension et l'expression des besoins d'information.

Marijuan *et al.* (2012) considèrent que l'évolution naturelle des sciences traduit une dynamique cognitive sociale déjà réalisée comparable à l'évolution des systèmes vivants par re-combinaison génomique. Les « scientomics » sont considérés comme un nouveau champ de recherche dans le domaine de l'organisation de connaissances. Cette science émergente appréhende le processus d'expansion et de re-combinaison scientifique. Par exemple, une analogie est faite relativement au séquençage génomique par des moyens informatiques pour « identifier les séquences informatives » que ce soit en génomique ou en classification scientifique. Elle peut expliquer l'évolution de la scientométrie et servir au développement d'outils conceptuels de planification scientifique et de gestion de recherche, mais aussi ouvrir des perspectives de sérendipité indispensables à un esprit à l'affut de potentiel d'évolution. Cette ouverture n'est pas sans rappeler la démarche de sérendipité provoquée par la capacité à écouter ses intuitions, quitte à s'écarter légèrement de son sujet de recherche initial.

C'est associé à cette interdisciplinarité que le web sémantique lié à des technologies de présentation, de recommandation et de classification avancées peut transcender les processus cognitifs classiques de RI vers une démarche plus globale d'avancée scientifique. Ce champ exploratoire permet une recherche d'information avec une méthodologie, débutant classiquement avec l'axiome du besoin d'information, mais s'ouvrant sur une évolution cognitive débordant peut être du champ disciplinaire initial. La question qui apparait alors est : de quelle manière peut-on modéliser la connaissance et comment amener les bases de connaissances éparses et structurellement différentes à communiquer avec les interfaces de recherche d'information ?

Deuxième partie

Pratiques de recherche bibliographique

Chapitre 6

Bibliographies, métadonnées et notices : *normes, formats et styles*

Je suis de ceux qui croient qu'un romancier, un écrivain n'a pas de biographie, il a une bibliographie.

Jean D'Ormesson

Introduction

DAns ce chapitre, nous allons nous interroger sur la manière de représenter une référence à un document. Les documents sont décrits de diverses manières selon le contexte de représentation. La base de la recherche d'information est l'indexation des références documentaires qui représentent les documents. Au sein d'un ouvrage ou d'un article, l'auteur appuie ses opinions et positionne épistémologiquement son propos en citant ses collègues et prédécesseurs. L'ensemble des références proposées par les citations est regroupé au sein d'une partie formalisée nommée bibliographie. En préalable à sa bibliographie, l'auteur effectue sa recherche documentaire et repère les notices des ouvrages relatifs à son besoin d'information.

Définition

Le terme bibliographie est formé de l'affixe d'origine grecque *biblio-*, « livre, et *grapho-*, « écrire ». Ce substantif signifie « liste des ouvrages cités dans un ouvrage ». Une bibliographie peut être également relative à un sujet donné [1] et donc dans ce cas être un document autonome.

Notice bibliographique

> Ensemble des éléments comprenant la description bibliographique d'un document et des points d'accès à celle-ci, généralement rédigée en suivant les prescriptions de normes nationales ou internationales.

Nous pouvons décrire une notice comme un ensemble structuré d'informations descriptives à propos de ressources physiques (imprimées) ou en ligne (électroniques). Une notice bibliographique est une fiche (cartonnée à l'origine, puis majoritairement électronique depuis la fin du 20e siècle) constituée d'un ensemble d'attributs, ou éléments, nécessaires pour décrire la ressource. Par exemple, un système commun de métadonnées dans les bibliothèques – le catalogue de bibliothèque – contient un ensemble de notices de métadonnées comprenant des éléments spécifiques pour décrire un livre ou tout autre document que l'on trouve en centre documentaire : auteur, titre, date de création ou de publication, sujet et cote, permettant de le retrouver dans les rayonnages. Dans ce

1. définition du Larousse : n.f. Connaissance des ouvrages et de leurs éditions, envisagés dans un domaine déterminé. Liste de titres, de références d'ouvrages, de livres ou de périodiques relatifs à un domaine général ou spécialisé.

chapitre, nous allons présenter les aspects normatifs liés aux bibliographies, les logiques de description, d'échange et de présentation. Enfin nous entamerons une discussion sur l'intérêt de la formalisation bibliographique et de l'automatisation des processus associés.

6.1 Les métadonnées et notices bibliographiques

Une notice constituée de métadonnées peut être incluse dans la ressource. La ressource sera dans ce cas réputée « auto-décrite ». L'autre solution de mise à disposition de métadonnées est de référencer le document dans une notice au sein d'un catalogue.

Un des rôles fondamentaux des documentalistes est de recenser les différentes productions des laboratoires de recherche et de créer des notices bibliographiques d'indexation.

Avec l'avènement du tout numérique et l'élargissement des catalogues bibliographiques, la masse de métadonnées pour un catalogue devient titanesque. Pour accélérer et élargir le processus d'affichage et de distribution du catalogue documentaire, les notices doivent être informatisées. Une notice informatisée doit obéir à une norme pour être lue par un processus automatisé et affichée de manière standard. Selon Françoise Leresche (2004) de la BnF, cette notice obéit à un modèle de présentation structurée de l'information bibliographique en vue de son échange sur support informatisé et / ou de son traitement dans un système informatisé. Cette standardisation est appelée « format » de normalisation de données. Selon Arlette Boulogne (2004), en informatique documentaire un format est un « agencement structuré des données numériques sur un support lors de leur production, leur affichage, leur stockage sur ce support, leur compression, impression ou diffusion ». Nous appellerons ce type de structure les métadonnées, que nous définissons de manière synthétique de la manière suivante :

Métadonnées

> Ensemble structuré de données créées pour fournir des informations sur des ressources.

6.1.1 Nature et fonctions des notices

Ces notices permettent de classer et décrire les documents d'une collection en fonction du ou des sujets principaux traités. Ces notices sont constituées du titre du document, de sa collection, de l'auteur, de sa date de publication, de son contexte de publication, de l'éditeur et de la pagination. L'usage des notices bibliographiques sur fiches cartonnées a été rendu obsolète par l'arrivée de l'informatique grand public.

6.1.2 Nature et fonction des métadonnées

Les métadonnées sont, en contexte documentaire, des données à propos d'un document. Une métadonnée est littéralement une donnée décrivant une donnée. Les métadonnées décrivent une ressource d'information, dans notre contexte des documents scientifiques. Le terme grec « meta » dénote une autodescription. Les métadonnées remplissent les fonctions suivantes (NISO-PRESS, 2004) :

- L'évolution temporelle des ressources décrites (création, modification, archivage) ;
- l'information sur le contenu de la ressource pour en faciliter la localisation et l'accès ;
- les droits et conditions d'utilisation associés à la ressource.

En effet, le(s) auteur(s), date de publication et éditeur ne prêtent ni à confusion ni à interprétation. Certaines métadonnées sont cependant plus complexes selon Peccatte (2007), notamment celles plus subjectives de description par mots clés qui peuvent être sujet à caution et interprétation de la part de l'utilisateur final. En effet, selon la formation initiale ou l'école de pensée du chercheur, un terme peut être usité de manière différente, sans que pour autant l'on puisse parler de polysémie. Un même terme peut avoir plusieurs aspects selon les centres d'intérêt. Même si les métadonnées sont vitales dans le cadre d'une recherche par l'usager de SRI, notamment dans de vastes corpus, leur choix doit être le plus neutre possible.
L'ensemble de ces métadonnées forme une notice bibliographique qui peut être plus ou moins dense selon le nombre de métadonnées et la richesse du contenu. Nous allons examiner les principaux formats d'échange de données bibliographiques qui font autorité dans les sphères universitaires, scientifiques et industrielles

6.2 L'objet bibliographique

Boulogne (2006), ancienne directrice de l'Institut national des sciences et techniques de la documentation définit la bibliographie comme une liste de références ou de notices bibliographiques. Contextuellement, son usage permet aux lecteurs d'identifier les documents référencés dans une production écrite. Ainsi, une bibliographie est une partie d'un ensemble plus complexe, le document, qu'il s'agisse d'un livre, d'un article ou d'une thèse.

6.2.1 Définition de l'objet bibliographique

bibliographie

> Liste de références bibliographiques classées selon certains critères pour en permettre le repérage.

La présentation des éléments de la bibliographie peut être organisée de manière signalétique et classée :

- Par nom d'auteur ;
- Par ordre d'apparition dans le texte ;
- Par date de publication.

La bibliographie peut aussi être présentée de manière thématique, par sujets abordés ou par courants de pensée.

6.2.2 Nature et fonction de la bibliographie

On utilise parfois par abus de langage le terme de bibliographie pour tout autre chose qu'un ensemble thématique de notices bibliographiques. Il arrive de faire référence dans un ouvrage à des documents autres que des ouvrages imprimés. Il peut s'agir d'archives vidéos, audio, de planches d'art graphique, site Internet et autres Web-log (blog) ou tout autre « document ». Il arrive parfois que ces autres supports soient séparés de la bibliographie proprement dite et soient inscrits sous forme alternative avec le type e support suivi du suffixe « graphie ». Nous trouverons ainsi des webographies (ou sitographies), discographies, filmographies etc.

Par convention, la bibliographie d'un ouvrage est placée en fin de document et les références incluses dans le texte sont suffisamment identifiables pour distinguer les idées

de l'auteur de celles des références. Dans ce cadre, la bibliographie est le fil conducteur retraçant les travaux antérieurs qui ont permis de structurer la réflexion aboutissant à la rédaction du document en cours de lecture (Boulogne, 2006). Une bibliographie peut être présentée sous forme :

- d'un document (ou produit bibliographique) autonome ;
- d'une annexe à un document ;
- de la partie finale d'un document, pour recenser toutes les références faites dans le développement.

Il est également possible de mettre les bibliographies en avant comme un objet indépendant qui met en exergue (en relation avec des indices bibliométrique) les auteurs phares autour d'une thématique, mais aussi les contradicteurs. L'examen des bibliographies illustre aussi par recoupement les revues et conférences clés d'une branche scientifique, c'est pourquoi il est courant que les enseignants du supérieur distribuent des bibliographies relatives à leur cours en début de semestre. Les étudiants peuvent ainsi se faire une idée *a priori* du contenu des cours. Cela leur permet de s'imprégner de la substance du cours, de mieux saisir les cours magistraux et de se préparer aux travaux dirigés.

6.2.3 Normes liées aux bibliographies

Une bibliographie, pour être compréhensible doit être normalisée ; ainsi, les entrées qui la composent sont clairement identifiables. Cette condition nécessaire n'est pas suffisante pour rendre une bibliographie limpide à un lecteur. Il faut également que la structure de chaque élément bibliographique obéisse à des règles de présentation, parfois implicites, avec pour objectif de présenter de manière distincte les différents types de documents. Ces règles secondaires de présentation sont dépendantes des prescriptions d'une discipline, d'un éditeur, d'une conférence. Il arrive même que les universités prescrivent leur propre style bibliographique. Nous consacrerons une sous partie aux styles de présentations bibliographiques dans le présent chapitre (cf. 6.5).

Il est courant pour un chercheur de partager des bibliographies avec une équipe ou un laboratoire lors de travaux communs. Les éditeurs de littérature scientifique, les chercheurs, les moteurs de recherches scientifiques, les plateformes de dépôt pour le

partage et l'archivage de documents scientifiques ont dû normaliser les méthodes pour exposer et échanger des données bibliographiques de manière électronique.

Elles peuvent être exprimées dans le même format technique de codage que celui de la ressource qu'elles accompagnent et être disponibles en même temps que celle-ci. Les métadonnées peuvent être écrites actuellement selon plusieurs standards : RDF (*Resource description framework*), TEI (*Text encoding initiative*), syntaxe *meta* en HTML et Dublin Core, DTD EAD (*Encoding archival description*), etc.

C'est ce dernier aspect que nous allons principalement essayer de développer dans cette partie du chapitre.

6.3 Les formats de notices bibliographiques et de classification

6.3.1 Les formats d'échange de données bibliographiques

La cadre général : La norme ISO 2709

La norme de l'organisation internationale de standardisation 2709 et baptisée « Information et documentation – Format pour l'échange d'information » spécifie les exigences d'un format générique d'échange bibliographiques [1]. Cette norme permet de décrire des documents de tous types susceptibles de constituer une bibliographie. Dans ce cadre, la communication entre SI est privilégiée au détriment de l'exploitation humaine. L'ISO 2709 a donc pour objectif de normaliser des échanges bibliographiques automatisés entre systèmes d'information.

6.3.2 Le Dublin CORE

Le Dublin Core Metadata Initiative est connu sous son acronyme DCMI. Il est aussi appelé simplement « Dublin Core » ou « DC ». Il s'agit d'un un format permissif de description générique, qui peut être appliqué à tous types de documents.

1. Disponible en français et en anglais à l'URL : `http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41319`

Historique du Projet

La paternité de ce projet revient à la communauté *Dublin Core Metadata Initiative* dont les premiers échanges datent de 1996 et font suite à la réunion tenue aux États-Unis, plus précisément au siège de l'OCLC à Dublin (Ohio) en 1995 autour des métadonnées. Des représentants de communautés issues d'horizons aussi divers que la documentation et les bibliothèques avec l'OCLC (Online Computer Library Center), l'informatique avec le NCSA [1], les sciences de l'information et de la communication se réunirent pour établir des bases communes pour les métadonnées.

Vers une normalisation du DCMI

Le Dublin Core est un ensemble de 15 éléments de description qui a émergé de cet effort dans la construction d'un consensus interdisciplinaire et international. La première formalisation officielle du groupe de réflexion fut la recommandation proposée par les membres de l'IETF [2] Weibel *et al.* (1998) dans la RFC 2413, aussi nommée *Dublin Core Metadata for Resource Discovery.*
L'événement majeur de ce groupe de réflexion est l'*International Conference on Dublin Core and Metadata Applications* [3]. Ce rendez-vous existe sous sa forme actuelle depuis 2001. L'ensemble de ses recommandations tend vers une normalisation, puisque depuis 2003 l'Organisation Internationale de Normalisation (ISO [4]) propose des recommandations officielles, remises à jour en 2009 et baptisées « Information et documentation – L'ensemble des éléments de métadonnées Dublin Core » sous le sigle « ISO 15836 :2009 » [5] puis rapidement corrigée en « ISO 15836 :2009/Cor 1 :2009 » [6]. Certains de ces des-

1. Le National Center for Supercomputing Applications / Centre National pour les Applications des Super-calculateurs est un centre américain de recherche et d'exploitation des super-calculateurs. Il est situé sur le campus de l'université de l'Illinois dans les villes jumelles de Champaign et Urbana. Ce centre de recherche est à l'origine du serveur web Apache, du client de terminal telnet et du navigateur Mosaic

2. Internet Engineering Task Force, groupe informel, international qui produit la plupart des nouveaux standards de l'Internet

3. `http://dublincore.org/workshops/`

4. `http://www.iso.org/iso/fr/catalogue_detail.htm?csnumber=37629`

5. `http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52142`

6. `http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=54784`

cripteurs ont trait au contenu, d'autres à la propriété intellectuelle, d'autres, enfin à la version. Dans les faits, comme nous l'avons souligné, le Dublin Core est une norme depuis 2003. Il s'agit cependant plus d'un standard composé d'un ensemble ouvert (extensible) de recommandations relatives à la mise à disposition de métadonnées que d'une norme. La référence ISO, qui sert de tronc commun, fait état d'un ensemble de 15 descripteurs de portée très large et de sens générique. Le résumé offert sur le site français de l'ISO stipule en effet :
« *L'ISO 15836 :2009* » est une norme de description de ressources interdisciplinaires, connue sous le nom de « L'ensemble des éléments des métadonnées Dublin Core ». Comme le « RFC 3986 » , la norme ISO 15836 :2009 n'est pas limitée à ce qui peut constituer une ressource. Cela définit les éléments généralement utilisés dans le contexte d'un profil d'application qui en cible ou en spécifie l'utilisation conformément aux préconisations et pratiques locales ou de communautés professionnelles. Cependant, cette norme ne définit aucune précision de mise en application, ceci ne relevant pas du champ d'application de l'« ISO 15836 :2009. » Chaque organisme qui utilise le Dublin Core est donc libre dans le cadre de sa mise en œuvre.

Spécification technique du DCMI

Dublin Core simple et qualifié Pour faciliter une mise en œuvre technique dans un cadre professionnel du Dublin Core, deux formats de Dublin Core sont apparus.

- Le Dublin Core simple est composé des 15 éléments fondamentaux, il sert pour une description générique, généralement pour un usage non professionnel.
- Le Dublin Core qualifié offre une granularité plus fine. Comme le Dublin Core simple, il est composé des 15 éléments initiaux, avec la possibilité d'affiner avec des spécifications quand au fond (au niveau du sens) ou des précisions (techniques) sur la forme de la ressource. [1]

Le Dublin Core simple Dans le tableau 6.1, les 15 descripteurs fondamentaux formant le socle du Dublin Core, tel que défini par l'ISO sont explicités et brièvement décrits. Les attributs ont été séparés en trois sous ensembles de métadonnées ayant trait

1. La définition du Dublin Core qualifié est disponible sur le site du DCMI à l'URL `http://dublincore.org/documents/usageguide/qualifiers.shtml`

Désignation	Attribut DC	Définition
Titre	Title	Titre de la ressource
Couverture	Coverage	Portée géographique et temporelle de la ressource
Source	Source	Ressource primaire dont est dérivé le document
Sujet et termes clés	Subject	Objet de la ressource et / ou termes clés associés
Type de contenu	Type	Nature du document
Description	Description	Description du contenu de la ressource
Relation	Relation	Référence à une ressource en rapport avec ce document
Créateur	Creator	Auteur du document
Contributeur	Contributor	Personne ayant contribué à l'élaboration de la ressource
Éditeur	Publisher	Entité responsable de la diffusion de la ressource
Droits	Rights	Information sur la propriété intellectuelle du document
Date	Date	Date de publication
Format	Format	Spécification du type digital (ou matériel) de la ressource
Identifier	Identifiant	Référence unique de la ressource dans un contexte donné
Langue	Language	Langue du document

Tableau 6.1: Les 15 attributs du Dublin Core

respectivement au Contenu, à la Propriété intellectuelle et aux indicateurs de version. Dans notre définition des attributs, nous utiliserons les termes ressource et document comme étant équivalents.

Le Dublin Core qualifié se compose donc de deux types distincts de spécifications ayant chacun une fonction distinctes :

1. Spécification d'éléments.
   Ces éléments aident à l'interprétation d'une valeur de l'élément. Ces schémas comprennent des vocabulaires contrôlés et des notations formelles ou des règles d'analyse. Ainsi, ces éléments plus spécifiques permettent de désambiguïser un attribut pour un utilisateur néophyte, par une spécification de portée plus limitée.
2. Schéma de codage
   Une valeur exprimée en utilisant un schéma de codage sera donc un jeton sélectionné à partir d'un vocabulaire contrôlé (par exemple, un terme à partir d'un système de classification ou un ensemble de rubriques) ou une chaîne formatée conformément à la une notation formelle (par exemple, "2000-01-01" comme l'expression normalisée d'une date). Si un schéma de codage n'est pas compris par un client ou un agent, la valeur peut encore être utile à un lecteur humain. La description définitive d'un schéma de codage pour les qualificatifs doit être clairement identifié et disponible pour une utilisation publique.

6.4 Les formats d'échanges bibliographiques

6.4.1 Les formats de la bibliothèque du Congrès

Les formats *MAchine-Readable Cataloging* (MARC), proposés par la bibliothèque du Congrès américain [1] et conformes à la norme ISO 2709 (voir le paragraphe 6.3.1 page 145), servent à l'échange de données bibliographiques. Ils permettent d'informatiser les catalogues de certaines bibliothèques. On distingue, dans cette famille de formats, le

1. Tous les formats d'échange de la *Library of Congress*(LOC) sont décrits en détail à l'URL `http://www.loc.gov/standards/`, accédé en ligne le 1er septembre 2012

DCMES Element	Element Refinement(s)	Element Encoding Scheme(s)
Title	Alternative	-
Creator	-	-
Subject	-	MeSH, DDC, LCC, UDC, LCSH
Description	Table Of Contents, Abstract	-
Publisher	-	-
Contributor	-	-
Date	Created, Valid, Available Issued, Modified, Date Accepted Date Copyrighted, Date Submitted	DCMI Period, W3C-DTF
Type	-	DCMI Type Vocabulary
Format	- Extent Medium	Audio, image, video, text... - -
Identifier	Bibliographic Citation -	- URI
Source	-	URI
Language	-	ISO 639-2, RFC 3066
Relation	Is Version Of, Has Version Is Replaced By, Replaces Is Required By, Requires Is Part Of, Has Part Is Referenced By, References Is Format Of, Has Format Conforms To	URI
Coverage	Spatial Temporal	DCMI Point, ISO 3166, DCMI Box, TGN DCMI Period, W3C-DTF
Rights	Access Rights	-

Tableau 6.2: Tableau du Dublin Core qualifié

Source de : file:///Users/admin_spi/Desktop/MODS.xml

```
1  <?xml version="1.0"?>
2  <modsCollection
3          xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-2.xsd"
4          xmlns="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
5    <mods>
6      <titleInfo>
7        <title>A model of cross language retrieval for IT domain papers through a map of ACM Computing Classification System</title>
8      </titleInfo>
9      <typeOfResource>undefined</typeOfResource>
10     <genre authority="local">conferencePaper</genre>
11     <name type="personal">
12       <namePart type="family">Kembellec</namePart>
13       <namePart type="given">Gérald</namePart>
14       <role><roleTerm type="code" authority="marcrelator">aut</roleTerm></role>
15     </name>
16     <name type="personal">
17       <namePart type="family">Saleh</namePart>
18       <namePart type="given">Imad</namePart>
19       <role><roleTerm type="code" authority="marcrelator">aut</roleTerm></role>
20     </name>
21     <name type="personal">
22       <namePart type="family">Sauvaget</namePart>
23       <namePart type="given">Catherine</namePart>
24       <role><roleTerm type="code" authority="marcrelator">aut</roleTerm></role>
25     </name>
26     <part>
27       <extent unit="pages">
28         <start>162</start><end>168</end>
29       </extent>
30     </part>
31     <originInfo>
32       <place><placeTerm type="text">Ouarzazate</placeTerm></place>
33       <publisher>IEEE</publisher>
34       <dateCreated>2009-09-22</dateCreated>
35       <issuance>monographic</issuance>
36     </originInfo>
37     <identifier type="isbn">1424437563</identifier>
38     <relatedItem>
39       <titleInfo>
40         <title>Multimedia Computing and Systems, 2009. ICMCS '09. International Conference on</title>
41       </titleInfo>
42     </relatedItem>
43   </mods>
44 </modsCollection>
```

Vocabulaires MODS

Figure 6.1: Article décrit au format MODS, adaptation simplifée du marcXML

MARC traditionnel qui n'est pas intuitivement compréhensible, des versions XML dont la logique structurelle est plus accessible à l'esprit humain [1] :

- Le MARC traditionnel est décliné selon les pays.
- Le MARC21 - Fusion des principales moutures anglo-saxonnes du MARC.
- L'UNIMARC (*UNIversal MARC*) est un format international développé et maintenu par l'IFLA [2].
- MarcXML - Permet de représenter sous forme XML l'ensemble des champs du format MARC21.
- Le MODS (*Metadata Object Description Schema*) est une adaptation du format MARC XML simplifié et orienté usage avec l'inclusion de possible du Dublin Core (voir figure 6.1).

1. Le bibliothécaire suisse Pierre Gavin propose une méthode pour afficher le MARC21 de manière compréhensible par l'homme grâce à une feuille de style : `http://www.pierregavin.ch/documents-1/5-formats/marc21-iso-2709-marcxml-xslt`, accédé en ligne le 1er septembre 2012.

2. *International Federation of Library Associations and Institutions* : `http://www.ifla.org/`, accédé en ligne le 1er septembre 2012.

Les communautés de bibliothèques cherchent à s'émanciper des formats MARC vieillissants, pour s'orienter vers des formats séparant données et description, typiquement les formats XML comme le MarcXML et le MODS.

6.4.2 BibTEX

Lorsque l'on rédige un grand nombre de documents avec LaTEX, la rédaction des bibliographies peut vite apparaître fastidieuse quand on doit, pour chaque article, reporter à la main dans le fichier .tex ses références, et ce d'autant plus que les mêmes documents relatifs à une problématique sont régulièrement cités dans la production écrite d'un chercheur. Avec BibTEX un seul fichier, que le chercheur enrichit au fur et à mesure de sa carrière, est utilisé pour l'ensemble de ses écrits.

Bref historique du BibTEX

BibTEX est un format de notation de bibliographie conçu par Oren Patashnik en 1985 en collaboration avec Leslie Lamport. BibTEX est le plus souvent associé au format et au logiciel libre LaTEX. LaTEX est principalement utilisé en sciences mathématiques ou informatiques pour sa capacité à générer des formules ou des tableaux de manière très fine en séparant le fond de la forme (Florczak, 2005). Avec les programmes classiques de traitement de texte, l'auteur définit la mise en page du document de manière interactive pendant la saisie du texte. À contrario, les formats BibTEX et LaTEX sont des fichiers de textes bruts qui seront compilés conjointement. Dans ce cadre, le fond est donc bien séparé de la forme, comme avec les couples HTML et CSS ou XML et XSL. Le format BibTEX (associé à LaTEX) est très répandu dans les milieux scientifiques et mis à disposition nativement sur quasiment toutes les plateformes de diffusion de littérature scientifique en informatique, médecine, mathématiques, chimie et autres sciences dites « dures ».

Terminologie du format BibTEX

Ce format possède une terminologie et une syntaxe propre qui en font un langage de stockage de données complet et complexe. Le langage BibTEX, portant uniquement sur les données, ne permet pas de traiter de la présentation des références. L'affichage de la bibliographie passe par un vocabulaire de présentation en annexe. C'est la compilation

du fichier LaTEX qui mettra en relation le texte, la bibliographie et la feuille de style. La mise en page sera automatisée avec un style particulier et les appels de bibliographie seront hyperliés à la bibliographie. Lors de la lecture du fichier au format électronique des liens cliquables seront visibles pour faire des aller-retours entre le contenu du document et la bibliographie (Markey, 2009, Patashnik, 1988, Peyre, 2007).

Les types de documents descriptibles au format BibTEX

Comme le montre le tableau 6.3 résumé de la partie bibliographie du cours de Denis Bitouzé (2012), il existe un vaste panel de documents auxquels il est possible de se référer dans un écrit.

Exemple de notice d'article scientifique au format BibTEX

```
@inproceedings{kembellec2009model,
        title     = {A model of cross language retrieval for IT
                     domain papers through a map of ACM
                     Computing Classification System},
        author    = {Kembellec, G\'erald and Saleh, Imad
                     and Sauvaget, Catherine},
        publisher = {IEEE},
        booktitle = {International Conference on Multimedia
                      Computing and Systems, 2009},
        year      = {2009},
        isbn      = {978-1-4244-3756-6},
        doi       = {10.1109/MMCS.2009.5256709},
        pages     = {162--168}
}
```

La bibliographie se trouve dans un fichier séparé au format BibTEX [1]. Pour une meilleure compatibilité des caractères accentués et des acronymes, il est préférable d'intégrer les contenus des champs entre accolades (comme dans l'exemple de bibliographie de la page 153).

1. Il est également possible de positionner les éléments de bibliographie directement dans le texte en LaTEX, mais ce n'est pas considéré comme une bonne pratique (Bitouzé et Charpentier, 2010).

Dénomination	Description	Champs obligatoires	Champs optionnels
article	Article de journal ou de revue	author, title, journal, year	volume, number, pages, month, note
book	Livre	author, editor, title, publisher, year	volume, series, address, edition, month, note, pages
booklet	Document imprimé, sans éditeur ou institution	title, author	howpublished, address, month, year, note
proceedings	Actes de conférence	title, year	editor, publisher, organization, address, month, note
inproceedings	Article de conférence	author, title, booktitle, year	editor, pages, organization, publisher, address, month, note
inbook	Une partie d'un livre souvent un chapitre	author, editor, title, pages, publisher, year	volume, series, address, edition, month, note
incollection	Partie d'un livre qui possède son propre titre, plus grand qu'un chapitre	author, title, booktitle, year	editor, pages, organization, publisher, address, month, note
manual	Documentation technique	title, author	organization, address, edition, month, year, note
mastersthesis	Une thèse de Master	author, title, school, year	address, month, note
phdthesis	Une thèse de doctorat	author, title, school, year	address, month, note
techreport	Rapport technique, publié par une école ou un autre institution	author, title, institution, year	type, number, address, month, note
patent	Brevet ou demande de brevet	nationality, number, year, yearfiled	address, type, number, day, dayfiled, month, monthfiled
unpublished	Document non publié avec un auteur et un titre	author, title, note	month, year
misc	Documents qui ne correspondent à aucune des catégories ci-dessus	aucun	author, title, howpublished, month, year, note

Tableau 6.3: Typologie des documents décrits au Format BibTEX

Conclusion sur le BibTEX

Malheureusement la prise en main du LaTEX, bien que démocratisé par son portage sur MAC OS et Windows et des interfaces quasi WYSIWYG [1], reste peu aisée pour un utilisateur issu des SHS. De ce fait jusqu'à il y a peu, le format d'échange et de stockage bibliographique BibTEX était l'apanage d'une partie des STM : les sciences dites dures comme les mathématiques, l'informatique, la chimie ou la biologie. Il existe depuis l'arrivée de la suite Microsoft Office 2007 des solutions pour les chercheurs pluridisciplinaires permettant de transformer une bibliographie du format BibTEX vers un autre format compatible avec l'éditeur de texte Microsoft Word ou LibreOffice : le format bibWord.

6.4.3 bibWord, le format bibliographique de Microsoft Word

Depuis la version Office 2007 de a suite de bureautique Microsoft, comprenant le célèbre logiciel de traitement de texte Word, les bibliographies peuvent être gérées de manière automatisée. Grâce à l'arrivée de l'OpenDocument et l'intégration de ce format par la firme Microsoft dans sa suite Office (Office Open XML), la tâche de référencement des enseignants chercheurs, documentalistes et étudiants se trouve simplifiée.

Office Open XML est une norme ISO/IEC (IS 29500 créée par Microsoft, destinée à répondre à la demande d'interopérabilité dans les environnements de bureautique et à concurrencer la solution d'interopérabilité OpenDocument. Ce format est normalisé par l'ISO en collaboration avec l'ECMA [2].

Les notices bibliographiques peuvent être intégrées manuellement grâce à l'interface dédiée à cet effet dans le menu bibliographie de Word. Les notices sont ensuite enregistrées dans le répertoire dédié au paramétrage de Word de l'espace personnel de l'usager, qu'il soit utilisateur d'un système d'exploitation Microsoft ou Apple. Si nous observons l'exemple de bibliographie au format Office Open XML en figure 6.2, nous observons

1. *What you see is what you get*, Interface graphique qui offre le rendu final en temps réel.
2. Respectivement : `http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=61750` et `http://www.ecma-international.org/memento/TC45-M.htm`, accédés en ligne le 1er septembre 2012.

Source de : file:///Users/admin_spi/Desktop/fichiers%20XML%20ou%20BIB/sourcesWORD.xml

```
1  <?xml version="1.0"?>
2  <b:Sources
3      SelectedStyle=""
4      xmlns:b="http://schemas.openxmlformats.org/officeDocument/2006/bibliography"
5      xmlns="http://schemas.openxmlformats.org/officeDocument/2006/bibliography">
6      <b:Source>
7          <b:SourceType>JournalArticle</b:SourceType>
8          <b:Tag>Kembellec2010OntologyNavigator</b:Tag>
9          <b:Title>OntologyNavigator: WEB 2.0 scalable ontology based CLIR portal to IT scientific corpus for researchers</b:Title>
10         <b:Year>2010</b:Year>
11         <b:Author>
12             <b:Author>
13                 <b:NameList>
14                     <b:Person><b:Last>Kembellec</b:Last><b:First>Gérald</b:First></b:Person>
15                     <b:Person><b:Last>Saleh</b:Last><b:First>Imad</b:First></b:Person>
16                     <b:Person><b:Last>Sauvaget</b:Last><b:First>Catherine</b:First></b:Person>
17                 </b:NameList>
18             </b:Author>
19         </b:Author>
20         <b:Pages>115--119</b:Pages>
21         <b:JournalName>International Journal of Design Sciences and Technology</b:JournalName>
22         <b:Volume>16</b:Volume>
23     </b:Source>
24     <b:Source>
25         <b:SourceType>ConferenceProceedings</b:SourceType>
26         <b:Tag>kembellec2008ontologie</b:Tag>
27         <b:Title>Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques</b:Title>
28         <b:Year>2008</b:Year>
29         <b:Author>
30             <b:Author>
31                 <b:NameList>
32                     <b:Person><b:Last>Kembellec</b:Last><b:First>Gérald</b:First></b:Person>
33                 </b:NameList>
34             </b:Author>
35         </b:Author>
36         <b:Pages>213-231</b:Pages>
37         <b:BookTitle>Terminologie et Ontologie : Théories et applications, Actes de la deuxième conférence</b:BookTitle>
38     </b:Source>
39 </b:Sources>
```

première notice

deuxième notice

Figure 6.2: Bibliographie Word au format Office Open XML

que dans le système de balisage, une bibliographie est notée **<sources>** (au pluriel) et qu'un élément de bibliographie est noté **<source>** au singulier. Le fichier nommé « *sources.xml* » par convention contient l'ensemble des données bibliographiques de l'utilisateur. Ce dernier pourra grâce à son traitement de texte intégrer des appels à la bibliographie, qui se générera automatiquement dans le style défini. La démarche d'intégration manuelle des références est chronophage, Word ne proposant pas de filtre d'import, il faut donc saisir manuellement les données dans un formulaire.

Cependant, si l'usager possède une bibliographie au format BibTEX, il lui est possible de la formater en XML OpenDocument grâce au logiciel BibDeskToWord et d'écraser le fichier « *sources.xml*[1] ». Cette méthode requiert toutefois de l'utilisateur des aptitudes avancées aux manipulations d'outils informatiques.

1. Méthode complète : `http://mahbub.wordpress.com/2007/03/22/deciphering-microsoft-office-2007-bibliography-format/`, accédé en ligne le 1er septembre 2012.

6.4.4 Le RIS

Le RIS (*Research Information Systems*) est un format de métadonnées bibliographique initialement développé par la société éponyme [1], pour permettre aux logiciels de gestion de bibliographies d'échanger des données avec les éditeurs de texte ou les traitements de textes et les bases de connaissances.

Bref historique du RIS

La société RIS qui possédait le logiciel de gestion bibliographique (LGRB) Reference Manager avait également acheté le LGRB ProCite. RIS a ensuite fusionné avec Niles logiciels, les créateurs de EndNote. Le résultat de la fusion a été la création de l'Institute for Scientific Information (ISI) ResearchSoft, une filiale de Thomson Reuters, qui produit Reference Manager, EndNote et ProCite. Les spécifications d'utilisation de ce format résolument propriétaire sont accessibles au sein d'un manuel fourni par l'éditeur (Thomson-Reuteurs, 2008). Ce format d'échange est principalement popularisé par les produits commerciaux de Thomson Reuters : RefMan et EndNote.

Terminologie du format RIS

Les principaux champs renseignés au format RIS sont :

- TY : Le type de référence ouvre obligatoirement la notice. Cet élément de la notice doit être renseigné pour spécifier le type du document. Ce choix intervient au sein d'une liste prédéfinie très fournie. Si le type de référence n'est pas reconnu à l'importation par le logiciel de gestion de références bibliographiques (LGRB), il sera étiqueté comme générique. Le typage de ce champ influera sur la manière dont les autres champs seront interprétés par le LGRB.
- ER : Dernier champ de la notice, sert à la clore et n'est donc pas valué.
- AU, A1-A_n : Auteurs, rédacteurs, traducteurs. Chaque auteur doit être sur une ligne distincte, précédé de la balise qui correspond au rôle auteur. Chaque référence peut contenir un nombre illimité d'auteurs chacun composé d'un nombre maximal de 255 caractères. La syntaxe de ce champ est la suivante : Nom, Prénom (noms complets, des initiales, ou les deux), Suffixe.

1. Research Information Systems Incorporated

- PY ou DA : L'année de publication au format numérique de 4 chiffres : AAAA.
- DO : L'identifiant d'objet numérique (DOI).
- KW : Termes clés, un par ligne.
- IS : Numéro de série (pour une revue).
- NV : Numéro de volume (pour une revue).
- PB : Editeur
- SP et EP : Page de début et de fin, ou nombre de page en cas de livre ou monographie.

L'ordre des balises est libre, sauf pour le marqueur typologique TY et celui de fin ER.

Les types de documents descriptibles au format RIS

Voici une sélection des principaux types de documents susceptibles d'être catalogués dans une bibliographie au format RIS. Nous avons sélectionné les plus couramment utilisés pour la communication scientifique [1] en SHS et STM.

- BOOK Book, Whole
- CHAP Book chapter
- CONF Conference proceeding
- GEN Generic
- ICOMM Internet Communication
- JFULL Journal (full)
- JOUR Journal
- RPRT Report
- SER Serial (Book, Monograph)
- THES Thesis/Dissertation
- UNPB Unpublished work

1. L'ensemble de la typologie RIS est disponible en annexes

Exemple de notice d'article scientifique au format RIS

TY	-	CONF
ID	-	896
T1	-	A model of cross language retrieval for IT domain papers through a map of ACM Computing Classification System
AU	-	Kembellec, Gérald
A2	-	Saleh, Imad
A3	-	Sauvaget, Catherine
PB	-	IEEE
PY	-	2009
SP	-	162
EP	-	168
SN	-	978-1-4244-3756-6
DO	-	10.1109/MMCS.2009.5256709
ER	-	

Conclusion sur le format RIS

La typologie décrite et illustrée dans ce paragraphe propose un modèle souple et complet de description de références bibliographiques. Les éléments au sein de ces fichiers ne sont pas liés par une syntaxe complexe à comprendre pour un béotien, comme ce peut être le cas en XML ou BibTEX. L'utilisation de ce format de stockage et d'échange de références bibliographiques est très lié au type de logiciel de LGRB employé. Cependant, même sans posséder une licence d'un logiciel utilisant par défaut et nativement le RIS, il est courant d'employer ce format car les logiciels payants de Thomson Reuters RefMan et EndNote ont longtemps été les plus côtés par les documentalistes de l'enseignement supérieur et de la recherche. De ce fait, les notices à ce format se sont répandues et sont devenues une norme tellement incontournable, que la communauté de développeurs de LGRB libre a du intégrer des passerelles d'import et d'export entre ce format et leurs outils. Il est raisonnablement possible d'affirmer que le RIS partage avec BibTEX une compatibilité bibliographique universelle.

6.5 Les styles de bibliographies scientifiques

Un style (du latin *stilus* : façon d'écrire), est une manière d'écrire ou de parler. Ce substantif trouve une forte résonance dans le cadre de l'informatique et particulièrement dans les interfaces homme/machine. Dans le cadre d'Internet par exemple, la manière d'exprimer la charte graphique d'un hypertexte est appelée feuille de style. Il est ainsi possible de présenter de diverses manières un document unique [1]. Appliqué aux bibliographies et dans le cadre de l'information-documentation nous définissons le style de la manière suivante :

Les styles de mise en page appliqués aux bibliographies revêtent une importance capitale aux yeux des producteurs de littérature scientifique. Cette importance est donc répercutée sur les chercheurs qui doivent rendre un document non seulement épistémologiquement cadré, scientifiquement étayé, mais aussi irréprochable quant au style. De plus, l'auteur doit respecter une charte très stricte de mise en page. De manière générale, l'organisme responsable de l'édition du futur document propose un gabarit ou *template* d'article. Ce gabarit se présente sous forme de fichier de modèle de traitement de texte ou de classe LaTeX. Le plus souvent, en sciences dures le style du document est directement intégré au fichier LaTeX. La mise en page du document, comme celle de la bibliographie est accompagnée, voire automatisée. En sciences humaines et sociales, le contexte est bien différent. Si l'éditeur propose un fichier de gabarit aux formats *Word* ou *OpenOffice*, l'accompagnement reste très flou pour ce qui est de la bibliographie. Bien sûr, le type de bibliographie souhaité est identifié, soit par sont titre, soit par une description fine du résultat attendu. Cependant, si la méthode de présentation n'est pas parmi celles proposées par le traitement de texte, il est rare que la feuille de style associée soit « offerte » en annexe. Cet outil associé aux pratiques de mise en page est cependant indispensable pour une mise en page automatisée des références et de la bibliographie. L'auteur a alors le choix de chercher un fichier de mise en page compatible sur Internet ou de l'écrire lui-même. Il arrive qu'un autre enseignant chercheur membre d'une communauté comme Zotero par exemple, ait écrit une feuille de style compatible et l'ait partagée. Pour ce qui est de la création d'une feuille de style,

1. L'exemple parfait pour illustrer ce propos est le cite web Zen Garden qui propose de modifier de manière radicale l'affichage d'une même page : `http://www.csszengarden.com/tr/francais/`, accédé en ligne le 1er septembre 2012.

Kolski C., *Interfaces homme-machine*, Paris, Hermès, 1997.

Demeure I., Farhat J., « Systèmes de processus légers : concepts et exemples », *Technique et Science Informatiques*, vol. 13, n° 6, 1994, p. 765-795.

Lallouet A., « DP-LOG : un langage logique data-parallèle », *Actes des 6e journées francophones de programmation logique et programmation par contraintes JFPLC'97*, Orléans, 26-28 mai 1997, Paris, Hermès, p. 53-68.

Braun T., Diot C., Hoglander A., Roca V., An experimental user level implementation of TCP, Rapport de recherche n° 265, septembre 1995, INRIA.

Nawrocki A., Contribution à la modélisation des câbles monotorons par éléments finis, Thèse de doctorat, Université de Nantes, 1997.

Figure 6.3: Style Lavoisier

il s'agit d'une entreprise qui ne s'improvise pas, surtout sans compétences pointues en informatique, plus précisément en feuilles de style associées au XML. Dans le cas où le chercheur n'aurait pas les connaissances (peu répandues en SHS...) lui permettant de créer une feuille de style compatible et s'il ne la trouve pas sur internet, il ne lui reste qu'une seule solution : intégrer et numéroter les citations manuellement. De plus, il lui faudra également « rédiger » la bibliographie lui même. Le terme de rédaction est terminologiquement inadéquat pour un travail de mise en forme. Nous l'avons employé ici pour exposer le fait que manuellement, ce travail est loin d'être trivial. Il faut effectuer un tri sur la collection de citations, que ce soit par chronologie, ordre d'apparition dans le texte ou alphabétique. Selon le style, il faudra également passer les noms de famille en petites minuscules. Une fois la rédaction de la bibliographie terminée, elle ne pourra pas être transposée directement dans une autre revue, car les styles différent d'une revue à l'autre, même dans le même champ disciplinaire.

6.5.1 Exemple d'instruction bibliographique

Si nous prenons le cas des instructions données par la revue française Hermes [1], la bibliographie sera composée en Times New Roman 9 romain, interligné 11 points, les références sont rassemblées en fin d'article par ordre alphabétique, espacées les unes des autres de 6 points. Leur référence est du type (Kolski, 1997) pour un auteur, (Kolski

1. Instructions présentées à l'url `http://www.hermes-science.com/fr/cons_ouvr.html`,accédé en ligne le 1er septembre 2012.

et al.,1998) pour plusieurs auteurs. Elles sont justifiées avec un alinéa négatif de 5 mm (Format>Paragraphe>Retrait de 1re ligne négatif de 0,5 cm).

- Pour les ouvrages : titre en italique, le reste en romain.
- Pour les revues et actes de conférences : titre de l'article entre guillemets, titre. de la revue ou de la conférence en italique, le reste en romain.
- Pour les rapports internes et les thèses : texte tout en romain.

La figure 6.3, donne un aperçu du rendu de la bibliographie. Le problème est que cet exemple, pourtant tiré de la page dite « Feuille de style pour Word/OpenOffice » ne respecte pas scrupuleusement la description associée. En effet, le nom d'un des auteurs se retrouve en petites majuscules alors que les autres sont en minuscule. Cela est sans doute du à un copier-coller depuis une mise en page différente. Cela prouve à quel point la mise en forme bibliographique non assistée est une expérience périlleuse. Cependant, en cas d'usage du fichier LaTEX, également proposé sur le site en alternative aux modèles de traitement de texte traditionnels, la mise en forme bibliographique n'est pas seulement assistée, elle est automatisée. Ce type d'erreur ne devrait logiquement pas exister avec une mise en page LaTEX.

6.5.2 Les principaux styles bibliographiques

Il existe un grand nombre de styles typographiques de mise en page pour présenter la bibliographie d'un document. Chaque éditeur doit choisir le style de bibliographie associé à une revue, un livre ou à des actes de conférences. Le choix s'offre alors à lui de se baser sur les modèles existants ou d'en créer un. Voyons quelques exemples de styles universellement reconnus et utilisés, compatibles avec les principaux outils de gestion bibliographiques et de SRI.

Dans l'exemple proposé dans la figure 6.4, nous avons utilisé la version en ligne du LGRB Mendeley (voir le paragraphe 7.2.3 page 171 dédié à l'étude de cet outil) pour proposer des éléments bibliographiques dans quelques uns des principaux formats de mise en page. Les différences d'un style à l'autre sont assez minimes, on notera des constantes comme les auteurs en premier ; puis le titre ; le nom des actes de la conférence en italique, enfin la pagination.

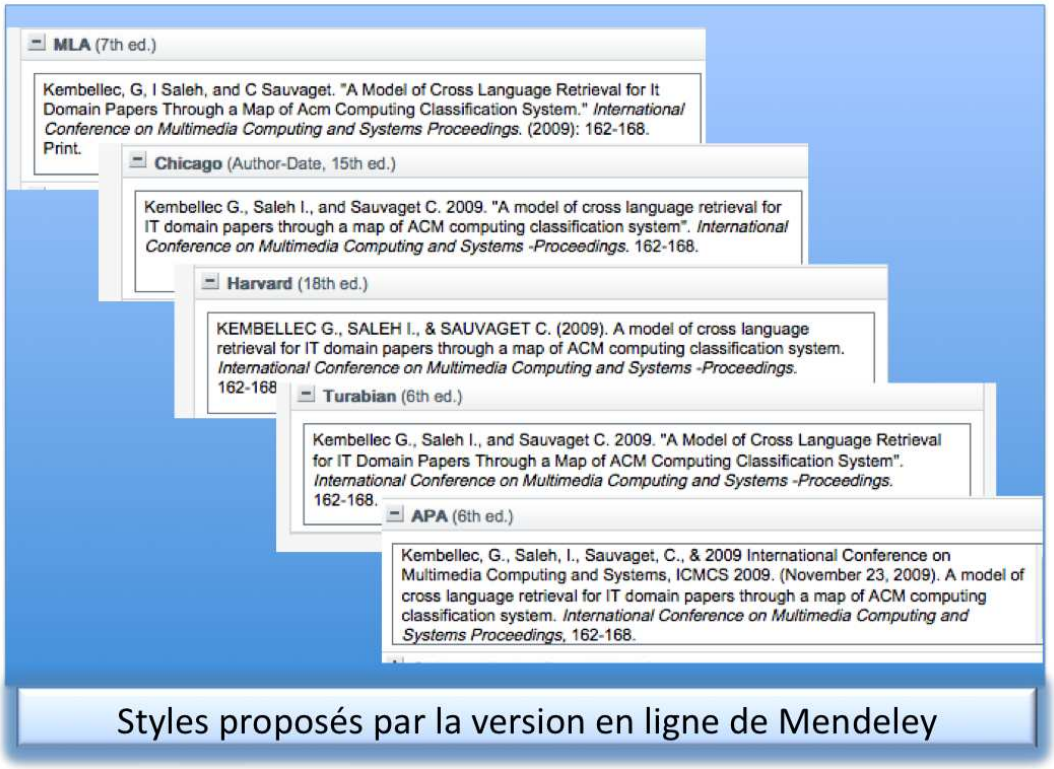


Figure 6.4: Principaux styles bibliographiques proposés par Mendeley

6.5.3 Discussion conclusive sur les bibliographiques

Nous avons présenté Kurt Schick [1] professeur de littérature à l'université James Madison dans notre partie de contextualisation (voir l'introduction page 8 et suivantes). Cet enseignant chercheur s'intéresse à la pédagogie, la stylistique et la méthodologie de recherche. Il propose sa vision de l'usage des styles bibliographiques imposés dans le cadre universitaire au travers d'un article dans « *Chronicle of Higher Education* » (Schick, 2011). Au prisme de son expérience d'enseignant, l'attention particulière des enseignants à l'usage des styles bibliographiques est scandaleusement chronophage pour les jeunes chercheurs qui délaissent le fond au profit de la forme. Cet article a fait polémique sur le blog du journal [2], avec 176 commentaires, car l'auteur désacralisait l'objet bibliographique académique, en minimisant l'intérêt du soin apporté au style bibliographique, principalement dans le cadre de l'enseignement. Nous approuvons

1. Pour plus d'information, voir `http://www.jmu.edu/learning/admin.html`, accédé en ligne le 1er septembre 2012.

2. Voir `http://chronicle.com/article/Citation-Obsession-Get-Over/129575/`, accédé en ligne le 1er septembre 2012.

l'idée que le temps consacré à la formalisation des appels bibliographiques, de la bibliographique ampute le temps total consacré à l'écriture scientifique, ou au produit documentaire. Cela vaut autant pour un documentaliste qui réalisation une synthèse documentaire que pour l'étudiant rédigeant un mémoire,ou pour le chercheur en cours d'écriture d'un article.

La formalisation bibliographique est indispensable pour la cohésion d'un ouvrage ou de tout écrit scientifique. De plus, le style bibliographique permet de distinguer les éléments de bibliographie (Type de citation, auteur, édition...) y compris dans des langues étrangères, ce qui facilité la compréhension du lecteur. Cependant, comme nous le présentons dans notre hypothèse principale (page 12), la démarche bibliographique doit pouvoir être automatisée de bout en bout, depuis l'acquisition des notices bibliographiques en contexte de recherche jusqu'à la création de la bibliographie, en passant par l'intégration dans le document de références. Pour que cette démarche soit envisageable, il faut que les populations cibles soient prêtes à utiliser les logiciels de gestion de ressources bibliographiques (LGRB), ce que nous tentons de vérifier (voir hypothèse secondaire page 12). Une autre condition nécessaire à la résolution de ce problème réside dans l'implication des instances officielles de l'enseignement supérieur et de la recherche. Les organisateurs de conférences, les éditeurs scientifiques, les responsables scientifiques universitaires doivent fournir une feuille de style dans un format compatible avec les outils récents de gestion bibliographique et les principaux logiciels de traitement (ou de compilation) de texte. Enfin, il faut que les LGRB soient compatibles à la fois avec les sources documentaires électroniques en ligne et les logiciels traitements (ou compilateurs) de texte. Examinons maintenant en détail les logiciels de gestion de références bibliographiques et leurs potentialités, notamment en terme d'automatisation du traitement de l'ensemble du processus documentaire.

Chapitre 7

Panorama des logiciels de gestion de bibliographie

Face à la croissance explosive des techniques de communication de l'information, les capacités de notre cerveau d'acquérir, de stocker, d'assimiler et d'émettre de l'information sont restées inchangées..

Pierre Joliot

Introduction

Pour la gestion des bibliographies, comme dans toute activité humaine, un processus de sérialisation et d'automatisation a été établi par la sphère informatique pour faciliter l'usage et la ré-exploitation des bibliographies. Un logiciel de gestion bibliographique est un logiciel destiné à établir, trier et utiliser des listes de citations relatives à des revues, des articles, des sites web, des livres principalement dans le cadre de publications scientifiques. Ce type de logiciel est principalement utilisé par les étudiants et enseignants et chercheurs de l'enseignement supérieur. Ces logiciels sont généralement composés d'une base de données qui peut s'alimenter de différentes façons. Une des méthodes les plus courantes est l'intégration de notices bibliographiques directement depuis les serveurs de données scientifiques. L'interaction avec ces bases de connaissances peut se faire en interrogeant directement le serveur hypertexte attaché à la base au travers d'un navigateur.

Dans le cas d'utilisation de greffons logiciels associés au navigateur, la détection de notices bibliographiques ou de documents hypertextes scientifiques avec métadonnées intégrées sera automatique. L'utilisateur pourra alors choisir d'intégrer les métadonnées dans sa bibliographie sans avoir à quitter son navigateur. Cet usage nécessite un travail de mise à disposition des métadonnées dans un des formats compatibles décrits dans le chapitre 6 sur les bibliographies. Parfois, l'éditeur du logiciel de gestion bibliographique a un accord pour interroger la base de connaissance à travers sa propre Interface Homme machine (IHM). Dans ce cas, un protocole de communication permet d'utiliser un langage de requête, avec des opérateurs, pour interroger la base. Les réponses offertes par le portail de données sont également affichées dans l'interface logicielle du logiciel de gestion bibliographique. Il est alors possible d'enregistrer tout ou partie des réponses directement dans la bibliographie.

Une autre approche de constitution de bibliographie est l'intégration de références bibliographiques uniques ou de listes déjà établies et formatées. Dans ce cas les fichiers d'origines doivent être obligatoirement établis dans un des formats détaillés dans la section 6.3 (page 145). Nous classons le copié-collé de citations dans cette catégorie d'intégration pré-formatée. Certains logiciels n'autorisent qu'un ou deux formats (souvent propriétaires), d'autres autorisent une ouverture plus large.

La dernière manière de construire une bibliographie est d'utiliser un formulaire. Tous

les logiciels permettent d'intégrer des notices de cette manière. Il s'agit de la méthode la plus précise pour un professionnel de la documentation habitué à manipuler les fiches documentaires et les subtilités associées. Cette méthode tend à disparaître, remplacée par l'automatisation logicielle. Avec la base interne de ces programmes informatiques, il est possible d'effectuer une sélection de documents par recherche à facettes. Ces outils se distinguent souvent par leur capacité à importer et exporter des formats bibliographiques différents comme RIS ou BibTeX.

7.1 Objectifs et protocole de test

Ce chapitre propose de présenter des logiciels de bureau, ainsi que les *Rich Internet Applications* RIA, dont l'objectif commun est la réalisation et l'édition de bibliographies. Cette étude comparative a pour objet de sélectionner les applicatifs les plus complets dans le cadre de notre objectif conceptuel et logiciel. Cette partie est donc cruciale pour orienter les fonctionnalités ultérieures de notre projet applicatif. Ainsi, dans le cahier des charges de notre future application, nous inclurons une compatibilité optimale avec notre sélection. Pour connaître les pratiques en matière de création assistée de bibliographies, nous avons fait appel à un panel d'utilisateurs spécialistes. Ce groupe est constitué de conservateurs et de documentalistes de *Services Communs de Documentation*[1] d'universités parisiennes, ainsi que d'enseignants-chercheurs en mathématiques, informatique, sciences de la communication et même un maitre de conférence de lettres ayant fait sa thèse en LaTEX. Nous les avons interrogé pour connaître leur utilitaire de gestion bibliographique de prédilection et les usages associés. Dans l'optique de mener notre étude comparative avec impartialité, nous avons mis au point un protocole d'évaluation. Pour juger de la valeur d'un outil de gestion bibliographique, nous avons, sur une URL donnée, intégré des données bibliographiques sous différents formats. Les formats testés sont :

- Le BibTEX ;
- Le Dublin Core embarqué sous forme de métadonnées dans le *Header*[2] ;
- Le Z39.50.

1. Centres documentaires universitaires.
2. <HEAD /> est une balise du langage HTML permettant de positionner les métadonnées de la page.

Grâce à cette URL, nous pourrons jauger la compatibilité du logiciel avec la navigation en mode glanage d'information. Nous avons également prévu de tester les logiciels en mode moissonnage. Pour y parvenir nous proposerons aux logiciels de charger un panel de documents scientifiques au format PDF. Le corpus est composé de PDF bien formés et correctement annotés au niveau métadonnées. Les logiciels que nous sélectionnerons seront à même de générer une bibliographie aux notices soigneusement renseignées à partir de ce corpus.

7.2 Les logiciels « lourds » de gestion bibliographique

Un logiciel « lourd », en terme d'interface hommes machines, est une interface permettant de piloter une base de connaissances locale ou distante, que ce soit en consultation ou en écriture. Le code de l'application s'exécute sur une machine locale ou sur un serveur d'applications après avoir été chargé en mémoire. Ce code peut soit être interprété, soit compilé et exécuté. Il s'agit dans tous les cas d'un logiciel. Dans cette sous partie, nous allons nous familiariser avec des outils pouvant être installés sur une station de travail pour gérer la bibliographie constituée au cours du processus de recherche documentaire. Il s'agit d'une alternative aux RIA que nous verrons dans un deuxième temps.

7.2.1 JabRef

JabRef est une interface graphique de gestion de bibliographie qui permet de maintenir exclusivement des bibliographies au format BibTEX. Ce logiciel est multi-plateforme car développé en java. Cette programmation autorise donc une compatibilité maximale en terme de systèmes d'exploitation. Ainsi le portage se fera aussi bien sur MAC, PC que Linux.

Historique de JabRef

Le projet JabRef est né de la rencontre de Morten O. Alver et Nizar Batada. Le premier avait développé JBibtexManager et le second BibKeeper. La première version de JabRef est né de la fusion des deux logiciels et a été disponible en 2003. Le nom JabRef est apparu comme signifiant J pour Java, a pour Alver, b pour Batada et Ref

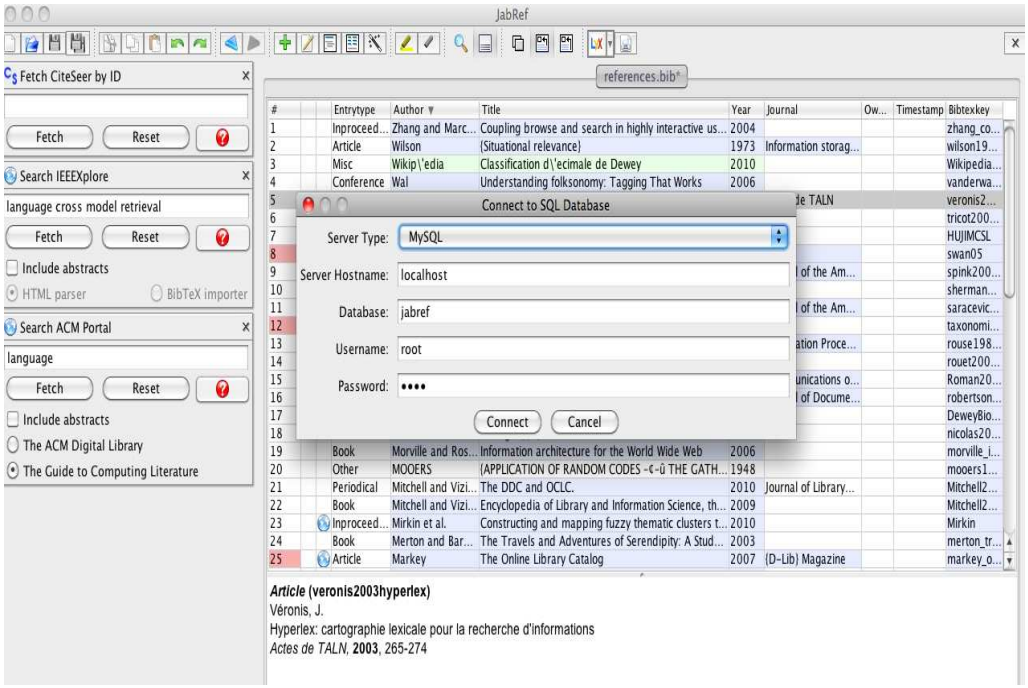


Figure 7.1: Création d'un fichier bibTEX à partir de JabRef

for Référence. Ce projet a été intégré comme un package disponible sous Linux depuis 2007.

Formats d'import de JabRef

JabRef importe principalement les notices grâce à des fichiers BibTEX et BibTEX ML, mais aussi RIS et Refer. Bien que ce logiciel s'appuie sur une interface graphique, son moteur peut être alimenté en ligne de commande. Cette opération peut trouver son intérêt lors de traitements par lot en shell script. Les adeptes de la ligne de commande, tels les membres de la communauté Linux apprécieront particulièrement cette possibilité de charger des bibliographies issues de divers fichiers, de formats différents pour créer une base unique.

Formats d'export de JabRef

Les principaux formats d'export normés supportés sont le RIS, le MODS, le RDF. Outre ces quelques formats d'export classiques, JabRef offre la possibilité de se connecter

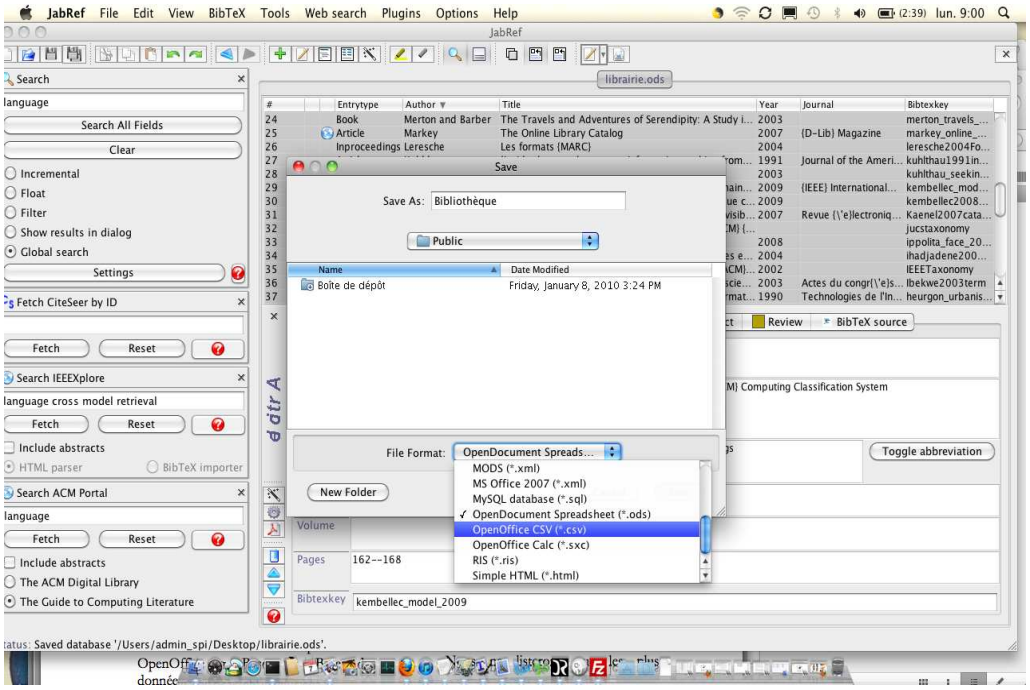


Figure 7.2: Export à partir de JabRef

à une base de données et d'exporter sa bibliographie sous MySQL. Cette possibilité est particulièrement attractive après un moissonnage ou un glanage intensif d'information avec export en format normé. Le processus est donc presque automatisé entre navigation avec le plug-in Firefox Zotero et la génération d'une base de connaissance de tous les articles intéressants sur un sujet. Il suffit pour cela de configurer une connexion à un serveur MySQL (qu'il soit local ou distant) comme l'indique la figure 7.1 page 169.

Fonctionnalités de JabRef

Pour les scientifiques en sciences dures, l'utilisation de LaTEXl'usage BibTEXavec un fichier bien formalisé est pré-requis pour soumettre des articles dans la plupart des revues. JabRef offre l'opportunité de formaliser des bibliographies de manière très stricte. En effet, si les champs obligatoires ne pas renseignés, ou sont mal renseignés, le logiciel le signale de manière claire en mettant un point rouge sur les lignes. Un autre avantage de ce logiciel est qu'il est gratuit et facile d'installation. Toulin *et al.* (2009) présentent également JabRef comme étant compatible avec plusieurs traitements de texte. Les

plus usités sont Word et OpenOffice (et sa version redevenue libre : LibreOffice) via le tableur OpenDocument au format *.ods* que JabRef est capable de générer. Le point faible de ce logiciel est le faible choix de formats d'export. Cependant, comme son usage est clairement orienté pour le couple BibTEX, LaTEX, l'inconfort reste mineur.

7.2.2 EndNote

EndNote est un des logiciels les plus anciens du marché des LGRB et donc le plus connu. Il bénéficie d'un fort recul sur la technologie et les pratiques associées à la bibliographie, notamment scientifique. Ce logiciel est édité par la société Thomson Reuters qui est également solidement implantée dans l'édition scientifique. Il a donc de nombreux partenariats commerciaux avec des sociétés d'éditions, des bases de connaissances comme SciVerse Scopus d'Elsevier, des bases de l'agrégateur EBSCO comme Francis, ou Eric. Ceci explique le fait que EndNote soit privilégié par les professionnels de la documentation. De plus, pendant longtemps, EndNote fut leader du marché en condition de quasi monopole. En important un PDF dans la base, EndNote extrait les éléments bibliographiques, crée la référence et attache le PDF. Ceci fonctionne avec les PDFs qui contiennent l'information du DOI (Digital Object Identifier) et pas ceux qui contiennent des métadonnées traditionnelles. Nous n'avons pas eu la possibilité de tester de version complète d'EndNote, nous proposons pour de plus amples informations de consulter se site de l'Urfist de Lyon. En effet ce dernier propose une introduction aux fonctionnalités avancées de l'avant dernière version d'EndNote [1].

7.2.3 Mendeley

Mendeley Desktop [2] est comme le substantif anglais « *Desktop* » l'indique un logiciel installé localement sur le disque dur de l'ordinateur. Cette application est comme les autres, spécialisée en gestion bibliographique et permet également de partager des références via son interface web.

1. Présentation de Frédérique COHEN ADAD proposée à l'URL `http://urfist.univ-lyon1.fr/servlet/com.univ.collaboratif.utils.LectureFichiergw?ID_FICHIER=1320397710677`, accédée en ligne le 1er octobre 2012.
2. `http://www.mendeley.com/`

Historique de Mendeley

Mendeley Ltd. est une société londonienne. La première mouture du logiciel éponyme a été publiée mi 2008. Cette solution est gratuite, mais pas « *Open Source*[1] ». Même si la version actuelle du logiciel est gratuite, de nouvelles fonctionnalités sont prévues sous forme de « *plugins* » payants. Le programme quand à lui doit rester gratuit.

Formats d'import de Mendeley

Mendeley Desktop permet d'importer directement les métadonnées depuis les documents, notamment les fichiers PDF. Les autres formats d'import sont l'Ovid, le RIS, le BibTEX, le EndNote XML, le txt (texte brut), mais aussi et c'est un point très appréciable la base sqlite du plugin Zotero de Firefox.

Formats d'export de Mendeley

Mendeley propose de créer des bibliographies sous divers formats, tant libres que propriétaires : BiBTEX, RIS, texte brut, XML, Zotero (zotero.sqlite), PDF. Le logiciel offre la possibilité de copier/coller des citations vers bibTEX directement de Mendeley via le presse papier du système d'exploitation. Le module de rédaction de citation peut être intégré directement dans Word et Open Office sous forme d'une barre d'outils.

Fonctionnalités de Mendeley

Ce logiciel offre la possibilité de partager des références et leurs documents attachés au sein de librairies partagées entre plusieurs utilisateurs au sein d'un même groupe. Dans ce cadre, l'espace de stockage est limité à 500 méga-octets par compte. Les groupes ne peuvent pas intégrer plus de 10 participants par « *Shared Collection* ». De plus, dans ce cadre la détection des doublons n'est pas gérée. Des options payantes permettent d'augment l'espace de stockage. Mendeley permet de partager ses références et les documents attachés mais aussi les métadonnées personnalisées. Ce logiciel offre une gestion des fichiers PDF performante basée sur la technologie PDFN et SDK capable de lire les métadonnées des documents afin d'auto générer les notices bibliographiques.

1. L'openSource est l'autorisation des auteurs de modifier les sources d'un projet pour une redistribution.

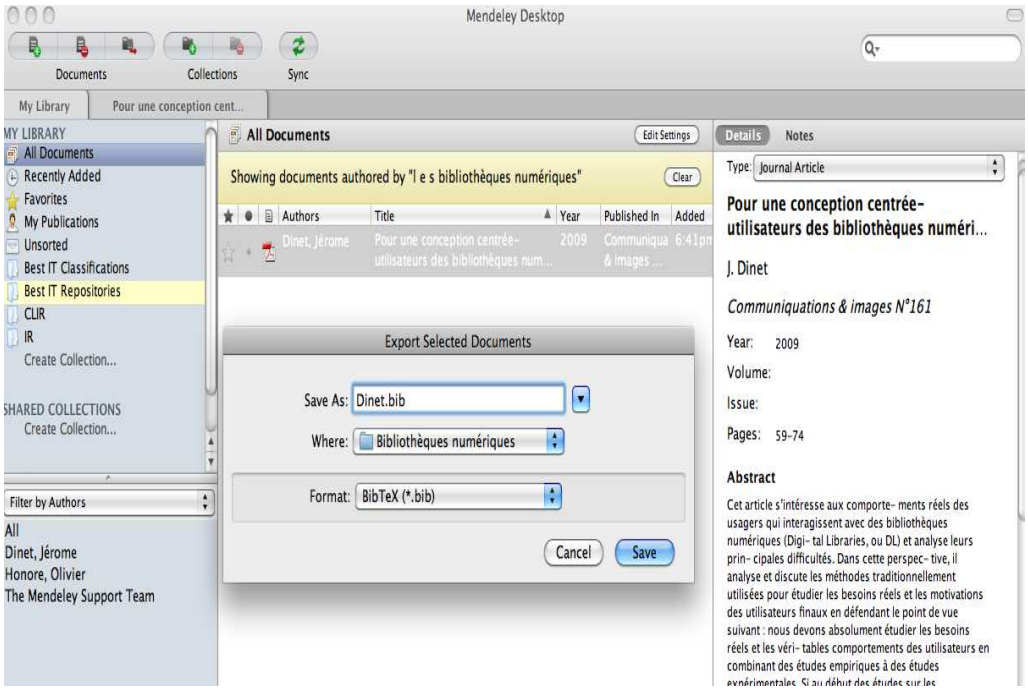


Figure 7.3: Création d'un fichier bibTEX à partir de Mendeley Desktop

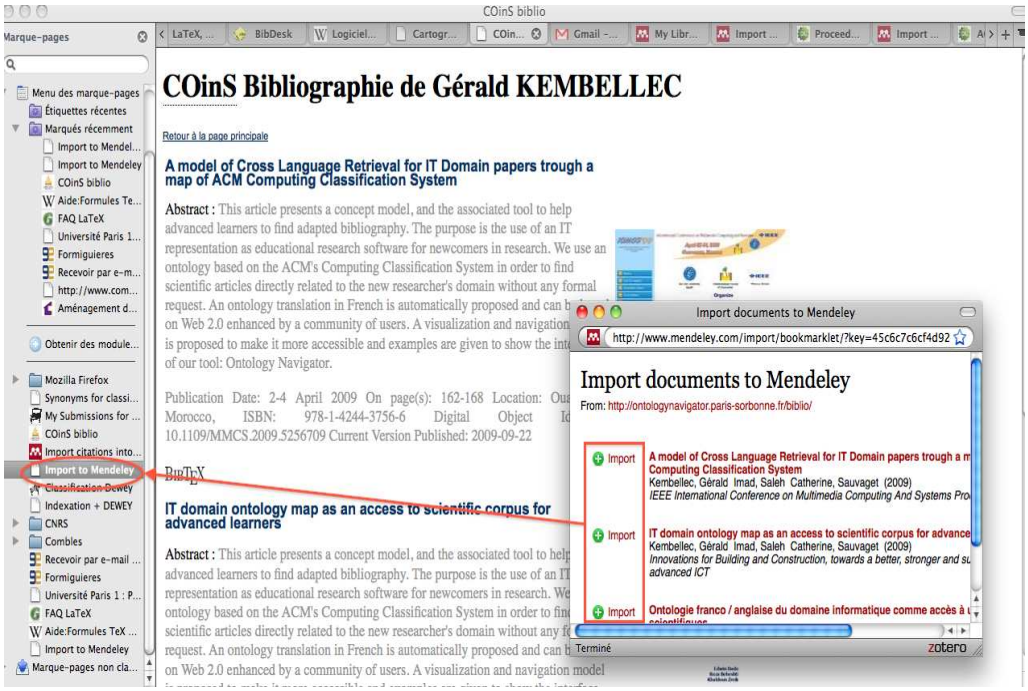


Figure 7.4: Import de notices bibliographiques avec bookmarklet vers Mendley via le navigateur web.

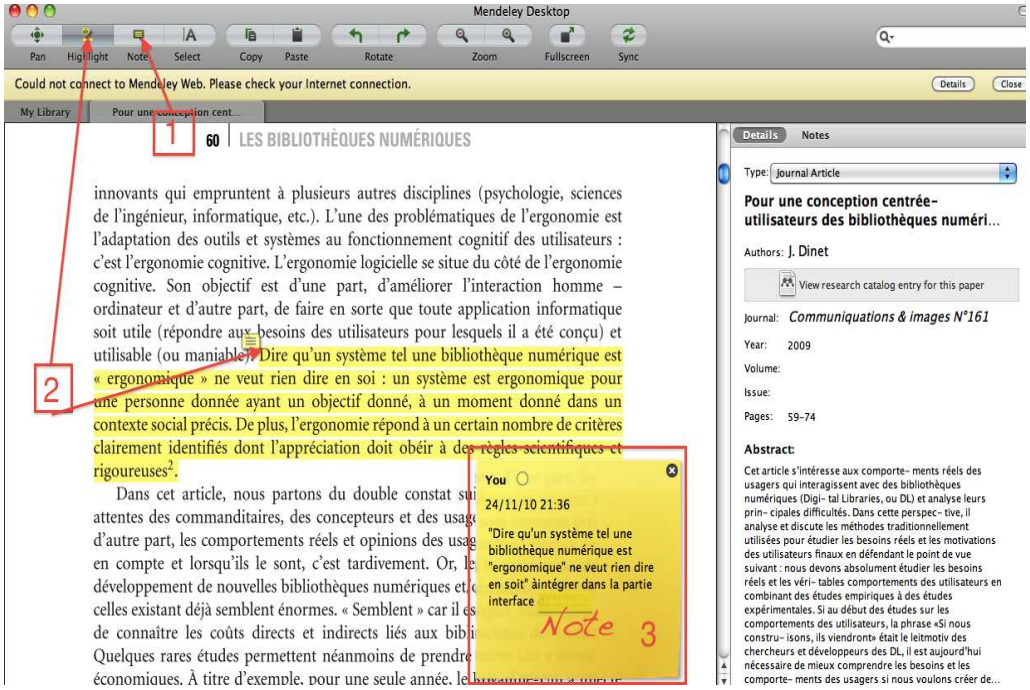


Figure 7.5: Annotation d'un document pdf par Mendley.

Cette fonctionnalité, particulièrement appréciable, nécessite tout de même une relecture systématique. En effet, selon les éditeurs de textes et les modèles fournis par les organisateurs de conférences l'organisation des métadonnées peu être altérée. Mendeley est compatible avec les logiciels MS Word et OpenOffice pour l'intégration de bibliographies.

Pas d'export direct des bases bibliographiques vers Mendeley, pas vu d'outil de création de styles. Nouvelles fonctionnalités ou capacités payantes. Un plug-in bookmarklet à intégrer au navigateur (dans les marques pages, à gauche de la figure 7.4 page 173) permet de détecter les notices au format COinS et de les intégrer à Mendeley Desktop, sous condition d'avoir un compte en ligne.

D'un point de vue fonctionnel, un élément très appréciable est la possibilité d'annoter un document PDF *au sein même* du logiciel qui réalise la bibliographie. Dans l'exemple explicité par la figure 7.5 page 174, le document de Dinet cité ailleurs, dans la thèse, a pu être annoté(1) et réutilisé ultérieurement car au cours de la lecture nous avions pu surligner(2) les passages clés et annoter le document grâce à Mendeley Desktop. Par la suite, la note apparait comme un *post-it* flottant dans le document près de son point d'ancrage(3)

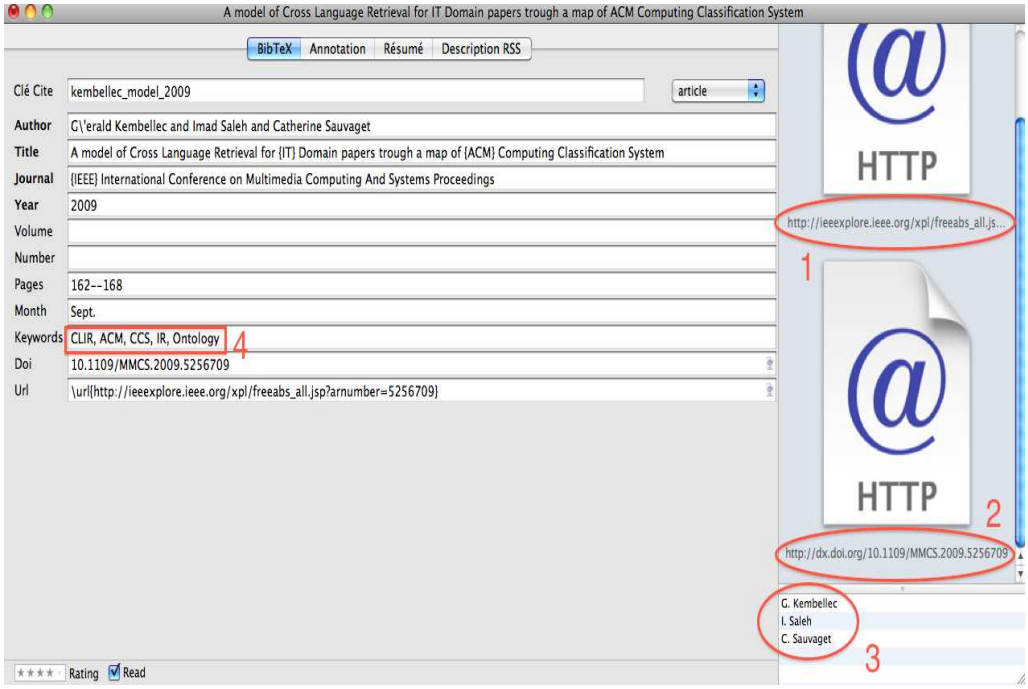


Figure 7.6: Détail d'édition d'un article dans Bibdesk

7.2.4 BibDesk

Un format d'import unique

Bibdesk est un éditeur graphique dédié aux bibliographies LaTEXc'est à dire au format BibTEX. Il ne gère donc qu'il seul type de fichier et donc d'extension, le *.bib*. Cette volonté est clairement définie dans le nom du logiciel. Le segment de clientèle de cet outil est, comme pour LaTEX, les mathématiciens, les scientifiques ou encore ceux qui ont des travaux de mise en page avancée avec des polices de caractère intégrant des langues mortes ou rares. L'interface est simple, l'illustration 7.6 page 175 montre l'édition d'un article avec sur la partie de droite les hyperliens vers DOI (2) et la bibliothèque numérique (1) ou l'article est officiellement référencé. Les auteurs sont également clairement identifiés en bas à droite (3). La partie de gauche est réservée aux métadonnées de l'article, dont les mots clés qui revêtent une importance particulière (4).

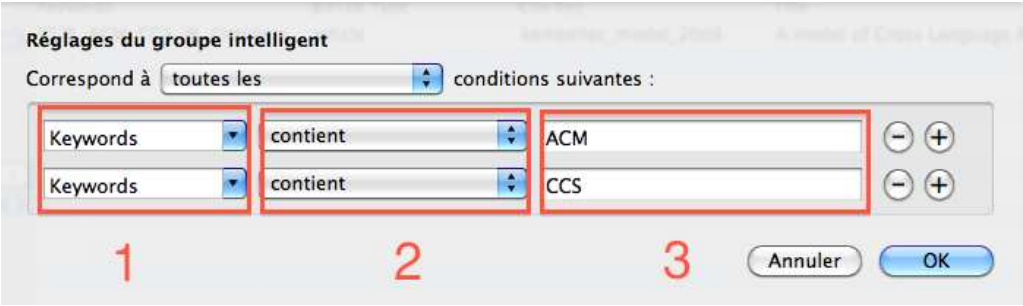


Figure 7.7: Politique de filtre de classement des entrée bibliographique de BibDesk, ici sur les mots clés

Formats d'export de BibDesk

Le premier et principal format d'export de BibDesk est bien sur le BibTEX, avec une option pour ne pas intégrer les mots clés et résumés au fichier bibliographique. Cela allège le document et limite les possibilités d'erreur à la compilation du document LaTEX. En effet, pour mémoire tous les caractères accentués doivent être re-formatés pour être valides. Cette option de « nettoyage » n'est pas négligeable lors de la compilation d'une thèse intégrant plus d'une centaine de citations. Un temps conséquent de déboguage peut ainsi être dégagé. BibDesk propose égale d'exporter des notices bibliographiques ou des bibliographies complètes dans les formats RIS,

Fonctionnalités complémentaires de BibDesk

Ce logiciel ne s'arrête pas à l'édition des notices bibliographiques contenues dans un une bibliographie au format LaTEX, il intègre outre un système d'auto-complétion des mots clés, des fonctionnalités de classement « intelligent » grâce à ces mêmes mots clés. Les politiques de classement s'effectuent également à partir de filtres sur des opérateurs de type « est » , « contient » et peuvent s'appliquer sur n'importe quel champs BibTEXde l'auteur au titre en passant par es mots clés. Dans l'exemple de la figure 7.7, page 176 les éléments 1 représentent les champs tester avec l'opérateur 2 et les arguments 3. Cela peut être traduit par « création d'un groupe filtrant les documents dont les mots clés sont ACM ET CCS. »

BibDesk possède son propre moteur d'indexation au sein duquel il est possible de définir des règles de classement en « groupes intelligents ». Le moteur de recherche de

Figure 7.8: Détection automatique de notices bibliographiques BibTEX et COinS dans un site Web par BibDesk

BibDesk permet de parser, outre des fichiers bibliographiques au format BibTEXbon nombre de bases de connaissances prédéfinies. Parmi ces bases de connaissances citons :

- Des sites à l'accès gratuits en consultation comme ACM (méta données et résumés en accès libre), arXiv, CiteULike, Google Scholar, IACR (Cryptology) ;
- Des sites par abonnement comme : IEEE Xplore, MathSciNet, Project Euclid, SpringerLink, Zentralblatt Math ;
- Mais aussi, les sites respectant les normes bibliographiques : BibTeX, COinS, HCite.

Cette option permet d'utiliser BibDesk comme un navigateur web dédié à la recherche scientifique et d'intégrer à la volée les notices bibliographiques à sa bibliographie, voir figure 7.8 page 177.

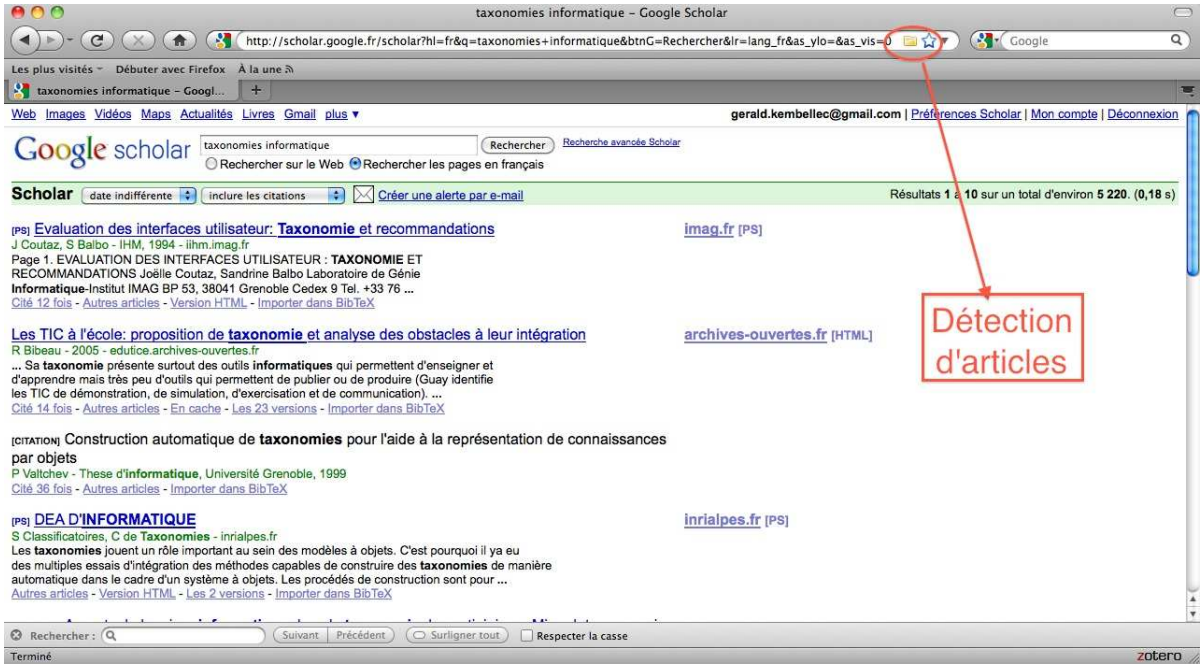


Figure 7.9: Détection massive d'articles par Zotero

7.2.5 Zotero, le module du navigateur firefox

Historique de Zotero

Zotero est un module logiciel, ou plug-in, développé le *Center for History and New Media* de la *George Mason University*. Zotero est à part dans le panorama des logiciels de gestion bibliographique. En effet, ce module s'exécute au sein du navigateur Firefox dont il est une composante optionnelle. Zotero permet de rentrer manuellement des notices bibliographiques au sein d'une base locale, de les classer et de les exporter sous différents formats bibliographiques.

Fonctionnalités de Zotero

Ce module possède également un moteur de recherche intégré pour parser les références enregistrées dans la base de l'utilisateur. Cependant, son réel intérêt est de proposer la possibilité de détecter sur le web des ressources documentaires de différents types, d'en proposer l'enregistrement en masse à la volée en cas de recherche au sein d'une base documentaire ou d'un portail. La figure 7.9 page 178 donne une exemple

de détection des métadonnées générées par le portail de recherche scientifique Google Scholar. La réponse à la requête est constituée d'éléments bibliographiques dont chacun est détectable par Zotero. L'icône Zotero de la barre de navigation signale la présence de multiples entrées bibliographiques. En cliquant sur cette icône, un pop-up propose d'enregistrer une ou plusieurs notice bibliographiques dans la base. Dans le cas de références isolées, Zotero permet également de détecter et d'enregistrer un seul élément, comme sur une page personnelle d'un chercheur qui propose la lecture de son dernier article. Dans le d'un document unique décrit sur une page Web, le plugin précisera le type du document lors de sa détection, à savoir un article , un livre ou une vidéo. Cette précision ce fera par le biais de l'icône apparaissant à droite dans la barre de navigation, par une iconographie adaptée. Par exemple un livre, un chapitre, un article d'encyclopédie sera symbolisé par un petit livre bleu.

Formats d'import de Zotero

Bien sûr, la capacité de Zotero à détecter un document ou une notice dépendra de la méthode de mise à disposition. Une explication technique détaillée sur la mise à disposition de notices et de métadonnées sera fournie plus tard dans le chapitre Urbanisation de Systèmes d'information page 223. Il faut à minima que le document soit en-capsulé de manière correcte sur la page xHTML, que ce soit en métadonnées (via les informations encodées en RDF ou en Dublin Core) dans la partie non affichée de la page (le HEADER) dans le cas d'un document unique auto-décrit ou COinS de la norme Z3950 [1], [2] dans le cas d'une base de connaissances répondant à une requête. Les résultats visibles à l'œil, c'est à dire les liens vers les ressources sont accompagnés par des métadonnées descriptives normalisées. Zotero est à même de comprendre ces métadonnées et de les proposer au lecteur, soit en import massif soit en ne sélectionnant que les documents intéressant le chercheur. Les notices glanées peuvent ensuite être éditée et classées en bibliographies. Un moteur de recherche intégré accède intuitivement au catalogue ainsi réalisé. La force de ce logiciel est d'autoriser la création et l'exportation de bibliographies aux formats d'éditeurs de textes les plus rependus avec de plus la

1. Norme Z3950 et Firefox`http://www.mozilla.org/rdf/doc/z3950.html`
2. `http://www.loc.gov/z3950/agency/`

Figure 7.10: Détection d'un document auto décrit ou d'une notice bibliographique par Zotero

mise en forme bibliographique adaptée au domaine de recherche de l'utilisateur, voire même du type de revue.

Formats d'export de Zotero

Notons que Zotero permet la génération de fichiers au format BibTeX. Il semble que le gros avantage de Zotero soit de pouvoir détecter intelligemment aussi bien les références bibliographiques que les documents eux mêmes. Zotero permet donc une chaine de production quasi automatisée allant de la détection ou l'import, jusqu'à la génération de bibliographie en de nombreux styles sous de nombreux formats. Il est même compatible, outre LaTeX via BibTeX à Word qui est l'autre outil de production scientifique. Il est même possible d'exporter une citation bibliographique par un simple cliquer déplacer vers un éditeur ou un client de messagerie.

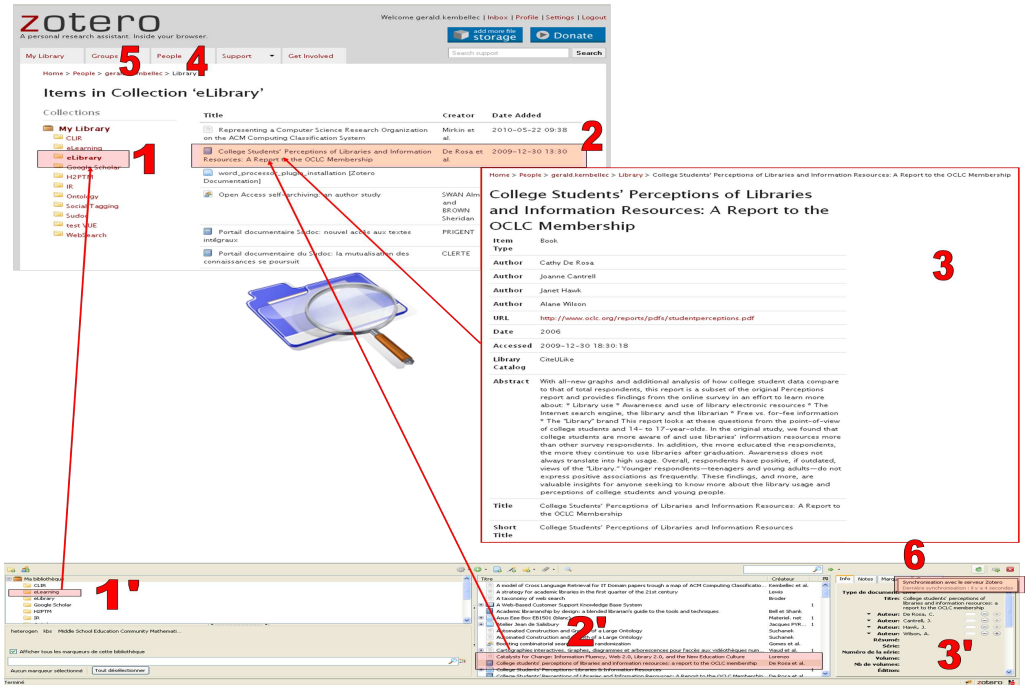


Figure 7.11: Zotero en ligne

7.3 Les applications « en ligne » de gestion bibliographique

Une *Rich Internet Application* ou RIA peut être décrite de deux manière différentes. Nous pouvons voir ce type de service comme un site internet offrant des fonctionnalités avancées d'interface homme machine. L'autre vision est à contrario de voir ce type de service comme un logiciel qui a été pensé pour fonctionner sur internet via un navigateur. Au cours de la section précédente nous avons examiné quelques applications et plug-ins. Certains de ces logiciels ont leur équivalent (même nom, mêmes fonctionnalités) sur des sites web. D'autres RIA ont été développées exclusivement pour un usage en ligne, au travers d'un navigateur.

7.3.1 Zotero, le site internet

Zotero.org possède une RIA capable de créer et gérer des bibliographies. Cet outil en ligne n'a pas pour vocation de se substituer à EndNote ou RefWork, mais plutôt de

créer un entrepôt de données auquel il est possible de se syndiquer via RSS Atom. Ce flux est visualisable au travers d'un navigateur et il est compatible avec le glanage et moissonnage d'information grâce à une API Zotero décrite dans le *x*ml name space[1] *d*e zotero. Pour bénéficier de service, il faut créer un compte et une clé privée (qui sera appelée userID et valuée de manière décimale). Le nom et l'identifiant permettront de paramétrer le plug-in Zotero pour qu'il se synchronise avec le serveur. L'intérêt de créer son compte en ligne Zotero est triple :

- sauvegarder sa base bibliographique ;
- synchroniser les bases de tous ses postes de travail ;
- partager ses références bibliographiques.

Il faut donc voir Zotero.org comme une extension du module Zotero, une option permettant de partager des notices bibliographiques, mais également de les sauvegarder. Non seulement les notices sont sauvegardées, mais également la structure des dossiers de classement (cf : Figure 7.11 repères 1 et 1'). Les notices sont présentées différemment en ligne que sur le plug-in, mais ils contiennent les mêmes informations (cf : Figure 7.11 repères 2 et 2'). Cependant la lecture est beaucoup plus aisée sur la page web. La notice détaillée (cf : Figure 7.11 repère 3) permet de visualiser le détail d'une notice avec l'hyperlien associé au document référencé s'il est disponible en ligne. dans le cas contraire, il nous est possible de disposer gratuitement d'un espace de stockage de 100 mo pour y déposer des documents à lier à la notice. Il nous est également possible de suivre les activités d'un utilisateur ou d'un groupe d'intérêt (cf : Figure 7.11 repères 4 et 5). Bien que la synchronisation entre le client et le serveur soit automatique il est également possible de l'effectuer manuellement (cf : Figure 7.11 repère 6).

7.3.2 Refworks

RefWorks est une application en ligne fonctionnelle et très complète qui s'adresse à des organismes plutôt qu'à des particuliers. Il arrive régulièrement que des universités fassent appel à cette RIA pour gérer les bibliographies des enseignants, mais aussi celles des laboratoires ou équipes de recherche.

Les SCD s'équipent également de cette interface pour proposer aux étudiants, enseignants et chercheurs une interface souple et nomade de gestion de références.

1. `http://zotero.org/ns/api`

L'Université Paris-Sorbonne (Paris 4) s'est dotée de cette solution, nous avons pu l'étudier en détail lors de notre passage.

Historique de RefWorks

La société RefWorks a été fondée en 2001 par une équipe d'experts dans le domaine de la gestion de base de données bibliographiques avec une participation de la société *Cambridge Information Group*. RefWorks déclare fournir un service de recherche et de gestion bibliographique basé sur les technologies Web. Cette offre est dédiée l'écriture collaborative pour les communautés académiques de recherche, le gouvernement et les services recherche et développement des entreprises. Cette interface est utilisée quotidiennement par des milliers de chercheurs au sein de plus de 900 organisations à travers le monde. RefWorks est à même de communiquer avec des centaines de bases de données en ligne (Vivarès, 2009). RefWorks collabore avec certains des plus importants portails d'information en ligne (ProQuest, EBSCO, Elsevier, HighWire, l'ISI, OCLC...) [1]. ProQuest, une société du consortium *Cambridge Information Group*, a d'autant des rapports étroits avec RefWorks qu'elle l'acquis en janvier 2008. Refworks a été fusionné avec *Scholar Universe*, le site de réseautage scientifique pour donner *COS Scholar Universe*, ce qui offre un accès direct potentiel à plus de 1,4 million de chercheurs dans plus de 200 disciplines.

Formats et méthodes d'import de RefWorks

RefGrab-It et le glanage contextuel RefGrab-It est une fonctionnalité optionnelle sous forme de plug-in Firefox et Internet Explorer ou de *Bookmarklet*. Ce greffon logiciel a pour optique de capturer contextuellement l'information bibliographique à partir des pages Web et de la mettre à disposition pour une intégration. RefGrab-It recherche des informations bibliographes sur les pages Web parmi aux formats suivants :

- numéro ISBN
- PubMed ID
- Digital Object Identifier (doi)
- Context Object in Span(COinS)

1. Source :`http://www.refworks.com/content/about_us.asp`

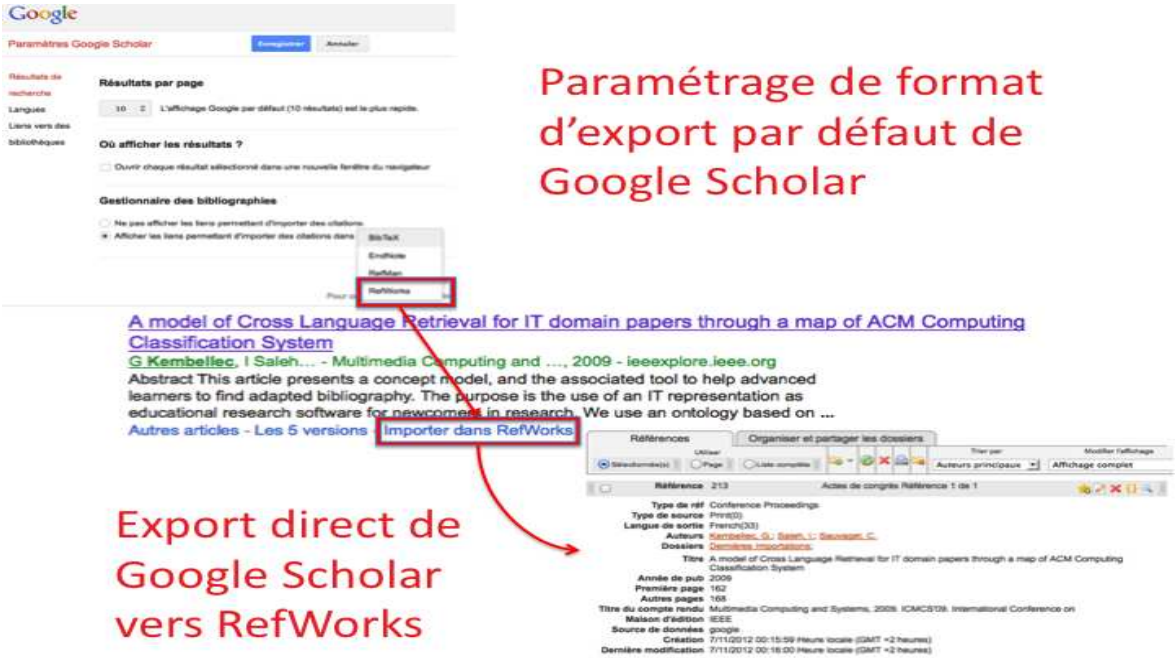


Figure 7.12: Utilisation de Google Scholar avec RefWorks

– Codage spéciaux intégrés dans la page web qui lieront directement cette référence vers RefWorks

Il prendra automatiquement que les ressources d'information sur Internet de recherche différents (les coulisses) pour obtenir des informations supplémentaires qui peuvent être d'intérêt pour vous que vous pouvez également importer. RefWorks peut même localiser les flux RSS lié à cette page Web auxquels on peut s'abonner pour obtenir des informations ultérieurement. Remarque : RefGrab-It ne peut parcourir que le code xHTML et ne pas donc pas être utilisé pour analyser les métadonnées des fichiers PDF hyperliés dans les sites Web.

Lorsque RefGrab-It détecte de la matière première documentaire, il ouvre automatiquement une fenêtre temporaire des résultats sous forme de « *pop-up* ». Il est alors possible de consulter les métadonnées des documents repérés par le système puis de décider d'importer les données ou non.

Les hyperliens de compatibilité RefWorks Un certains nombre d'éditeurs scientifiques de bases de connaissances et de portails offrent la possibilité d'intégrer directement

les informations bibliographiques de chaque fiche à travers un hyperlien spécifique explicitement affiché avec le logo RefWorks. Citons parmi les plus connus indiqués par l'éditeur :

- CAIRN
- EBSCOhost
- Google Scholar
- HubMed
- IEEE
- JSTOR
- Microsoft Academic Search (le lien RefWorks se contente d'afficher une notice en RIS)
- OCLC
- ProQuest
- ScienceDirect
- Scopus

Cette fonctionnalité est compatible avec le plugin RefGrab-It, comme vu précédemment mais permet également d'entrer une fichue dans la base bibliographique personnel avec un simple click sur un hyperlien sans installation logicielle complémentaire.

import de notices Pour importer massivement des bibliographies depuis une plateforme de contenu ou un autre outil de gestion de référence bibliographie. RefWorks accepte un vaste panel de formats de fichiers grâce à des filtres de conversion :

- BibTEX
- MARC
- Ovid
- OvidMARC
- RefWorksXML
- RIS
- UniMARC

Il est possible d'entrer une notice complète dans un de ces formats (ou de leurs nombreux dérivatifs propriétaires), dans un formulaire d'import. L'autre méthode est l'import unitaire ou massif depuis un fichier formaté dans un des ces formats. Il est vivement

conseillé de se limiter aux formats standards avec les champs had hoc. Notons que le RefWorksXML est un version propriétaire de XML dérivée du RIS.

Formats et méthodes d'export de RefWorks

RefWorks exporte ses notices sous forme d'hyperliens ou de fichiers. Il est également possible de se les faire envoyer par courriel. Les formats possibles sont les suivants :

- Refworks Export Tagged Format (fichier plat similaire au RIS)
- RIS
- Liste de citation (fichier plat non formaté)
- RefWorks (fichier XML reprenant le modèle RIS)
- BibTEX

Styles et fonctionnalités avancées

RefWorks propose des centaines de styles bibliographiques pour MS Word, dont l'APA, MLA, Chicago, Vancouver et Turabian. Les administrateurs locaux en SCD ou centre de recherche peuvent offrir leurs propres styles ou personnaliser un modèle existant en utilisant le gestionnaire de style de présentation [1]. Cette fonctionnalité est utilisée par le SCD de l'université Paris-Sorbonne qui propose à l'usager d'utiliser soit tous les styles, soit ceux acceptés par l'Université. Ce service de documentation universitaire offre à travers cet RefWorks une solution complète de gestion de bibliographe, depuis la recherche d'information (tant au sein de son OPAC que dans les bases) jusqu'à la production normée et stylisée de bibliographie. Les étudiants et les enseignants chercheurs n'ont pas dans ce cadre à se préoccuper des contingences matérielles liées à la bibliographie. Un hyperlien permet en outre de contacter par mail l'administrateur local de RefWorks, qui se trouve être un conservateur de bibliothèque.

Write-n-Cite : l'intégration de références bibliographique dans MS Word

Write-n-Cite est un plug-in offrant l'accès à la base bibliographique de RefWorks au sein d'un document rédigé avec Microsoft Word. Ce greffon permet d'insérer des appels de citation dans le texte puis de générer une bibliographie à la fin du document. Cette

1. notice explicative de l'éditeur à l'URL :`http://www.refworks.com/tutorial/RefWorks%20Basics%20Creating%20a%20Custom%20Output%20Style%20List.pdf`, consulté le 10 juillet 2012

fonctionnalité formate aussi bien les appels de citations dans le corps du texte que la bibliographie selon le style choisi.

RefWorks et mobilité

La société ne propose pas de *Widget* de gestion de références, cependant le site offre une section baptisée RefMobile [1] qui présente l'avantage d'un formatage adapté aux smartphones. Son utilisation, comme celle de la version classique, requiert l'identifiant groupe fourni par le service de documentation de l'organisme de rattachement, l'identifiant de l'utilisateur ainsi que son mot de passe. L'interface mobile est fonctionnellement réduite pour être utilisable sur un écran de petite taille. Pour une utilisation nomade, l'intégration de références bibliographies par hyperlien est toujours efficace. Il en va de même pour la recherche par interface, que ce soit dans la base du groupe, la base individuelle. Il est également possible de faire une recherche par ISSN (utile dans une bibliothèque ou une librairie), ISBN ou DOI. Cependant l'intégration manuelle de notices est malaisée. L'export de notices est sans objet sur un périphérique nomade qui n'a pas pour vocation de formater du texte.

Possibilités de recherche avancée de RefWorks

RefWorks propose une interface de recherche avancée permettant d'effectuer des recherches dans des champs spécifiques, de créer des requêtes booléennes et de limiter la recherche à un ou plusieurs sujets. Les résultats d'une recherche avancée sont affichés par auteur et classés par ordre alphabétique, avec les termes de la recherche mis en surbrillance.

Pour conclure sur RefWorks

Cette interface de gestion de bibliographies en ligne est une des solutions globale payantes les plus complètes du marché. En effet, RefWorks capitalise tous les services liés à la documentation avec une gestion centralisée accessible par un gestionnaire local désigné par la société ou l'université payant le service. Ce produit est particulièrement adapté aux centres de documentation en R&D d'entreprises, en service commun de documentation universitaire. Le fait que la souplesse de gestion soit alliée à une gestion

1. à l'URL : `http://www.refworks.com/mobile`

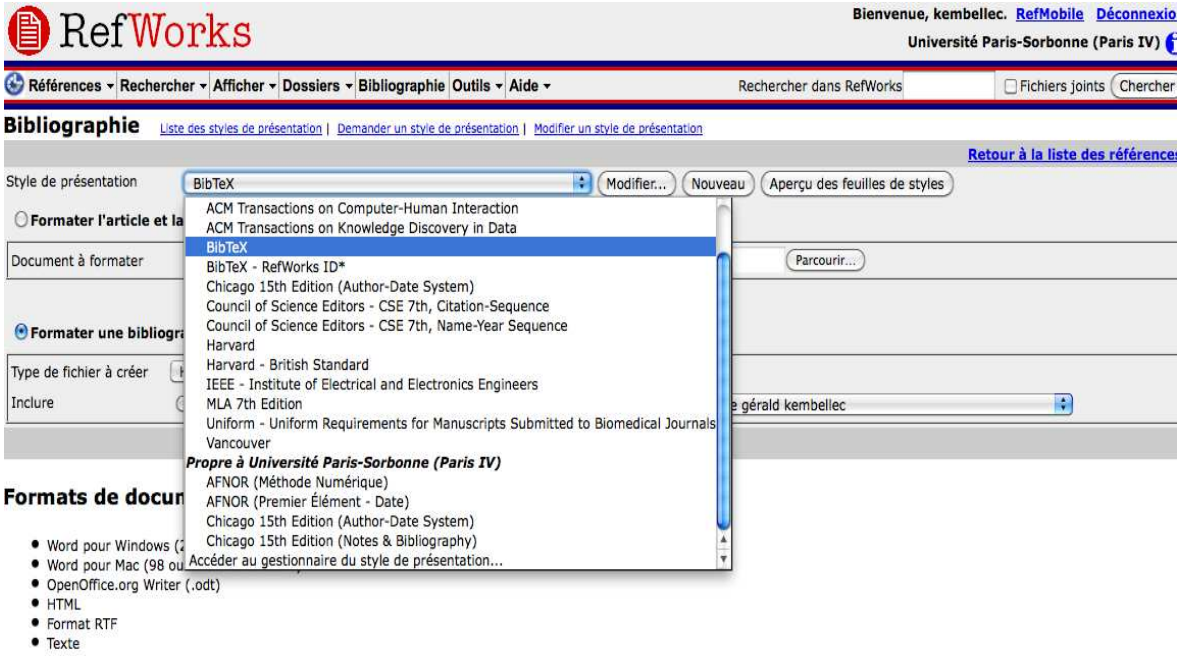


Figure 7.13: Export RefWorks

technique externalisée offre un service pérenne et n'est pas chronophage pour la structure d'accueil. De plus le vaste choix d'import et d'export, ainsi que les greffons logiciels associés rendent son usage quasi transparent une fois la prise en main effectuée. Il n'en demeure pas moins que cette solution a un coût non négligeable en terme d'abonnement et en formation pour les personnels et étudiants.

7.3.3 Discussion

L'écriture scientifique moderne s'accompagne indubitablement d'une assistance à la gestion bibliographique. Deux visions idéologiques se partagent la production de logiciels, en particulier d'écriture et gestion de référence. Ces points de vues sont liés à des visions opposées de la monétisation et des droits d'utilisation de ces programmes et de leur sources. Le monde du libre désire mettre en avant l'ouverture et le partage de la propriété intellectuelle. Cette vision s'accompagne souvent d'une volonté de gratuité, mais pas obligatoirement. Comme nous l'avons montré, l'usage d'un logiciel libre n'est pas forcément gratuit (Zotero avec l'option de stockage), de même qu'un

logiciel propriétaire n'est plus forcément payant (Mendeley). Ces évolutions en terme d'usage fait parfois régner une certaine confusion quand au choix d'un LGRB. Une certaine « bataille » entre les éditeurs et le monde du libre qui se professionnalise pour les parts de marchés d'utilisateurs de LGRB. La vielle garde d'outils professionnels payants comme Procite, EndNote ou RefWorks conserve une solide notoriété. Les sociétés et les administrations apprécient de pouvoir faire appel au service après vente en cas de besoin. Cependant, des éditeurs de logiciels libres commencent également à se positionner sur le segment service payant tout en conservant leur éthique première sur l'ouverture des sources. L'écart de qualité entre les produits d'éditeurs logiciels commerciaux et de communauté s'est considérable réduit. De même, l'interopérabilité des systèmes est devenue quasi universelle. La mentalité du « tout libre » au sein des communautés est moins forte, ce qui offre une démocratisation des produits libres. En effet, ces deux dernières années, un effort considérable a été fourni par les communautés de développeurs du libre pour offrir une compatibilité accrue entre les LGRB et les traitements de textes, notamment propriétaires. Cette volonté d'ouverture a provoqué chez les éditeurs de suites propriétaires en bureautique une réaction de transparence avec des formats « ouverts » de présentation de documents. Le fonds et la forme sont ainsi maintenant séparés dans les documents au format OpenDocument (MS Office XML). La question qui émerge de ces réflexions est celle de la rencontre entre ces logiciels aux modèles si différents en termes économiques et de structures de développement et de mise à jour. Les communautés libres sont-elles en train trouver un modèle économique viable ? Les sociétés d'édition de logiciels de production de documents et d'accompagnement bibliographique pourraient-elles trouver une parade dans l'ouverture de leur données ? Le bénéfice de cette mutation devrait se faire au bénéfice de l'utilisateur final, tant d'un point de vue qualitatif que financier.

7.4 Conclusion

L'une des questions les plus déterminantes pour l'usage des logiciels bibliographiques est l'interopérabilité entre catalogues, bases de données bibliographiques et plateformes de journaux (Vivarès, 2009). Ce questionnement s'étend à la capacité d'un logiciel à respecter les normes bibliographiques de cataloguage de données, tant en import qu'en export. En conséquence, après examens des solutions de gestion en bibliographie, un

service de production et/ou de recherche de contenus scientifique va inévitablement devoir se poser la question omniprésente sur le Web de données :

- Les notices exposées sur l'interface à l'intention de l'utilisateur sont-elles également accessible par les LGRB ?
- Le choix des formats d'export est il un enjeu majeur de la plateforme ?
- Les différents formats de notices sont-ils utilisés dans toutes les sciences ?
- Quels sont les choix stratégiques à effectuer pour offrir un service de qualité du point de vue du système d'information ?

Voici autant de points d'ombre que la partie suivante va s'efforcer d'éclaircir par l'observation des usages et pratiques courantes. Les conclusions permettront une meilleure approche des solutions techniques. Ainsi, la synthèse des usages et outils disponibles permettra de proposer une solution d'urbanisation de système d'information dans le cadre d'un SRI scientifique.

Chapitre 8

Étude d'usage de système d'information documentaire scientifique

Mais si on fonctionne au sondage, c'est le serpent qui se mord la queue, on tournera en rond, on proposera toujours aux gens ce qui leur plaît déjà.
C'est l'ennui assuré.
- Mieux vaut être ennuyeux que téméraire.

Paradis sur mesure, Bernard Werber

Introduction

Nous ambitionnons de modéliser un système de recherche d'informations (SRI) scientifique pour les enseignants chercheurs, les étudiants de troisième cycle et les personnels spécialisés en documentation qui les aident dans leur phase de recherche documentaire au sein des services communs de documentation (SCD [1]). Cette étude exploratoire vise à réunir de l'information autour des pratiques documentaires pour en faire émerger des pistes de réflexion afin d'optimiser les fonctionnalités du SRI. En effet, placer les pratiques des usagers comme fondement de la conceptualisation d'un système d'information (SI), c'est créer un SI à « valeur ajoutée globalisée Ruiz et Noy (2007) ». Selon de Kaenel et Iriarte, « les dernières évolutions du web, avec l'entrée en jeu de XML, des nouveaux usages et nouveaux outils, ainsi que le déplacement du centre de gravité qui s'est fortement rapproché des utilisateurs, ouvrent de nouvelles voies et de nouveaux champs d'application pour les catalogues en ligne »de Kaenel et Iriarte (2007).

8.1 Technologie et pratiques bibliographiques en milieu universitaire

Nous organiserons une étude autour de plusieurs axes techniques et sociologiques pour mieux appréhender les besoins liés à la problématique. L'axe principal est technique, il se propose de faire émerger les protocoles de recherche des universitaires (doctorants, documentalistes et enseignants chercheurs). Ces protocoles seront établis sociologiquement, sur un deuxième axe, pour les usagers en fonction de leur niveau d'expertise et de leurs centres d'intérêts scientifiques.

1. Un Service Commun de la Documentation rassemble les bibliothèques d'un établissement de l'enseignement supérieur.

8.2 Problématique

8.2.1 Constat

Bermès et Martin déclaraient qu'une « collection numérique ne peut pas être appréhendée directement. Elle requiert une médiation technique entre l'usager et la collection Martin et Bermès (2010) ». L'objet de la présente étude est précisément de déterminer des profils d'usagers de l'information scientifique et technique (IST) au travers de l'observation de leurs choix techniques. Ainsi, munis de précieuses informations relatives aux usages, nous tenterons de synthétiser les besoins des utilisateurs en terme de « médiation technique ». Ces données permettront ultérieurement de créer un outil répondant aux besoins en terme d'usage technique.

8.2.2 Postulat

Pour les doctorants et les enseignants chercheurs principalement, nous avons tenu à distinguer principalement les sciences dites « dures », des sciences humaines et sociales. Cette distinction est importante en terme d'usages, car les outils et protocoles de production scientifique diffèrent. En effet, en sciences dures, il est courant que la demande des revues et conférence impose un protocole d'écriture précis. Les documents doivent ainsi parfois être rédigés en utilisant le format LaTeX, ce qui induit l'usage du format bibliographique BibTeX. Nous allons tenter de spécifier les pratiques autour de ces contraintes par la présente enquête. En sciences humaines et sociales, les usages diffèrent, nous allons en définir la mesure. Les différences ne s'arrêtent pas à la production normée de documents, les usages en terme d'outils varient également, notamment en matière de recherche d'informations et de stockage de notices bibliographiques.

8.2.3 Résultats attendus

En théorie, cette étude devrait permettre de spécifier des profils donnant lieu à des préconisations d'usage, utiles surtout aux novices. Ces profils sont déjà pressentis, notamment sur l'attachement aux outils libres et gratuits dans l'enseignement supérieur (Berizzi et Zweifel, 2005). Cependant, l'intérêt d'une telle étude est de préciser sur un segment précis de population, à un instant donné la véracité de ces résultats « attendus ».

». D'autres études précédentes donnaient une idée des logiciels et formats les plus usités pour la gestion bibliographique. Une étude suisse de 2009 présente les logiciels proposés par les SCD des universités suisses (Masur, 2009). Masur concluait que le public universitaire connaissait encore largement mal en 2009 les LGRB [1], que les outils libres peinaient à se faire une place parmi l'offre. Masur terminait sur la prédiction que les LGRB prendront plus d'importance dans un proche avenir. Nous nous baserons principalement, pour formuler nos questions en ce domaine, sur l'étude de Carole Zweifel qui présente sept des principaux logiciels de gestion bibliographiqueZweifel (2008). Nous avons sélectionné parmi ces logiciels ceux dits autonomes (qui ne nécessitent pas l'installation d'un serveur hypertexte et d'une base de données en pré requis). Nous avons rajouté RefBase qui est parfois installé par les services informatiques universitaires comme outil de partage bibliographique pour une équipe ou un laboratoire. Nous avons complété notre panel avec les deux outils propriétaires parmi les plus connus : EndNote et RefWorks. Nous offrons donc au sondé de choisir le ou les outils de gestion bibliographique qu'il utilise le plus parmi six des plus connus. Il a également la possibilité de proposer un outil alternatif s'il le désire. Il en va de même pour les sources d'informations scientifiques et techniques dans le milieu de la recherche. Nous allons réactualiser et mettre en contexte ces données. Mais également et surtout, nous obtiendrons des informations précieuses sur les caractéristiques techniques de compatibilité que doit proposer un système de recherche d'informations en adéquation avec ses usagers. De plus, cette étude pourra être utilisée pour mesurer l'adéquation technique des SRI en information scientifique et technique avec les outils communément usités par les communautés de chercheur. Dans un premier temps, nous allons proposer les éléments de notre questionnaire et les objectifs induits par leur recoupement.

8.2.4 Méthodologie d'enquête

Pour encourager les utilisateurs à répondre au questionnaire, nous avons volontairement simplifié le questionnaire en réduisant la typologie des questions. Ainsi, le temps des sondés est épargné au maximum. Nous ne ferons pas intervenir ici de questions sur l'âge ou le sexe, ce qui nous semble sans objet, voire déplacé. Typologie des questions :

– Questions fermées : 1 choix parmi 3 ou 4.

1. Acronyme de Logiciels de Gestion de Références Bibliographiques, imputé à Zweifel

- Questions semi-fermées : 1 ou plusieurs choix parmi plusieurs.
- Questions semi-ouvertes : 1 ou plusieurs choix parmi plusieurs ou la possibilité de fournir sa propre réponse.

Nous avons établi le plan de sondage de la manière suivante :

1. Questionnaire filtrant pour établir des profils scientifiques avec des questions fermées (statut, expérience, science étudiée).
2. Questionnaire fermé sur le contexte technique d'écriture scientifique (système d'exploitation et logiciels connus et utilisés).
3. Questionnaire plus spécifique sur les pratiques bibliographiques avec des questions fermées et semi-ouvertes.
4. Questionnaire fermé et semi-ouvert sur l'impact des aspects de logiciels libres et de gratuité sur le choix de l'utilisateur.
5. Questionnaire des priorités dans le choix d'un logiciel bibliographique.
6. Question semi-ouverte sur le choix des sources d'informations scientifiques.

8.2.5 Population cible et Panel

La question de la représentativité d'un panel de sondés est cruciale pour la crédibilité d'une enquête. À partir d'une vingtaine d'observations (par échantillon de la population observée), des tendances émergent. Cependant, plus le public est hétérogène, plus il faut élargir le panelRanjard (2000). Notre étude, extrêmement ciblée, est quantitative. Dans notre cadre, contrairement aux sondages d'opinion, il n'est pas nécessaire d'interroger beaucoup d'individus de chaque catégorie. Pour chaque population étudiée, 20 à 25 personnes suffisent pour obtenir de bons résultatsWahnich (2006). Quand à la marge d'erreur, pour une population entre 100 et 200 personnes, dans le cadre de ce type de sondage, elle se situe entre 3.1 et 4.4 points si l'on se réfère au tableau d'intervalle de confianceRipon et Evans (2011). Cette marge d'erreur nous semble acceptable.

Cette enquête a été menée sur une période de 15 jours en décembre 2011. Ce formulaire proposé sur la base du bénévolat et de l'anonymat proposait aux usagers de recevoir un compte-rendu du rapport après publication des résultats, s'ils laissaient leur courriel. Notre enquête a été proposée en ligne sur Google Formulaires et soumise :

	Doctorant début de thèse	Doctorant fin de thèse
Sciences humaines et sociales	34	45
Sciences dures	5	6
Autres	5	3

Tableau 8.1: Répartition des doctorants par expérience et domaine de recherche

- à la liste de diffusion de l'association EGC [1] ;
- à l'école doctorale Cognition Langage Interaction de l'Université Paris 8 ;
- aux laboratoires d'informatiques LIAFA et LIASD respectivement rattachés aux Universités Paris 7 et de Paris 8 ;
- à de grands établissements comme Sciences Po et à l'ensemble des doctorants de l'école doctorale Abbé Grégoire du CNAM ;
- aux conservateurs et bibliothécaires des universités parisiennes ;
- à la liste de diffusion de l'ADBS ;

Le panel est constitué de 195 participants, dont la distribution est la suivante :

- 54 enseignants chercheurs dont 16 en sciences dures, 36 en sciences humaines et sociales et 2 dans d'autres domaines ;
- 98 doctorants dont 11 en sciences dures et 79 en sciences humaines et sociales, 8 dans d'autres domaines ;
- 28 personnels de SCD ;
- 4 post doctorants
- 11 personnels hors de ce classement comme des ingénieurs de recherche.

Nous n'étudierons pas en détail le statut de post doctorant, car le segment interrogé ne présente pas un corps suffisamment élevé pour être représentatif. Nous avons segmenté le statut de doctorant, pour spécifier l'expérience, en début (44 personnes) et fin de thèse (54 personnes). Pour spécifier les profils utilisateurs, nous avons encore dégagé 4 sous-ensembles principaux en fonction de l'expérience et du domaine de recherche : Le tableau 1 semble montrer qu'il sera pertinent de faire une distinction sur l'évolution des usages technologiques liés aux pratiques bibliographiques pour les doctorants entre le début et la fin de la thèse pour les sciences humaines et sociales. Passons à l'examen des résultats de l'enquête.

1. EGC - Association Extraction et Gestion des Connaissances : http ://www.egc.asso.fr/

Système d'exploitation (OS)	population totale	SHS	Sc. Dures	Autres
Linux	15,00 %	7,80 %	36,21 %	0,00 %
Windows	63,00 %	70,43 %	39,65 %	83,36 %
Mac OS	22,00 %	21,74 %	24,14 %	13,64 %
Autre	0,00 %	0,00 %	0,00 %	0,00 %

Tableau 8.2: Contexte technologique des usagers, selon la répartition scientifique

8.3 Les résultats

Le dépouillement des résultats en matière de données a donné les résultats suivants :

8.3.1 Contexte technologique

Les sondés ont présenté leur contexte d'usage de l'outil numérique au travers de quelques questions. Ces questions portent sur l'environnement technologique entourant leur production scientifique. À la question du système d'exploitation favori, les chiffres de la population sont :

De manière générale, sans plus spécifier la population, il apparaît que de manière générale, un peu plus de 20 % des sondés utilisent des ordinateurs de marque Apple, que ce soit en sciences humaines et sociales ou en sciences dures. Les systèmes d'exploitation de Microsoft sont utilisés par plus des deux tiers des usagers de SHS. Le pourcentage d'usage de Mac 0S est le même en SHS que pour l'ensemble de la population de notre enquête, à savoir environ 22 %. Avec 8 %, l'usage de systèmes Linux reste marginal en SHS. En sciences dures, la répartition est plus équilibrée entre Linux (36 %) et Windows (40 %). Mac 0S est une fois de plus utilisé par environ un quart du segment. Il est à noter que le système Mac OS X est basé sur un ancêtre commun à Linux et peut être également être utilisé comme telJepson *et al.* (2008). Comme les chiffres montrent que seules les sciences dures ont un réel impact sur le choix d'un système d'exploitation, nous allons affiner par type d'usagers (Doctorants, enseignants chercheurs et documentalistes). Les résultats du tableau 3 montrent qu'en sciences dures, les chercheurs débutants utilisent Windows puis glissent vers l'utilisation de Linux ou Mac (qui est basé sur Linux) en fin de thèse. Les enseignants-chercheurs retourneront à 40 % sur système Microsoft, 36 % resteront sous Linux et 14 % sous Mac. En début de thèse, le jeune chercheur de sciences

OS	Doctorants débutants	Doctorants confirmés	Post-doctorants	Chercheurs	Documentalistes
Linux	40,00 %	66,00 %	50,00 %	36,21 %	0,00 %
Windows	60,00 %	0,00 %	0,00 %	39,65 %	100,00 %
Mac OS	0,00 %	33,00 %	50,00 %	24,14 %	0,00 %

Tableau 8.3: Systèmes d'exploitation utilisés par profil utilisateurs en sciences dures.

OS	Doctorants débutants	Doctorants confirmés	Post-doctorants	Chercheurs	Documentalistes
Linux	3,00 %	9 %	0,00 %	12,50 %	15,40 %
Windows	85,00 %	71,00 %	50,00 %	43,75 %	7,70 %
Mac OS	12,00 %	20,00 %	50,00 %	43,75 %	76,90 %

Tableau 8.4: Systèmes d'exploitation utilisés par profil utilisateurs en sciences humaines et sociales.

dures utilise Windows (60 %) et Linux (40 %). Il apparaît très nettement que les postes de travail sont Linux ou compatible (Mac) en fin de thèse et après pour les populations de chercheurs en sciences dures. Les documentalistes et bibliothécaires de sciences dures semblent utiliser exclusivement des systèmes Microsoft. En SHS, la proportion d'usage diffère notablement. L'utilisation de Linux est anecdotique chez les doctorants et marginale chez les Enseignants chercheurs. Ce qui est surprenant, c'est que plus de 15 % des documentalistes spécialisés en SHS utilisent Linux. De plus, plus des trois quarts de ce segment de la population utilisent Mac OS. Ce système d'exploitation semble être particulièrement apprécié par les jeunes docteurs et les enseignants chercheurs. Windows a plus les faveurs des doctorants et d'une bonne partie des enseignants chercheurs.

8.3.2 Usages d'outils de productions écrites en sciences

Examinons quels sont les outils utilisés pour l'écriture de littérature scientifique par notre population. Les données brutes d'utilisation d'un éditeur graphique (Word ou OpenOffice...) ou d'un compilateur de textes classent l'utilisation de la manière suivante :

Outil	Doctorants débutants	Doctorants confirmés	Post-doctorants	Chercheurs	Documentalistes	Autres
TEX ou LATEX	10,0 %	18,5 %	50,0 %	46,3 %	3,6 %	18,2 %
Éditeur graphique	90,0 %	81,5 %	50,0 %	53,7 %	92,8 %	71,7 %
Autre	0,0 %	0,0 %	0,0 %	0,0 %	3,6 %	9,1 %

Tableau 8.5: Outil de production de documents par profil d'utilisateurs.

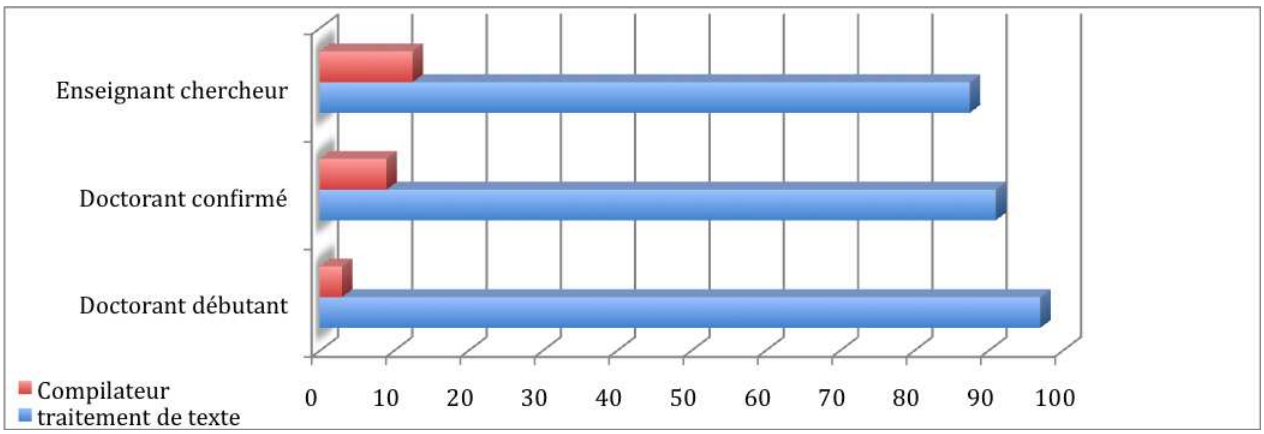


Figure 8.1: Méthode d'écriture en SHS selon l'expérience

1. Éditeur graphique : 149
2. Compilateur de texte : 44
3. Autre : 2

Analysons par répartition ces données.

De manière générale, les chiffres montrent que dans le cadre scientifique l'usage des éditeurs graphiques comme Word et OpenOffice est largement privilégié. Cependant pour le segment des enseignants chercheurs, l'écart entre le pourcentage d'usagers de compilateurs et d'éditeurs est très réduit. Dans les autres groupes, l'usage des compilateurs de texte est quasi nul, sauf pour les doctorants en fin de thèse et les ingénieurs (18 % dans les deux cas).

En sciences humaines, l'usage du compilateur évolue avec l'expérience de presque nul, 3 % en début de thèse à 12,5 %, un usage très modéré pour les enseignants chercheurs.

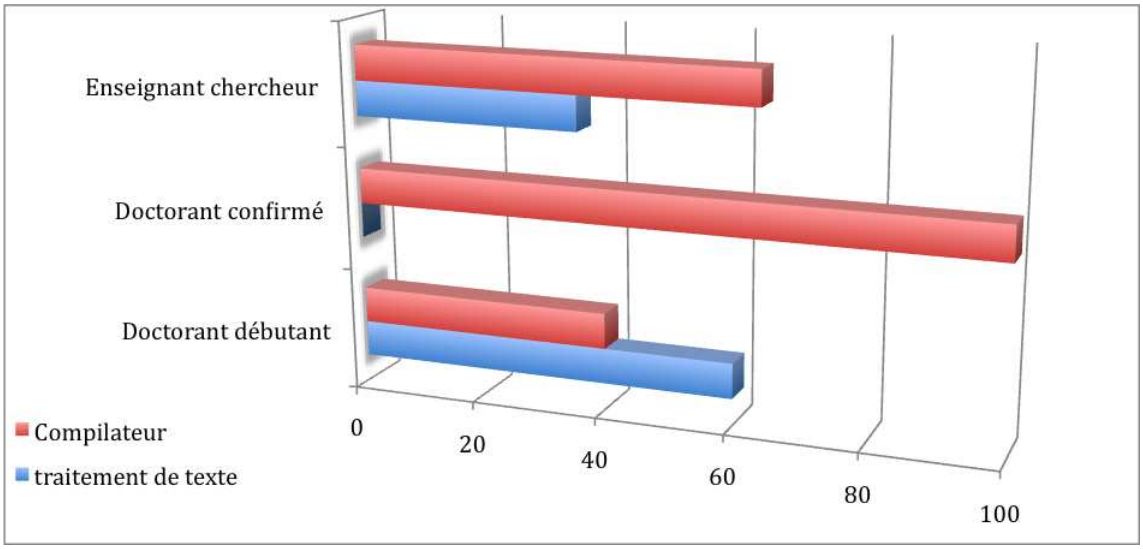


Figure 8.2: Méthode d'écriture en sciences dures selon l'expérience

L'usage d'un compilateur est beaucoup moins anecdotique en sciences dures, notamment pour les doctorants en fin de thèse. Cette partie de la population rédige son mémoire de thèse en sciences dures, avec des équations et une bibliographie complexe. Entre ces considérations techniques et des pratiques sociologiques universitaires très orientées vers LaTeX en sciences dures, l'usage d'un compilateur de texte est presque une thèse de doctorat.

Chez les documentalistes et bibliothécaires, il est plutôt courant d'utiliser un traitement de texte (97 %). Les 3 % restants sont les utilisateurs de compilateurs et autres. Si nous croisons les informations relatives à l'usage d'un compilateur de texte avec celles concernant le système d'exploitation utilisé, nous avons la surprise de constater que son usage n'est pas lié à un système d'exploitation en particulier : Les enseignants chercheurs (tous issus de sciences dures) usant de LaTeX ou TeX utilisent les 3 systèmes d'exploitation avec une majorité de Linux.

Le tableau 8.6 illustre le fait que les usagers de LaTeX utilisent principalement un système compatible Unix, à savoir Linux ou Mac 0S. Cette population édite les fichiers LaTeX de la manière suivante :

- Éditeur à compilation intégrée (TexWorks, Emacs...) : 72 %
- Éditeur basique (vi, ed, Notepad...) : 15 %
- Autre : 22 %

OS	Doctorants débutants	Doctorants confirmés	Post-doctorants	Chercheurs	Documentalistes	Autres
Windows	0 %	10 %	0 %	20 %	0 %	100 %
Mac OS	50 %	30 %	50 %	32 %	0 %	0 %
Linux	50 %	60 %	50 %	48 %	100 %	0 %

Tableau 8.6: OS en fonction de l'usage d'un compilateur de texte (LaTeX,TeX).

Les propositions pour les utilisateurs de TeX/LaTeX, ayant coché la valeur « autre » ont été : Kile [1], 4 personnes ont proposé *LyX* [2] dont 3 en sciences humaines, TeXnicCenter [3], usbTex [4] et Eclipse avec le plug-in TeXlipse.

8.3.3 Les usages et formats de bibliographique

Pour mieux cerner les usages scientifiques en matière de bibliographie, nous avons demandé à la population de sondés s'ils intégraient leurs citations et bibliographies manuellement où s'ils la généraient grâce au compilateur de leur outil de traitement de texte. Il semble que de manière globale, l'introduction des références et bibliographies se fasse encore beaucoup manuellement dans le texte (37 %). Pour ceux qui utilisent un fichier séparé, ils sont 19 % à l'enrichir manuellement et 18 % à utiliser un logiciel bibliographique. En tenant compte des 25 % de personnes qui utilisent l'outil interne à leur traitement de texte, 63 % des sondés en milieu universitaire utilisent une fonction automatisée de gestion de bibliographie. Ce chiffre mérite d'approfondir les pratiques liées à ces outils pour mieux définir les services que doivent rendre les SRI, pour être en adéquation avec les besoins des usagers.

Si l'on répartit ces résultats par sciences, 20 % des sondés en sciences dures intègrent manuellement leurs références dans le texte pour 47 % en sciences humaines. La

1. Kile est un éditeur graphique de TeX / LaTeX disponible sur Mac OS, Linux et Windows http ://www.framasoft.net/article2827.html et http ://kile.sourceforge.net/
2. Environnement TeX avec une interface entièrement graphique utilisable à la souris http ://www.framasoft.net/article1001.html
3. Environnement intégré de développement TeX en C++ sous Windows http ://www.framasoft.net/article1429.html
4. Environnement TeX complet transportable sur clé USB : http ://www.framasoft.net/article4641.html

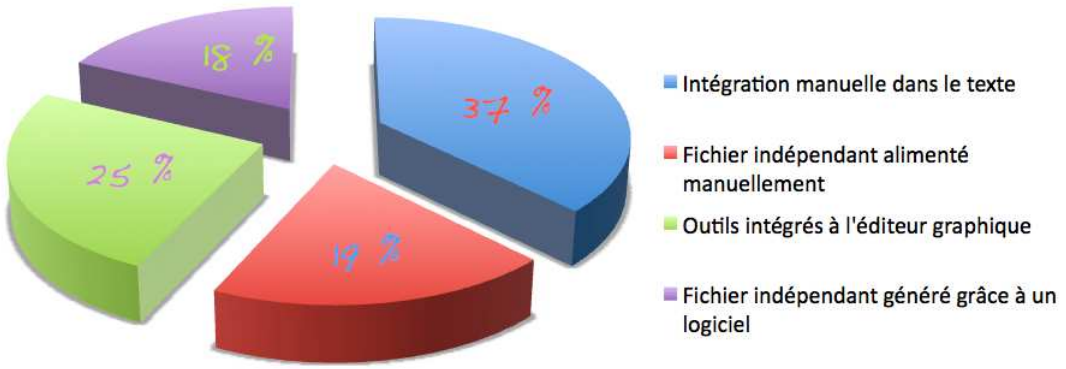


Figure 8.3: Méthode d'intégration bibliographique dans un document pour l'ensemble de la population.

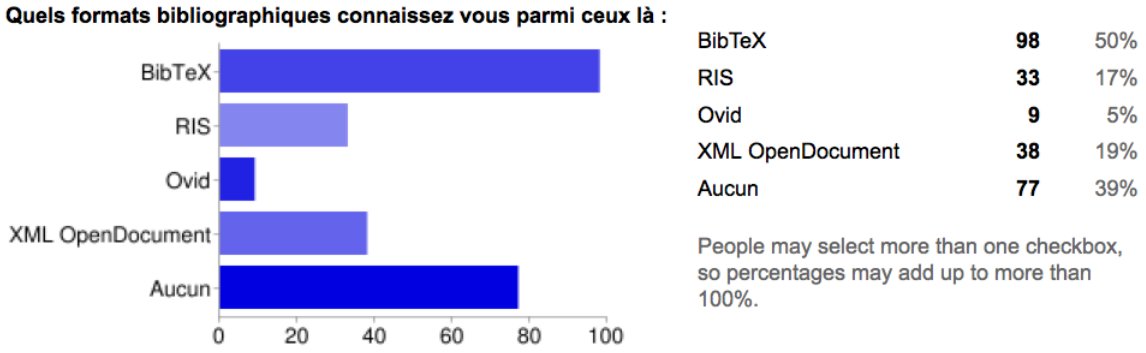


Figure 8.4: Formats de fichiers bibliographiques connus par l'ensemble de la population.

population la plus éloignée de cet usage est celle des personnels de SCD puisque seuls 17,86 % de ceux interrogés intègrent leur références manuellement dans le texte. Nous avons questionné les usagers sur les formats de fichiers bibliographiques qu'ils connaissent et ceux qu'ils utilisent. Nous voyons, grâce à la figure 8.3 que 39 % des sondés ne connaissent aucun des formats bibliographiques présentés. Il s'agit principalement des chercheurs et personnels en sciences humaines et sociales et autres. En sciences dures, seules 3 personnes (soit 5 %) ne connaissent aucun format de fichiers bibliographiques. Il s'agit d'usagers de Windows avec des éditeurs graphiques, comme Word ou OpenOffice. Il est à noter que seuls 41,7 % des sondés issus des SHS connaissent au moins un format de fichier bibliographique parmi ceux proposés.

Formats connus	BibTEX	RIS	Ovid	XML OpenDocument	Aucun
Connaissance unique	11,30 %	2,60 %	1,70 %	8,70 %	58,26 %
Connaissance multiple	16,50 %	5,20 %	2,60 %	8,70 %	-
Population totale	27,83 %	7,83 %	4,35 %	17,39 %	58,26 %
Population restreinte	66,67 %	18,75 %	10,42 %	41,67 %	-

Tableau 8.7: Formats bibliographiques connus en SHS

En sciences humaines toujours, nous allons examiner la connaissance de ces formats de fichiers bibliographique par les sondés de notre échantillon. Le tableau 8.7 se lit de la manière suivante :

- Le terme « connaissance unique » signifie que le sondé connaît ce format bibliographique et seulement celui-là.
- Le terme « connaissance multiple » signifie que le sondé connaît ce format parmi d'autres.
- Les valeurs en rouge indiquent le pourcentage de la population en SHS qui connaît le format bibliographique.
- Les valeurs en vert indiquent le pourcentage de personnes connaissant ce format bibliographique au sein du sous-ensemble composé de la population de SHS qui connaît au moins un format bibliographique .

Note explicative : Dans les tableaux suivants le terme **population restreinte** s'entend comme le **sous-ensemble des sondés connaissant au moins un format bibliographique**.

Le format le plus connu en SHS est le BibTEX avec presque 28 % des sondés. Le format bibliographique OpenDocument compatible avec Word et OpenOffice n'est connu que par moins de 20 % du segment. Le RIS et Ovid sont connus très à la marge, presque de manière anecdotique. Cela s'explique bien pour Ovid qui est un format dédié aux sciences médicales. Voyons si ces résultats sont comparables en sciences dures.

En sciences dures (voir tableau 8.8), l'écrasante majorité de l'échantillon connaît le format BibTEX et une proportion appréciable (22,41 %) connaît l'XML OpenDocument. Notons que la proportion de la population connaissant au moins un format tend à rejoindre le chiffre de la population, puisque seuls 5,17 % des sondés de sciences dures ne

Formats connus	BibTEX	RIS	Ovid	XML OpenDocument	Aucun
Connaissance unique	60,30 %	0,00 %	1,70 %	1,70 %	5,20 %
Connaissance multiple	31,00 %	15,5 %	3,4 %	20,70 %	-
Population totale	91,38 %	96,36 %	15,52 %	16,36 %	5,17 %
Population restreinte	5,45 %	22,41 %	23,64 %	5,17 %	-

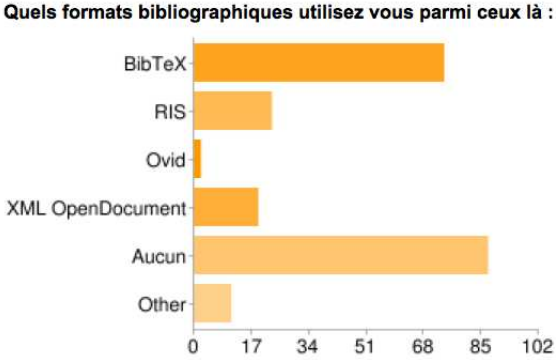
Tableau 8.8: Formats bibliographiques connus en Sciences dures

Formats	BibTEX	RIS	Ovid	XML OpenDocument	Aucun
Connaissance unique	3,60 %	6,00 %	6,00 %	10,70 %	14,30 %
Connaissance multiple	64,30 %	53,00 %	10,70 %	35,70 %	-
Population totale	67,86 %	79,17 %	53,57 %	62,50 %	14,29 %
Population restreinte	17,67 %	46,43 %	54,17 %	14,29 %	-

Tableau 8.9: Formats bibliographiques connus par les bibliothécaires et documentalistes.

connaissent aucun format. Aucun doctorant ou post-doctorant ne se trouvait dans cette situation et seuls 5,5 % des enseignants chercheurs de sciences dures ne connaissant aucun format. Pour ce qui est des bibliothécaires et documentalistes sondés, nous les avons placés à part. Leur statut leur permettant de se positionner au plus près des documents scientifiques, ils sont généralement au fait des normes et pratiques bibliographiques universitaires. Nous avons même constaté que certains connaissaient tous les formats présentés. Comme prévu, les documentalistes et bibliothécaires sont particulièrement sensibles aux formats bibliographiques. Une fois de plus, le format BibTEX est le plus connu avec plus des deux tiers des sondés qui le connaissent. Le format RIS est également connu par plus de la moitié des personnels de SCD interrogés (voir tableau 8.9). Le format de bibliographie XML OpenDocument est connu de presque la moitié de ce panel. Dans l'ensemble, les personnels de bibliothèques ont la connaissance la plus large (pas uniquement le BibTEX) de l'ensemble de la population.

Notons que toutes disciplines et corps confondus, la moitié des sondés connaissaient le format BibTEX. Nous avons ensuite interrogé la population sur les formats de



BibTeX	74	38%
RIS	23	12%
Ovid	2	1%
XML OpenDocument	19	10%
Aucun	87	45%
Other	11	6%

Les utilisateurs peuvent cocher plusieurs cases, donc les pourcentages peuvent être supérieurs à 100 %.

Figure 8.5: Formats de fichiers bibliographiques utilisés

bibliographie utilisés dans le cadre de la rédaction de documents scientifiques, thèse ou de rapports techniques. Plusieurs réponses à la question étaient admises, de plus, les sondés pouvaient proposer une autre réponse s'ils le souhaitaient. L'analyse des chiffres avant répartition indique que le format bibliographique le plus utilisé est le BibTEX. Cependant, ce que nous apprend la figure 3, c'est principalement que 45 % de l'échantillon n'utilise aucun fichier pour gérer les bibliographies. Nous avons également eu des réponses alternatives très pointues comme : Zotero RDF [1], APA [2](2 chercheurs), MARC (une documentaliste). En SHS, presque les deux tiers des sondés (61,74 %) n'utilisent pas de fichiers formatés pour la gestion de leur bibliographie. Parmi ceux qui n'utilisent pas de fichiers normalisés pour noter leur références bibliographiques, 29,58 % se servent néanmoins des fonctions intégrées à leur traitement de texte de gestion documentaire. Pour ceux qui se servent d'un fichier, c'est principalement BibTEX qui a leur faveur à presque 50 %. Pour le reste, 23,91 % des utilisateurs de fichiers bibliographiques formatés en SHS utilisent le format XML OpenDocument et presque 20 % le RIS. À la marge, 9 % des personnes interrogées, dans ce segment, utilisent d'autres formats.

Parmi les doctorants en début de thèse de SHS, sur une population de 34 personnes, 14,7 % utilisent le format BibTEX. Cela est étonnamment élevé, si l'on prend en compte que beaucoup n'ont pas encore commencé la rédaction de leur manuscrit et que ce format est plus facilement utilisé en sciences dures. 3,22 % des sondés de cette même population

1. Format XML de description de notices bibliographiques généré pas Zotero.
2. L'APA est un format de présentation pour les références bibliographiques, notes de bas de page, citations dans les publications défini par l'American Psychological Association (2010).

Formats utilisés	BibTEX	RIS	Ovid	XML OpenDocument	Autre	NSP	Aucun
Utilisation unique	13,00 %	5,20 %	0,00 %	6,90 %	2,60 %	-	-
Utilisation multiple	5,20 %	3,50 %	0,00 %	2,60 %	0,90 %	-	-
Population totale	18,26 %	8,69 %	0,00 %	9,56 %	3,48 %	1,5 %	61,74 %
Population restreinte	47,73 %	18,75 %	0,00 %	23,91 %	9,09 %	-	-

Tableau 8.10: Formats bibliographiques utilisés en SHS

Formats utilisés	BibTEX	RIS	Ovid	XML OpenDocument	Aucun
Utilisation unique	72,40 %	3,40 %	1,70 %	0,00 %	-
Utilisation multiple	6,90 %	5,20 %	1,70 %	6,90 %	-
Population totale	79,31 %	8,62 %	3,45 %	6,90 %	13,79 %
Population restreinte [1]	92,00 %	10,00 %	4,00 %	8,00 %	-

Tableau 8.11: Formats bibliographiques utilisés en Sciences dures

utilisent le format RIS et seulement 5,88 % utilisent le format OpenDocument, plus adapté aux éditeurs graphiques, notamment Word depuis sa version 2007. Pour les doctorants de SHS en fin de thèse, sur une population de 45 personnes, 30 n'utilisent pas de format bibliographique, soit les deux tiers (voir tableau 8.10). Cela peut s'expliquer en regardant le tableau 8.5 page 199. Plus de 80 % des doctorants rédigent sur éditeurs graphiques. Avec Word par exemple, il est possible d'introduire manuellement les centaines de références bibliographiques d'une thèse et d'y faire référence. Pour les enseignants chercheurs et post doctorants en SHS, l'échantillon s'élève à 18 individus dont 10 (soit 55,56 %) n'utilisent aucun fichier formaté de bibliographie. Pour ceux qui en font usage, 22,22 % choisissent le BibTEX, 11,11 % le RIS et 11,11 % l'XML OpenDocument.

Formats utilisés	BibTEX	RIS	Ovid	XML OpenDocument	Autre	Aucun
Utilisation unique	7,10 %	25,00 %	0,00 %	7,10 %	3,60 %	-
Utilisation multiple	1,40 %	21,40 %	7,10 %	17,90 %	0,00 %	-
Population totale	28,57 %	46,43 %	7,14 %	25,00 %	3,57 %	28,57 %
Population restreinte	40,00 %	65,00 %	10,00 %	35,00 %	5,00 %	-

Tableau 8.12: Formats bibliographiques utilisés par les documentalistes et bibliothécaires en SCD.

En sciences dures, l'usage d'un fichier séparé pour la bibliographie est globalement mieux ancré dans la culture scientifique (voir tableau 8.11). Seuls 13,8 % des sondés de ce groupe n'utilisent aucun fichier bibliographique, cependant dans ce sous-groupe seul un utilisateur sur 4 intègre ses citations manuellement, les autres utilisent les fonctionnalités prévues à cet effet par leur traitement de texte. Cependant, en sciences dures le format BibTEX est quasi exclusivement utilisé. 72,41 % de la totalité du groupe l'utilisent exclusivement et 92 % de ceux qui utilisent au moins un format de fichier en font usage. Seul le format RIS est également utilisé comme format unique par 5 % de la population. Nous pouvons remarquer un usage marginal du XML OpenDocument (4 % des sondés en sciences dures) comme format secondaire.

Les documentalistes et bibliothécaires utilisent un large panel de format de fichiers bibliographiques. Parmi les personnes qui n'utilisent pas de fichier, 65 % utilisent le RIS et 40 % LATEX. Le format XML est utilisé également par 35 % des utilisateurs de fichiers formatés de bibliographie. Ceux qui n'en utilisent pas se servent des outils bibliographiques intégrés à leur traitement de texte dans 75 % des cas. Sur l'ensemble de la population de SCD de notre panel, seulement 7,14 % des sondés intègrent leurs références bibliographiques manuellement dans les écrits.

8.3.4 Les usages de logiciels de gestion bibliographique (LGRB)

Nous allons nous intéresser à l'usage des outils qui permettent de générer et de gérer des bibliographies. Dans notre panel, 56 % en utilisent un, 38 % n'en utilisent pas et pour 6 %, cette question est sans objet. En SHS, la proportion est la suivante :

- utilise au moins un logiciel de gestion bibliographique : 53,91 %
- n'utilise aucun logiciel de gestion bibliographique : 38,26 %
- sans objet : 7,83 %

En sciences humaines et sociales, les résultats sont presque identiques à ceux des sciences dures. Seules les pratiques liées à la non-utilisation sont distinctes. En SHS, ceux qui n'utilisent pas de logiciel dédié pour la gestion de leur bibliographie utilisent plus volontiers les fonctionnalités de leur traitement de texte (20,45 %), mais plus souvent, ils intègrent les éléments bibliographiques sans assistance logicielle (75 %) dans les documents. Les autres entretiennent un fichier bibliographique manuellement.

En Sciences dures, les chiffres sont les suivants :

- utilise au moins un logiciel de gestion bibliographique : 53,45 %
- n'utilise aucun logiciel de gestion bibliographique : 41,38 %
- sans objet : 5,17 %

De manière générale, en sciences dures, ceux qui n'utilisent pas de LGRB intègrent leurs références manuellement dans un fichier formaté (58,46 %). L'autre possibilité est qu'ils intègrent directement les références dans le texte sans le concours du traitement de texte (29,17 %). Ils ne sont que 8,33 % de ceux qui n'utilisent pas de LGRB en sciences dures à utiliser les fonctionnalités dédiées d'un traitement de texte. Intéressons nous à la population de documentalistes et bibliothécaires :

- utilise au moins un logiciel de gestion bibliographique : 96,43 %
- n'utilise aucun logiciel de gestion bibliographique : 3,57 %
- sans objet : 0 %

Nous allons nous intéresser aux logiciels qui sont utilisés au sein de la population. Parmi les logiciels signalés par les sondés sous la catégorie « autres », nous avons eu la surprise de voir des personnes utiliser le logiciel de « *concepts map* » XMind pour noter les références bibliographiques. Cette pratique peut surprendre, mais cela permet de visualiser les rapports de co-écritures et de citations entre documents. Les autres logiciels proposés sont ReferenceManager (deux fois par des personnels de SCD) et le logiciel

LGRB utilisé	% global de la population	% en SHS	% en sc. Dures	% En SCD
Zotero	54,13 %	62,90 %	32,26 %	88,89 %
JabRef	20,18 %	9,68 %	45,16 %	14,81 %
RefWorks	0,92 %	1,61 %	3,23 %	0,00 %
BibDesk	6,42 %	3,22 %	16,13 %	0,00 %
Mendeley	21,10 %	11,29 %	38,71 %	25,93 %
EndNote	29,36 %	33,87 %	16,13 %	40,74 %
Bibus	0,92 %	0,00 %	0,00 %	3,70 %
RefBase	0 %	0 %	0 %	0 %
Autres	11,01 %	1,61 %	12,90 %	3,70 %

Tableau 8.13: Ventilation de l'usage des LGRB

libre Pybliographer, une fois. La version Web de EndNote a été également citée deux fois. Deux personnes en sciences dures, un enseignant chercheur et un post-doctorant ont déclaré utiliser le portail ACM comme gestionnaire de bibliographie.

Après vérification, le portail ACM propose effectivement une option soumise à abonnement « My Binders » pour stocker des hyperliens vers les notices des articles repérés sur le portail. Deux personnes utilisent le portail de dépôt d'article en ligne du CNRS HAL pour gérer des notices bibliographiques. Pour la répartition de l'utilisation des LGRB, de manière générale, le plug-in de Firefox Zotero est largement plébiscité. Parmi les progiciels, EndNote de Thomson Reuters bien que payant est également très largement utilisé (29,38 % de la population). Il arrive que les deux logiciels soient utilisés conjointement (12,84 % de l'ensemble du panel). Tous les pourcentages proposés pour la ventilation des usages de LGRB (tableau 8.13 et figure 8.6) s'entendent par rapport au nombre de personnes qui en utilisent dans l'échantillon et pas par rapport à la totalité de la population de référence. En sciences humaines et sociales, Zotero est très largement usité (presque 63 % des utilisateurs de LGRB en SHS). Vient ensuite EndNote (presque 30 % des utilisateurs de LGRB en SHS) qui est parfois utilisé conjointement avec Zotero. Cela s'explique par le fait que les utilisateurs peuvent utiliser Zotero de deux manières. Ce plug-in peut être un logiciel unique dont l'usage va de la détection sur les sites compatibles de notices bibliographiques à l'export dans Word ou OpenOffice en passant par le stockage et la gestion de la bibliographie. Mais Zotero peut également être utilisé

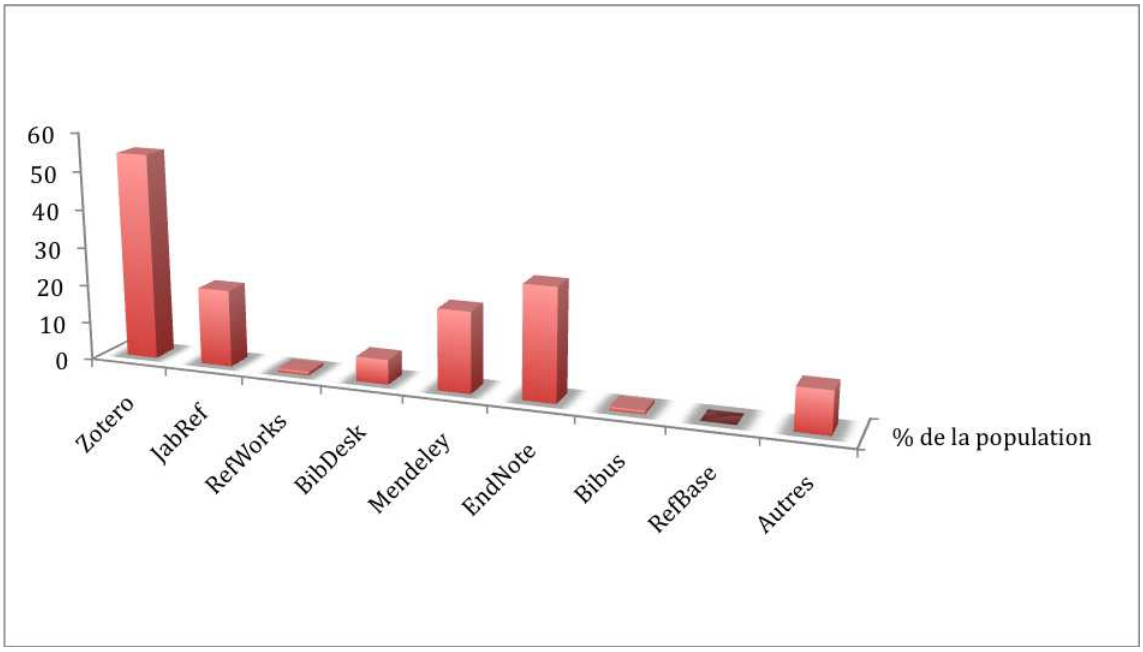


Figure 8.6: Logiciels utilisés pour la gestion documentaire, parmi ceux qui en utilisent au moins un.

juste pour détecter les documents scientifiques, les notices sont ensuite exportées dans EndNote qui se chargera et gestion et de l'intégration dans le traitement de texte.

L'écriture de documents en sciences dures nécessite souvent l'usage du format BibTEX, cela explique que JabRef arrive en tête des logiciels en terme d'utilisation avec presque 30 % des utilisateurs de LGRB en sciences dures. Mendeley (38,71 %) et Zotero (32,26 %) sont également largement utilisés. Zotero est d'ailleurs parfois utilisé en collaboration avec Mendeley et JabRef, pour les mêmes raisons et dans les mêmes conditions qu'en SHS. Mendeley offre la possibilité de communiquer avec Word grâce à un plug-in, alors que JabRef se spécialise dans la création et la gestion de fichiers BibTEX. JabRef par exemple permet de détecter les erreurs de typage dans un fichier BibTEX. Le chercheur s'épargnera ainsi une fastidieuse étape de déboguage à la compilation de la bibliographie du document.

Zotero est très largement utilisé par les documentalistes et la seule personne en SCD de notre panel qui n'utilise pas de logiciel de gestion de bibliographie, se sert de notices Marc individuelles générées par le système de gestion de la bibliothèque. Dans l'ensemble, les personnels de SCD utilisent un ou plusieurs LGRB. Zotero est utilisé

par presque 89 % des documentalistes et bibliothécaires dans l'enseignement supérieur. En complément et dans les mêmes conditions qu'en SHS ou Sciences dures, ils utilisent EndNote (40,74 %) ou Mendeley (presque 26 %).

8.3.5 Critères pour le choix d'un LGRB

Les facteurs qui orientent le choix d'un logiciel de gestion de références bibliographiques sont multiples. Outre ceux d'ordre technique induits par le domaine de recherche, nous allons essayer d'en dégager plusieurs autres. Dans un premier temps, intéressons-nous à la sensibilité aux logiciels libres ou Open Source qu'ils soient gratuits ou payants, opposés aux logiciels propriétaires qu'ils soient gratuits ou payants. Sur l'ensemble de la population, 62,6 % préfèrent utiliser la première catégorie. 7,7 % des sondés se déclarent plus sensibles à l'usage logiciels propriétaires. Enfin, 29,7 % n'ont pas de préférence par rapport à ce critère. Cette proportion est à peu de choses près équilibrée quelque soit le profil en catégorisant par matière ou personnel de SCD. Pour les doctorants en début de thèse, la donne est différente. Sur une population de 44 personnes, 4 (9,1 %) préfèrent les logiciels propriétaires, 17 (38,6 %) les logiciels libres et 23 (52,3 %) sont indifférents à ce critère. En fin de thèse nous assistons à une évolution : seuls 5,6 % continuent de préférer les logiciels propriétaires, 66,7 % se sont tournés vers le libre et 27,7 % n'ont pas d'opinion sur ce sujet. À l'étape suivante, pour les post doctorants et les enseignants chercheurs, le taux de préférence pour le logiciel propriétaire est 8,6 %. L'adhésion au logiciel libre monte à 72,4 % et l'absence de préférence descend à 20 %.

La figure 8.7 illustre très clairement qu'avec l'expérience l'attachement au libre augmente – pas forcément au détriment des logiciels propriétaires – et l'indifférence au critère libre/propriétaire diminue. L'opinion se tranche donc, en faveur la plupart du temps des logiciels libres. Il est possible d'imputer le manque d'intérêt pour ce critère en début de doctorat à une phase d'observation pendant laquelle le jeune chercheur utilise ce dont il dispose. En fin de doctorat, il se sera imprégné des méthodes de son environnement et aura également appris à se servir des outils.

Après une première question générique sur les logiciels en général, la question suivante portait sur choix potentiel d'un logiciel de gestion bibliographique. Nous désirions savoir sur quels critères principaux s'oriente plutôt le choix des sondés. Le choix se portait

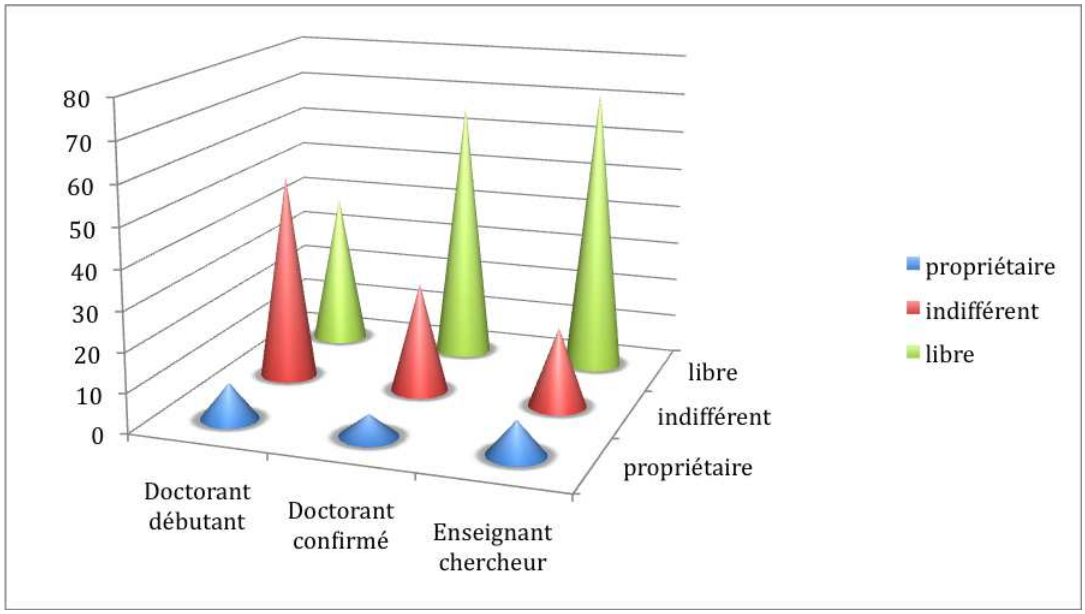


Figure 8.7: Évolution de la sensibilité aux logiciels libres ou propriétaires (en % des groupes sondés) chez les chercheurs.

sur les critères suivants : fonctionnalités du produit, facilité d'installation, gratuité, la disponibilité au sein de la structure (laboratoire, SCD), choix de vos collègues en la matière ou autre (avec invitation à préciser). Les sondés pouvaient choisir entre une et trois réponses ou ne pas se prononcer. De manière générale, 71 % des sondés s'intéressent avant tout aux fonctionnalités du produit, et ce avant la gratuité 57 %. Pour la moitié de la population, la facilité d'installation est un critère important. Notons à ce propos que parfois des logiciels libres nécessitent une compilation avec des dépendances logicielles. Cela les rend difficiles à installer pour des néophytes en informatique. Le choix des collègues et la disponibilité dans la structure sont des critères importants pour 20 % des sondés. Comme propositions alternatives, nous avons eu : « la facilité d'utilisation » et « l'habitude d'usage ». Parmi ceux qui disposent d'un LGRB, 78,1 % l'ont choisi et installé eux-mêmes, 14,5 % utilisent celui proposé par l'organisme de rattachement déployé par défaut (EndNote dans 62,5 % des cas). L'appartenance à un groupe scientifique ou une catégorisation personnel/doctorant n'est pas un facteur discriminant dans ce cas. Cela mène à penser qu'il s'agit de licences établissement négociées à grande échelle et donc installées aussi bien en laboratoire qu'en centre de documentation. Pour les 7,5 % restant, ils ont du installer eux même le logiciel choisi par l'organisme de rattachement (EndNote

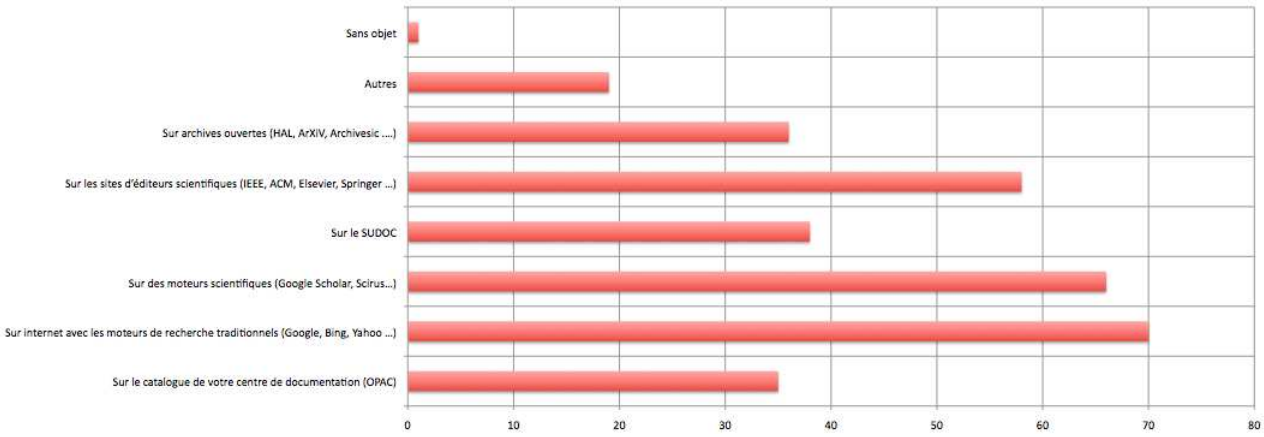


Figure 8.8: Sources d'informations scientifiques pour la recherche dans l'enseignement supérieur.

à 50 %, installation mixte de Zotero et d'EndNote à 50 %). Ceux qui ont choisi et installé eux-mêmes leur LGRB ont le plus souvent choisi Zotero (38 %), souvent associé avec un autre logiciel comme Mendeley, EndNote ou JabRef. Cela tend à confirmer que Zotero est surtout utilisé pour le glanage d'informations scientifiques, mais que la gestion de la bibliographie est plutôt confiée à un logiciel tiers dédié à cet usage. En effet seul 20 % de la tranche que nous étudions utilisent Zotero sans logiciel complémentaire. Mendeley, EndNote et JabRef sont installés et utilisés par 15 % d'utilisateurs chacun.

8.3.6 Recherche d'informations

Pour la recherche d'informations, les sondés devaient spécifier entre une et trois sources spécifiques parmi les suivantes :

- Les OPAC des SCD
- Sur le SUDOC
- Les moteurs scientifiques (Google Scholar, Scirus...)
- Les sites d'éditeurs scientifiques (Elsevier, Springer, IEEE, ACM...)
- Les archives ouvertes (HAL, ArXiV, Archivesic...)
- Les moteurs de recherche traditionnels (Google, Bing ou Yahoo...).

Nous observons en premier lieu que pour la recherche d'informations scientifiques et techniques, dans l'enseignement supérieur la source première d'information reste les moteurs

de recherche commerciaux traditionnels. Cette observation n'est pas nouvelle, elle avait déjà été observée dans des études américaines. Si l'on en croit Markey, les catalogues des bibliothèques sont tombés en disgrâce (Markey, 2007), c'est que semble confirmer notre étude. Notre étude montre que, sur un Panel de 195 doctorants, enseignants-chercheurs et personnel de SCD français, la recherche d'informations scientifiques et techniques ne se fait sur les OPAC que pour 35 % des publics sondés. Nous allons analyser nos chiffres sur cette question par population. Les jeunes doctorants utilisent leur OPAC pour un tiers d'entre eux. En fin de thèse ce chiffre chute à 28 % La surprise vient de la distinction entre sciences étudiées.

Utilisation des sources sciences dures

En sciences dures, les doctorants n'utilisent pas (du tout) le catalogue en ligne de leur SCD. Nous avons assimilé les post doctorants aux enseignants-chercheurs, du fait de leur faible nombre dans l'étude, pour les chiffres suivants. Seul 22 % de la population utilise l'OPAC de son établissement pour accéder à l'IST. Encore une fois la proportion est plus faible en sciences dures avec 7,5 %, là ou en SHS notre étude en dénombre 78,6 %. En sciences dures, la recherche d'IST se fait davantage au travers de moteurs de recherche scientifiques comme Google Scholar ou Scirus (78 % des sondés). La population de sciences dures fait usage des éditeurs scientifiques dans les mêmes proportions (76 %). Les moteurs traditionnels comme Google, Yahoo ou Bing sont également largement utilisés (71 %).

Utilisation des sources en sciences humaines et sociales

En sciences humaines et sociales, nous avons pu distinguer trois populations différentes du fait d'un plus grand échantillon. Les pratiques en recherche d'informations scientifiques et techniques (IST) évoluent avec l'expérience des usagers. En début de thèse les trois principales sources électroniques d'IST sont les moteurs de recherche scientifiques et classiques, mais aussi l'OPAC de l'établissement. Pour ce qui est de l'OPAC, les doctorants sont 44 % à l'utiliser en début de thèse et 33 % en fin de thèse. En fin de thèse, les moteurs traditionnels sont en tête d'utilisation, puis viennent le SUDOC et les moteurs scientifiques. Ces derniers arrivent à égalité avec les archives ouvertes. Enfin, les enseignants chercheurs utilisent aussi bien les moteurs scientifiques

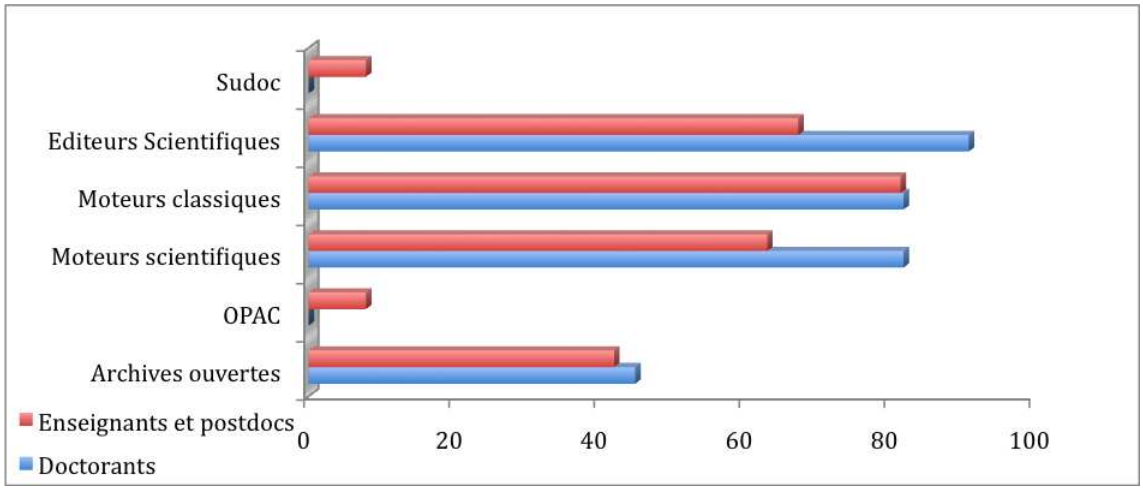


Figure 8.9: Sources documentaires en sciences dures consultées par profil d'expérience.

et classiques que le SUDOC en premier lieu. Ils utilisent moins les OPAC (44 %), ce qui reste très au-dessus de la moyenne d'utilisation des OPAC.

Il faut retenir de ces quelques données que, de manière générale, les chercheurs et étudiants chercheurs en sciences humaines usent plus volontiers leurs OPAC qu'en sciences dures.

8.3.7 Utilisation de sources pour les documentalistes et bibliothécaires

En SCD, la situation est bien différente puisque 78,6 % de l'échantillon se sert de son OPAC. Nous pouvons émettre l'hypothèse que les personnels de SCD maîtrisent parfaitement le catalogue en ligne de l'établissement pour lequel ils travaillent. Ce n'est cependant pas leur seule source d'IST.

Les documentalistes et bibliothécaires ont également une bonne pratique des éditeurs scientifiques, 71,5 % d'entre eux les utilisent pour la recherche d'IST. D'ailleurs, les centres documentaires sont souvent abonnés à plusieurs éditeurs scientifiques, ce qui représente un axe de dépense non négligeable. Les moteurs de recherche classiques (60 % d'utilisation en SCD) ou scientifiques (50% d'utilisation) ne sont pas les sources d'informations privilégiées par les personnels de documentation interrogés. Cela s'explique assez facilement par le fait que l'information est difficilement vérifiable et mal formatée

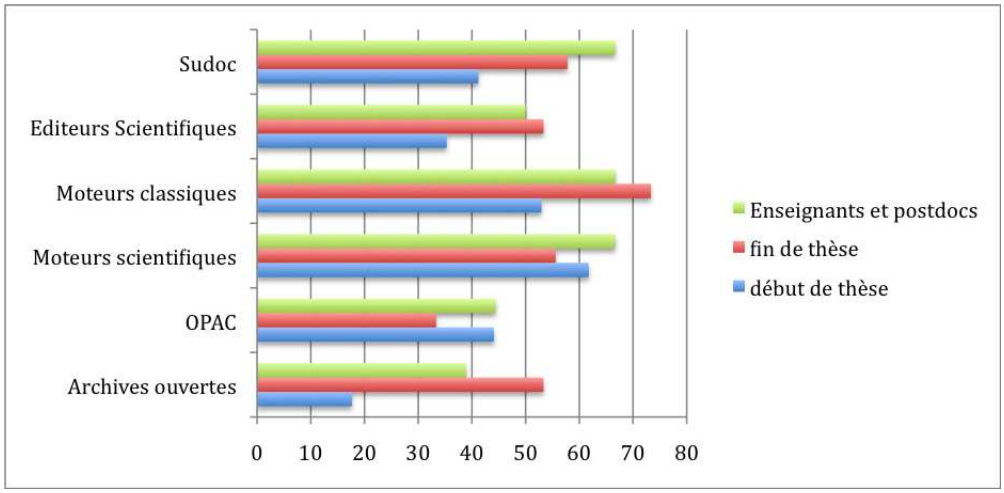


Figure 8.10: Sources documentaires consultées en sciences humaines par profil d'expérience.

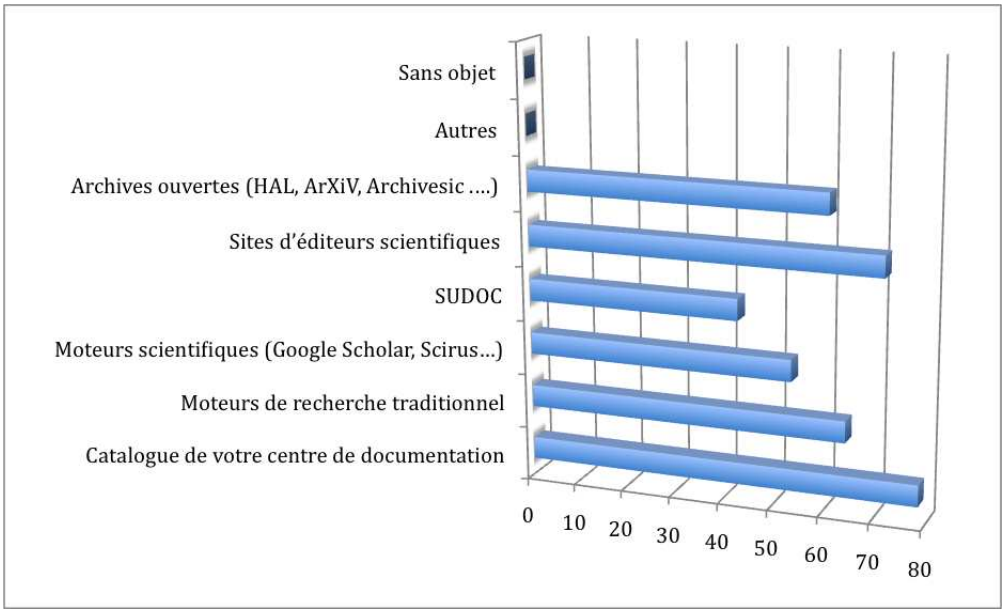


Figure 8.11: Sources de données en centre de documentation

sur un moteur de recherche standard. Pour les moteurs de recherche scientifiques, au premier rang desquels émerge Google Scholar, la fiabilité des données et métadonnées est trop contestable pour être appréciée par un documentaliste [1]. Nous avons noté au cours de nos essais des fonctionnalités de Google Scholar que le formatage des données bibliographiques est souvent approximatif. Les données elles-mêmes sont partielles, ce qui les rend inexploitables en l'état. Il faut notamment régulièrement réajuster le type du document. Le type « *Inproceedings* » (actes de conférence) par exemple est régulièrement remplacé par le générique « *article* ».

Avec 60 % d'utilisation en SCD, les archives ouvertes sont moyennement utilisées par les documentalistes et bibliothécaires. Les données sont souvent redondantes entre archives ouvertes et éditeurs scientifiques. Parfois, sur les archives ouvertes ne sont saisies que des notices bibliographiques. Ces notices sont parfois ne sont pas forcément saisies par les auteurs (Daphy et Ha-Duong, 2010), d'où des inexactitudes dans les métadonnées. Comme les données sont automatiquement répliquées d'une archive à l'autre, les potentielles erreurs sont parfois multipliées. Cela explique peut être, la plus faible utilisation des archives ouvertes chez la catégorie des documentalistes et bibliothécaires, très à cheval sur les la cohésion des données récoltées. Le SUDOC est assez peu utilisé (40% des professionnels de la documentation sondés), peut être en dernier recours, dans l'optique de localiser un ouvrage rare et demander un prêt inter établissement du document primaire. Globalement, les personnels de centres de documentation utilisent le plus large panel de sources documentaires des membres de la communauté de l'enseignement supérieur.

8.4 Analyse des résultats

Grâce à l'interprétation de cette collection de données, nous allons tenter d'établir des profils techniques d'utilisateurs de systèmes d'informations documentaires. Ces profils permettront ultérieurement de définir un cahier des charges pour la modélisation de SRI par population.

1. Lardy disait de Google Scholar : « Google Scholar est un bon point de départ mais qu'il n'a pas encore la maturité des outils de recherche documentaires commerciaux (Lardy, 2011). »

8.4.1 Établissement de profils

Profil Sciences humaines et sociales

En SHS, le profil technique type de l'utilisateur de SRI est composé d'un environnement équipé d'un système d'exploitation Windows. Les doctorants sont globalement équipés de Windows, cette tendance évolue après le doctorat avec un usage équilibré entre Mac OS et Windows. L'outil rédactionnel est très majoritairement le traitement de texte avec un petit pourcentage qui s'oriente vers le compilateur avec l'expérience. Pour ce qui est des usages typiquement liés à la bibliographie, une moitié intègre les citations et rédige sa bibliographie manuellement, ce qui est en accord avec le fait que moins d'un sondé en SHS sur deux connaisse au moins un format de fichier bibliographique. Néanmoins, le format BibTEX est connu par un quart de l'échantillon SHS. Il est utilisé par moins de 20 % de ce segment. Le format bibliographique XML OpenDocument et le RIS sont également utilisés chacun par presque 10 % de la population SHS. Pour la gestion de la bibliographie, la moitié utilise une assistance logicielle dédiée à cet usage (LGRB). Parmi ceux qui n'en utilisent pas, un sur cinq se sert des fonctionnalités intégrées au traitement de texte. Le plus souvent, les éléments bibliographiques sont intégrés et gérés sans assistance logicielle dans les documents. Pour ceux qui utilisent un LGRB en sciences humaines et sociales, la pratique la plus courante est d'utiliser Zotero pour la détection de notices et de l'associer à Mendeley ou EndNote. En SHS, des sources multiples sont largement utilisées, les moteurs classiques et scientifiques bien sûr, mais aussi les éditeurs scientifiques le SUDOC. Les OPAC sont en revanche moins utilisés, mais beaucoup plus qu'en sciences humaines. Les archives ouvertes sont assez peu utilisées, sauf chez les doctorants en fin de thèse.

Profil sciences dures

En sciences dures, pour les chercheurs, le choix des systèmes d'exploitation est plus hétérogène qu'en SHS. Les systèmes basés sur Unix comme Linux ou Mac OS sont privilégiés dès la fin de la thèse, le temps que les doctorants s'en approprient l'usage. Cet usage trouve probablement son explication dans la quantité d'outils scientifiques disponibles sous ces systèmes d'exploitation. Outils au rang desquels les compilateurs de texte de type TEX sont largement utilisés pour la production de documents scientifiques qui sont très majoritairement utilisés dès la fin de la thèse. Cela induit le choix de BibTEX

comme format de bibliographie dans la plupart des cas. Pour gérer leur bibliographie, les chercheurs en sciences dures utilisent pour plus de la moitié d'entre eux un logiciel dédié, souvent JabRef ou Mendeley en collaboration avec Zotero. Encore une fois, un usage conjoint de Zotero avec un autre outil permet de repérer et d'enregistrer les notices bibliographiques depuis internet puis de gérer la bibliographie avec un logiciel de bureau. L'usage de moteurs de recherche classiques ou scientifiques est très répandu pour la recherche d'IST en sciences dures, de même que les portails en ligne des éditeurs scientifiques.

Le profil documentaliste (ou bibliothécaire).

Les personnels dans les SCD de sciences dures sont équipés de systèmes Windows et de Mac OS pour ceux spécialisés en sciences humaines. De manière générale, les documentalistes et bibliothécaires dans l'enseignement supérieur utilisent quasiment exclusivement un traitement de texte. Ils connaissent et utilisent majoritairement les formats BibTEX et RIS qu'ils génèrent depuis Zotero et gèrent avec EndNote ou Mendeley. Pour la recherche électronique d'IST, cette population use d'un large panel de sources. Parmi ces sources, l'OPAC de leur établissement occupe une bonne place. Les éditeurs scientifiques font également partie des leur sources favorites.

8.5 Conclusion de l'étude

Selon le sociologue français Vourc'h, le pourcentage d'étudiants prétendant se rendre à la bibliothèque universitaire au moins une fois par semaine a baissé de 54% en 1997 à 49,9 % en 2006 (Vourc'h, 2010). Pour ce qui est des outils en ligne, l'usage des OPAC est quasi nul pour les doctorants en sciences dures et faible pour les enseignants chercheurs, plus habitués aux moteurs traditionnels et scientifiques, mais aussi aux éditeurs en lignes. En sciences humaines, la consultation de l'OPAC est plus élevée, mais pas au même niveau que les moteurs classiques et scientifiques. Pourtant, le catalogue en ligne des centres documentaires semble parfaitement convenir aux documentalistes et bibliothécaires qui maîtrisent parfaitement leurs arcanes, mais aussi les outils, méthodes et formats bibliographiques. Pour mieux couvrir les besoins technologiques en matière informationnelle, il est possible d'envisager de faire des points d'accès multiples à même

information scientifique et technique (comme pour DBLP [1] ou ISIDORE [2]). Selon l'axe de lecture scientifique et les choix techniques en matière de pratique bibliographique, les utilisateurs pourraient alors avoir un accès adapté à leurs attentes. Il est évident qu'une bonne connaissance des outils normes et formats utilisés par les usagers de l'information est un pré-requis pour modéliser un SRI compatible avec la population cible. On n'envisagera ainsi pas un outil de recherche d'IST en sciences humaines de la même manière qu'un SRI en sciences dures, les formats et styles bibliographiques étant différents. Cependant, la constante dans tous les cas est que l'information se doit d'être visible et exportable, mais aussi compatible avec les outils de glanage contextuel d'information comme Zotero.

1. La base de connaissance DBLP offre une interface « *human readable* » et une interface d'interrogation SPARQL

2. Isidore propose trois interfaces : http ://rechercheisidore.fr/ pour l'interface de recherche intuitive, http ://rechercheisidore.fr/api pour les inter-connections avec les outils logiciels et http ://rechercheisidore.fr/sparql pour les adeptes du langage de requête SPARQL.

Troisième partie

Modélisation et développement d'un outil global de création de bibliographies en informatique scientifique

Chapitre 9

Urbanisation de systèmes d'informations

We build too many walls and not enough bridges.
Isaac Newton

Introduction

Bermès et Martin déclarent qu'une « collection numérique ne peut pas être appréhendée directement. Elle requiert une médiation technique (...) entre l'usager et la collection » (Martin et Bermès, 2010). Les usagers des bibliothèques et plus globalement de l'information électronique s'attendent à passer de manière transparente du contenu documentaire aux références associées. Les systèmes d'informations (SI [1]) ont eu longtemps tendance à s'apparenter à des forteresses médiévales avec une entrée unique : une page au format HTML destinée uniquement à être lue par un humain. Cette vision caricaturale de l'époque épique de l'Internet documentaire est heureusement en train

1. Définition de Système d'information « Processus qui collectent des données structurées conformément aux besoins, qui stockent, traitent et distribuent l'information (Andreu *et al.*, 1992) »

de disparaître au profit de véritables serveurs de données. Ces derniers ressemblent de plus en plus à des mégalopoles informationnelles modernes équipées de nombreuses passerelles pour rendre leurs données accessibles aux autres systèmes d'informations du Web.

Il n'est plus possible de modéliser un SRI et plus généralement un SI sans se référer directement aux technologies de l'information (Kéfi et Kalika, 2004). Cette médiation s'entend sous forme de métadonnées relatives à chaque élément composant ladite collection. Selon Letrouit (2005), chaque ressource ou service possède sa norme d'exposition de métadonnées par une urbanisation *had hoc* du SI (Heurgon, 1990). par le terme d'urbanisation, nous désignons la mise à disposition de données par un système d'informations à destination d'un autre système au travers de formats normés. Il s'agit de l'interopérabilité au sein d'un système d'informations hétérogène complexe. L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre (Jarillon, Pierre, 2010).

L'interopérabilité se concrétise par des fonctions *Extract Transform Load* capables de fournir le flux informationnel dans le format attendu (Vassiliadis, 2009). Cette mise à disposition peut intervenir par la création de canaux informationnels normés, il peut par exemple s'agir formats provenant de la famille XML. Il peut également s'agir de l'exposition d'informations invisibles à l'œil humain dans le code source d'un document au format HTML. La mise en œuvre de métadonnées relatives aux documents scientifiques et techniques, au sein des bases de connaissances, donne un cadre structuré aux informations essentielles aux usagers de l'information et les guide dans leur quête face à une offre pléthorique (Nieuwenhuysen *et al.*, 2005). Dans le cadre d'une interaction directe entre l'utilisateur et le système, ces métadonnées se doivent d'être humainement visibles, idéalement hyper-liées à un processus de re-formulation comme un présentation utilisant des facettes. Cependant dans l'optique d'une interrogation inter systèmes (entre deux processus logiciels) les ressources proposées ne sont utilisables que si elles sont interopérables.

9.1 Historique

Le modèle initial du Web était établi comme un vecteur à la fois statique et unidirectionnel de communication selon des schémas historiquement établis (Jakobson et Ruwet, 1963, Shannon et Weaver, 1948). Le principe de communication sur l'Internet était jusqu'à la fin du siècle précédent une émission de connaissances, souvent institutionnelle ou académique, qui se propageait vers des lecteurs. Berners-Lee et Hendler ont planté le décor d'un espace communicationnel sémantique (Berners-Lee et Hendler, 2001) qu'ils ont redéfini plus tard comme « web de données » dans lequel les informations sont normées, reliées par le sens selon une logique compréhensible par des systèmes automatisés (Berners-Lee *et al.*, 2009). Google, de par sa position hégémonique, se révèle être une porte d'entrée quasi obligatoire pour le monde de l'information numérique (Simonnot et Gallezot, 2009). Avec Panda, nouvel algorithme d'indexation de l'Internet, Google a imposé à l'été 2011 un changement de paradigme dans la recherche d'information. Avec sa logique orientée web sémantique, la méthode Panda a révolutionné l'accès aux données numériques (voir chapitre 3, page 62). En effet, la valeur des sites web n'est plus attribuée uniquement grâce aux « *backlinks* » (liens entrants) et à l'analyse statistique du contenu. L'algorithme a été modifié pour prendre davantage en compte le contenu et la sémantique des sites web pour le classement d'une page par rapport à une thématique donnée. Les documents dont les métadonnées s'accordent avec les formats du Web de données sont largement privilégiés. Dans ce contexte, les méta-informations sont encapsulées au sein des documents hypertextes au format dit « web 3.0 », comme le RDFa et les micro-formats. Ce changement dans le calcul de la valeur d'un site par Google visait à améliorer l'adéquation entre les besoins de l'usager en quête d'information et les résultats offerts par l'interface de recherche. Cependant, cette évolution a eu pour effet de pousser les webmestres et les professionnels du net à améliorer leurs contenus mais aussi la manière de les mettre en valeur. Ainsi, la structuration et l'annotation des documents offrent aux outils d'indexation et d'interrogation la possibilité de sélectionner, filtrer et proposer une information plus en adéquation avec la requête de l'utilisateur (Benali *et al.*, 2009) mais aussi provenant des autres systèmes d'information. Dans notre contexte, nous n'aborderons que les méthodes et technologies d'urbanisation qui offrent un intérêt dans le cadre de la science documentaire.

9.2 Valorisation des métadonnées hétérogènes

L'architecture du Web a été initialement conçue pour s'appuyer sur l'emploi de langages de balisage (HTML) et de métadonnées. Cette tendance s'est confirmée avec les langages XML et xHTML. Le modèle de structure de données RDF (*Resource Description Framework*), défini par le W3C en 1999, fournit un cadre de description des ressources qui fonde l'interopérabilité entre les ressources disponibles sur internet, mais également avec d'autres ressources informatiques.

Dans cette section nous allons étudier les méthodes de valorisation de l'information au sein des systèmes d'informations. Nous pensons particulièrement aux chaînes de production de données automatisées, *Workflow* en anglais. Nous pensons particulièrement aux portails dédiés à la connaissance et à la recherche scientifique. En effet, comment modéliser un système de recherche d'informations (SRI) tout en ignorant le contexte d'usage ? Placer l'humain au centre du processus de modélisation permet de proposer un outil au plus près des besoins et attentes des usagers. Des études publiées précédemment synthétisent la littérature sur la méthodologie de recherche d'informations sous des aspects psycho cognitifs (Chaudiron et Ihadjadene, 2002, Kembellec, 2011). Carol Kuhlthau met en évidence les étapes du processus de recherche d'informations tout en y associant les sentiments, les pensées de l'usager. Selon elle, si l'utilisateur du SRI ne trouve pas rapidement l'information qu'il cherche lors des différentes étapes du processus de recherche, il va rapidement se décourager et être plongé dans un état d'insatisfaction et renoncer (Kuhlthau, 2005). Nous pensons que cet état de satisfaction peut être en corrélation, bien entendu avec la qualité des métadonnées recueillies, mais aussi avec la compatibilité entre le SRI et les outils que l'usager s'est choisis. L'usager incapable de charger de manière simple et intuitive le résultat de sa recherche - à savoir une ou plusieurs notices bibliographiques - se retrouverait dans un état d'insatisfaction qui le pousserait à changer de SRI. Ainsi, les aspects simplement quantitatifs et qualitatifs en terme de données fournies par le SRI, s'ils sont indispensables pour créer un système de recherche d'informations, ne sont pas suffisants pour le rendre acceptable par les usagers. Des normes ont été établies par les instances internationales pour encadrer le phénomène dans le cadre documentaire. La NISO [1] a par exemple émis les normes ISO

1. *National Information Standards Organization*

2789 [1] (Statistiques internationales des bibliothèques) et ISO 11620 [2] (Indicateurs de performance des bibliothèques). Une fois les métadonnées normalisées, il reste également à normaliser les échanges entre la source d'information et le logiciel destinataire. Selon Mkadmi et Saleh (2008), face aux dernières évolutions technologiques du web et à ses nouvelles applications (web 2.0 et web sémantique), les bibliothèques numériques doivent désormais s'adapter et redéfinir leur rôle dans les trois dimensions technique, architecturale et sociale.

1. Les ressources doivent être décrites avec une sémantique commune.
2. L'implémentation des fiches électroniques doit être standardisée dans un format interprétable par une machine.
3. Un (ou plusieurs) protocole(s) informatique(s) d'échanges pour ces données doit être structurellement établi.

Dans le champ de la bibliothéconomie, l'exposition hétérogène de données (*Library mashup*) est surtout utilisée pour les sites Web et les catalogues. Le site de la bibliothèque joue un rôle important dans sa valorisation au sein l'environnement numérique et la construction du pont entre les bibliothécaires qui transfèrent les connaissances et les usagers qui les reçoivent (Bach, 2010). Pour ce qui est des sites Web des bibliothèques universitaires et leurs catalogues en ligne (OPAC), une partie est spécialisée à l'usage des universitaires qui sont les utilisateurs dont la qualification documentaire est la plus élevée. Cette clientèle avec des besoins spécifiques nécessite des services dédiés de qualité plus élevée (Bach, 2010). Fichter définit le *mashup* (application composite) comme une application web qui utilise le contenu de plusieurs sources afin de créer un nouveau service affiché dans une interface graphique unique (Fichter, 2009). C'est cette notion de *library mashup* que nous appelons urbanisation de système d'information. Bryson (2010) reprenant Singer (2009) propose de classer les méthodes d'urbanisation de SI documentaire selon une organisation dichotomique basée sur le critère technique du langage utilisé pour mettre en œuvre l'exposition des métadonnées. De leur point de vue, la distinction des systèmes d'urbanisation se fait ainsi : d'une part les nouvelles méthodes

1. `http://www.iso.org/iso/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39181`, accédé le 1er août 2012.
2. `http://www.iso.org/iso/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=37853`, accédé le 1er août 2012.

sémantiques embarquées telles le RDF lié à du xHTML et d'autre part les méthodes dites *POSH (plain old semantic html)*, c'est à dire celles plus anciennes uniquement constituées de HTML traditionnel. Nous prenons toute la mesure de la justesse de cette différenciation, mais de notre point de vue il est plus judicieux d'effectuer le distinguo plutôt sur les méthodes d'usage des populations cibles. Ainsi, nous allons distinguer les méthodes contextuelles des méthodes systématiques d'acquisition d'informations documentaires. Voyons les différentes méthodes technologiques associées à ce concept d'urbanisation de SI ou de *library mashup*.

9.3 Méthodes orientées glanage

Le glanage d'information, opposé au moissonnage qui récolte toute l'information, est une méthode logiciellement assistée de détection et d'import sélectif de notice(s) bibliographique(s) au sein d'un document hypertexte. Pour l'illustration technique des protocoles suivants, nous avons mis en forme dans les différents formats présentés la référence bibliographique suivante :

> Kembellec, G. (2009). Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques. In I. Porphyre, ed., *Actes de la deuxième conférence Toth*, 213–231, Annecy, France.

9.3.1 Dublin Core intégré dans les métadonnées HTML

En 1999, à peine un an après la sortie de la première *Request For Comment* (RFC [1]) relative au Dublin Core, l'IETF [2] proposait une utilisation sur Internet du Dublin Core. Dans la RFC 2731, Kunze (1999) explique comment les descripteurs peuvent être exprimés en utilisant les balises **<meta>** et **<link>** du langage HTML. La balise **<meta>** est conçue pour en-capsuler des éléments de métadonnées dont le vocabulaire de typage est accessible en ligne au travers des *Uniform Resource Locators* (URL [3]) déclarées au sein des balises **<link>**. En utilisant tout, ou partie, des 15 attributs classiques préconisés par le DCMI (voir chapitre 6 page 145), et en les préfixant « dc »,

1. Série numérotée de documents officiels publiée par l'*Internet Engineering Task Force* et décrivant les aspects techniques d'Internet, `http://www.ietf.org/rfc.html`, accédé le 1er août 2012.
2. Groupe international qui participe à l'élaboration de la plupart des nouveaux standards d'Internet.
3. Voir la RFC relative : `http://www.ietf.org/rfc/rfc1808.txt`, accédé le 1er août 2012.

Source de : http://www.geraldkembellec.fr/html_embedded_metadata.html

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
2  <html xmlns="http://www.w3.org/1999/xhtml" xml:lang="fr" lang="fr">
3    <head>
4      <meta http-equiv="Content-type" content="text/html; charset=utf-8" />
5      <meta http-equiv="Content-language" content="fr" />
6      <title>Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques</title>
7      <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
8      <link rel="schema.MARCREL" href="http://www.loc.gov/loc.terms/relators/"/>
9      <meta name="DC.type" content="Text" />
10     <meta name="DC.identifier" scheme="URI" content="http://ontology.univ-savoie.fr/toth/TOTh2008_actes.pdf" />
11     <meta name="DC.identifier" scheme="ISBN" content="978-2-9516-4539-4" />
12     <meta name="DC.title" content="Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques" />
13     <meta name="DC.publisher" content="Porphyre Éditions" />
14     <meta name="DC.location" content="Porphyre Éditions" />
15     <meta name="DC.language" scheme="RFC3066" content="fr" />
16     <meta name="DC.type" content="inproceedings" />
17     <meta name="author" content="Kembellec, Gérald" />
18     <meta name="DC.creator" content="Kembellec, Gérald" />
19     <meta name="keywords" content="ontologie, informatique, CLIR, IR" />
20     <meta name="keywords" xml:lang="fr" lang="fr" content="ontologie, informatique, CLIR, IR" />
21     <meta name="DC.subject" xml:lang="fr" lang="fr" content="ontologie, informatique, CLIR, IR" />
22     <meta name="DC.date" scheme="W3CDTF" content="2008-05-01" />
23     <meta name="DC.relation.isPartOf" content="Actes de la deuxième conférence Toth" />
24     <meta name="DC.source" content="Actes de la deuxième conférence Toth" />
25     <meta name="DC.coverage" content="France" />
26     <meta name="citation_conference" content="Terminologie et Ontologie : Théories et applications" />
27     <meta name="citation_publisher" content="Porphyre Éditions" />
28     <meta name="citation_authors" content="Kembellec, Gérald" />
29     <meta name="citation_title" content="Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques" />
30     <meta name="citation_date" content="2008-05-01" />
31     <meta name="citation_year" content="2008" />
32     <meta name="citation_Event_place" content="Annecy" />
33     <meta name="citation_isbn" content="978-2-9516-4539-4" />
34     <meta name="citation_firstpage" content="213" />
35     <meta name="citation_lastpage" content="231" />
36     <meta name="citation_language" content="fr" />
37     <meta name="citation_keywords" content="ontologie, informatique, CLIR, IR" />
38     <meta name="citation_pdf_url" content="http://ontology.univ-savoie.fr/toth/TOTh2008_actes.pdf" />
39   </head>
40 <body>
41     Kembellec, G. (2009). Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques.
42     In I. Porphyre (Ed.), <i>Terminologie &amp; Ontologie : Théories et applications</i>
43     (pp. 213--231). Actes de la deuxième conférence Toth.
44 </body>
45 </html>
```

Liens vers les vocabulaires de description de métas

Figure 9.1: Exemple de métadonnées HTML « embarquées ».

un document hypertexte devient une ressource en ligne exposant ses métadonnées. Cette séquence de métadonnées est une auto-description pour la ressource hypertexte. Il s'agit d'une des méthodes dites *POSH* évoquées plus haut, c'est à dire du HTML sémantique traditionnel. Examinons son fonctionnement au travers du code source proposé en figure 9.1. Deux vocabulaires de description de métadonnées sont utilisés pour rendre cette page entièrement compatibles avec les logiciels de glanage. Dans le cadre de la figure 9.1, les autorités présentées au sein des balises **<link>** pour les vocabulaires de description sont :

1. Le *DCMI Metadata Terms* du dublin core classique [1]
2. Le *MARC Code List for Relators* de la librairie du congrès [2].

Cette technologie d'exposition de données est implémentée dans un cadre documentaire scientifique à une vaste échelle par revue.org et par HAL (voir chapitre 2, pages 51 et

1. `http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements`, accédé le 1er août 2012.

2. Le *MARC Code List for Relators* est décrit à l'URL : `http://id.loc.gov/vocabulary/relators.html` accédé le 1er août 2012.

49). Même si cette technique est vieillissante, elle reste encore très largement utilisée. De plus, elle offre l'avantage d'exposer les mots clés associés au document dans un format compréhensible par Zotero (Voir chapitre 7 page 178). Avec un logiciel comme Zotero, l'usager peut obtenir une notice bibliographique complète avec résumé et mots clés. L'usage des métadonnées intégrées au HTML présente tout de même un défaut majeur : il n'est en effet pas possible d'exposer les notices de plusieurs documents sur une même page, comme par exemple sur la page des réponses de Google Scholar. Cela s'explique par le fait que les métadonnées sont incluses dans l'entête du document HTML et qu'il ne peut y avoir qu'une seule entête par page. Ce défaut est corrigé en xHTML avec la possibilité, en utilisant RDF, de décrire plusieurs blocs de données au sein d'une même page.

9.3.2 Les méthodes basées sur le RDF « embarqué »

Avec l'avènement du xHTML, le DCMI a entamé une réflexion sur l'opportunité d'adapter l'usage du Dublin Core à cette évolution du HTML. En 2008, dans le document *Expressing Dublin Core metadata using HTML/XHTML meta and link elements*, Johnston et Powell (2010) ont décrit les possibilités d'intégration du Dublin Core dans le XHTML, notamment grâce aux triplets RDF. Cette idée était directement issue de la technologie baptisée GRIDDL (*Gleaning Resource Descriptions from Dialects of Languages*) normalisée par Dan Connolly (2007) pour le W3C en Septembre 2007 [1]. Kunze et Reschke (2010) de l'IETF ont donc révisé la RFC 2371 en 2010 pour tenir compte de cette évolution. Cependant, l'usage courant conserve largement les spécifications précédentes. Ces technologies sont particulièrement efficaces pour une indexation optimisée avec Google Panda. D'après nos tests, elles ne sont pour l'instant pas détectables par les outils comme Zotero ou Mendeley (Voir chapitre 7 pages 178 et 171). L'objectif était, et reste, de trouver une convention permettant une interopérabilité avec les logiciels d'indexation, d'affichage, de glanage et de moissonnage d'information documentaires dans les hypertextes. Comme ces technologies ne sont pas encore compatibles avec les outils dédiés, nous n'examinerons pas leur fonctionnement technique en détail. Gageons cependant que les communautés de développeurs autour des outils libres ouvriront sous peu la voie de la détection de ces technologies RDF que sont GRIDDL, le RDFa et les

1. `http://www.w3.org/TR/2007/REC-grddl-20070911/`, accédé le 1er août 2012.

Figure 9.2: Exemple d'implémentation de COinS

microformats. Nous reviendrons plus en détail plus loin sur les méthodes de description de données documentaires au format RDF (même chapitre, page 244).

9.3.3 COinS

L'un des services documentaires les plus innovants et révolutionnaires permettant d'évoluer dans l'ère de l'Internet de données est le *reference linking*. Il s'agit de la capacité de transmettre des données bibliographiques par des liens hypertextes (Caplan et Arms, 1999). Cette technique permet de connecter les utilisateurs (et leurs outils électroniques) avec toute la richesse des collections numériques de manière intuitive (ou grandement facilitée). *ContextObject in SPAN* d'OpenURL (COinS) permet d'intégrer en HTML les métadonnées nécessaires pour construire un lien compatible avec le protocole OpenURL et présente un taux d'adoption élevé sur le web (Romanello, 2008). Nous proposons de voir cette technique comme une méthode d'inclusion de métadonnées bibliographiques au sein de balises **** dans le code HTML des pages web. L'utilisation de COinS permet à un logiciel « résolveur de liens » (par exemple un outil de gestion bibliographique) de récupérer les métadonnées incluses dans une balise ****. Les métadonnées décrivant l'élément bibliographique (isbn, auteur, éditeur, DOI, etc.) n'apparaissent pas à l'affichage ou à l'impression de l'hypertexte. Si une page d'une interface homme machine (catalogue en ligne, SRI, OPAC) contient un ou plusieurs COinS, les logiciels résolveurs de liens vont « *glaner* » les informations bibliographiques. C'est de loin le plus simple à implémenter avec un code très court. Le principe à écrire les métadonnées de chaque référence dans une balise **** avec

des attributs **class** et **title** respectant un format particulier. Si l'on prend l'exemple du code source [1] présenté dans la figure 9.2, l'objet **rft**, qui préfixe tous les attributs bibliographiques, désigne le document référencé (ex : **rft.atitle** pour le titre de l'article ou **rft.pages** pour l'emplcement de l'article dans les actes de conférence). L'objet **rfr** désigne le *referrer*, l'application qui publie la référence, en l'occurrence Zotero. En effet nous avons généré le code de cette référence bibliographique au format COinS à partir de Zotero qui peut aussi bien détecter que créer les objets COinS [2]. Les attributs du champs **rft** sont spécifiés comme en dublin core avec le titre, les auteurs et autres métadonnées. Les attributs de champs doivent être séparés par le caractère **&** noté **&** en langage HTML. Nous avons constaté qu'en plus d'être compatible avec Zotero et Mendeley, COinS est également détecté par le navigateur incorporé au LGRB BibDesk.

9.3.4 unAPI

unAPI est un connecteur lié au protocole HTTP qui offre la possibilité à n'importe quelle application web de publier des objets identifiés de manière distincte dans des pages HTML. Dans notre optique, unAPI est particulièrement intéressant pour signaler des objets documentaires, principalement des notices bibliographiques. Cette technologie est utilisée comme alternative à COinS et aux métadonnées embarquées en xHTML aussi bien par des sociétés multinationales que par des organisations internationales ou encore par les catalogues des bibliothèques de grandes universités, par exemple :

- Amazon, Google, Microsoft
- Library of congress, National science foundation, NISO
- Georgia Institute of Technology, Oregon State, Alberta, Washington, Yale.

Le principe d'unAPI est différent des autres méthodes d'urbanisation de systèmes d'informations documentaires. En effet, les métadonnées ne sont pas intégrées à la page web, seule une URL est fournie comme identifiant pour chaque référence. Le navigateur va ensuite interroger toutes les URL signalées. La méthode consiste à offrir via une URL un résolveur d'identifiant qui prend des paramètres tels que l'identifiant ou le

1. Les sauts de ligne ne doivent pas être inclus au code, ils ne sont là que pour permettre de faire tenir l'ensemble du code sur une page imprimée

2. Le site *COinS generator* http ://generator.ocoins.info/ propose également un service gratuit de génération de code HTML pour la détection de notices, accédé le 1er août 2012.

Source de : http://www.geraldkembellec.fr/unAPI/index.html

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
2  <html>
3      <head profile="http://a9.com/-/spec/opensearch/1.1/">
4      <title>Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques</title>
5      <meta http-equiv="content-language" content="fr">
6      <meta http-equiv="content-type" content="text/html; charset=ISO-8859-1">
7      <meta http-equiv="Content-Style-Type" content="text/css">
8      <link rel="stylesheet" href="http://www.geraldkembellec.fr/refbase/css/style.css" type="text/css" title="CSS Definition">
9      <link rel="unapi-server" type="application/xml" title="unAPI" href="http://www.geraldkembellec.fr/refbase/unapi.php">
10     <script language="JavaScript" type="text/javascript" src="refbase/javascript/common.js"></script>
11     <script language="JavaScript" type="text/javascript" src="/refbase/javascript/prototype.js"></script>
12     <script language="JavaScript" type="text/javascript" src="/refbase/javascript/scriptaculous.js?load=effects,controls"></script>
13 </head>
14 <body>
15     <table>
16     <tbody>
17     <tr>
18         <td colspan="8" class="smaller" align="center">
19             <a href="http://www.geraldkembellec.fr/refbase/show.php?record=58" title="copy this URL to directly link to this record"></a>
20             <div class="unapi">
21                 <abbr class="unapi-id" title="http://www.geraldkembellec.fr/refbase/show.php?record=58"></abbr>
22             </div>
23         </td>
24         <td id="ref58" class="citation" valign="top">
25         Kembellec, G. (2009). Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques.
26         In I. Porphyre (Ed.), <i>Terminologie &amp; Ontologie : Théories et applications</i>
27         (pp. 213--231).Actes de la deuxième conférence Toth.
28         </td>
29     </tr>
30     </tbody>
31     </table>
32 </body>
33 </html>
```

URL du résolveur

identifiant de l'objet

Figure 9.3: Exemple d'implémentation d'unAPI

format d'export désiré. Ce lien résolveur peut se trouver sur le serveur hôte, hébergeur du site initial, mais ce n'est pas une obligation. En effet, la page peut faire référence à un autre résolveur situé sur une serveur distant. Prenons l'exemple d'un enseignant chercheur qui possède son propre site web. Il peut décider de présenter sa bibliographie en HTML et d'offrir au visiteur de l'exporter en intégralité au travers du résolveur de l'OPAC de son université. Le principe de cette API se décompose la manière suivante (Chudnov *et al.*, 2006) :

- un identifiant de document.
- un bloc de données au format « web de données » microformat spécifique, non approuvé par le projet `microformats.org`.
- l'URL d'un résolveur unAPI.

Techniquement, chaque référence doit être intégrée au sein d'une balise HTML **<abbr>** pour la déclaration de son identifiant. Dès lors, il sera possible d'intégrer dans une page HTML un bloc de données « connecté » au résolveur qui permettra au client de navigation ou au moissonneur de disposer d'informations complètes sur la, ou les ressource(s), dans un format compatible. La page web doit aussi contenir une balise

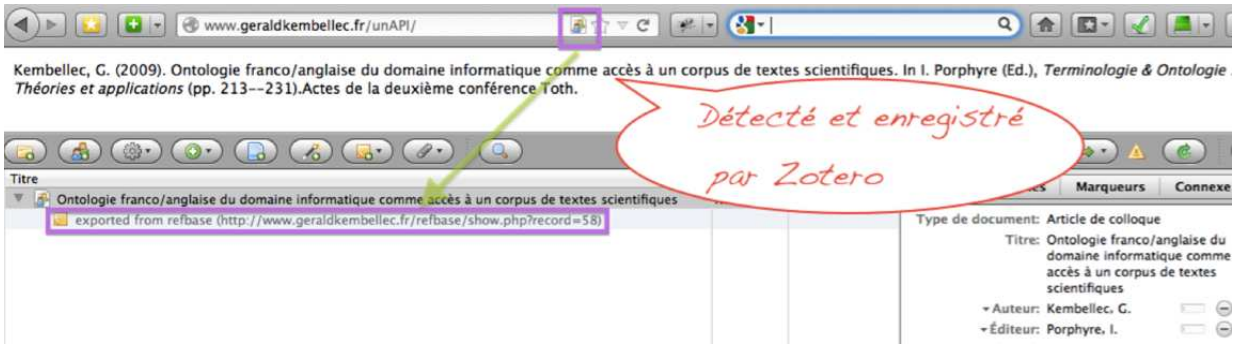


Figure 9.4: Utilisation d'unAPI par Zotero

hyperlien **<link>** qui définit l'URL à laquelle le navigateur pourra aller chercher la description de chaque référence, en passant son identifiant en paramètre (voir la capture du code source figure 9.3). Plusieurs formats de réponse sont possibles, selon ce que déclare le serveur et ce que demande le client.

Un des grands fournisseurs de résolveur unAPI est l'interface de gestion documentaire RefBase. Dans le cas de l'exemple 9.4, une page HTML utilise le résolveur du RefBase installé sur le serveur pour rendre le document (décrit sur une page tierce) détectable par Zotero. Il est également possible d'utiliser un gestionnaire de contenu compatible tels que WordPress ou Spip qui proposent des plugins unAPI. Outre sa compatibilité avec Zotero, en mode d'interaction avec navigation, ou glanage, unAPI propose de nombreuses options d'export de fichier bibliographique, ce qui le rend compatible avec un grand nombre de logiciels bibliographiques au moyen d'import de fichiers. Les formats d'export sont les suivants : BibTEX, endnote, atom, mods, oai_dc, dc, mods et ris. Le site officiel du logiciel Zotero positionne unAPI comme une alternative à COinS pour rendre un site compatible avec Zotero. Toujours selon ce site, ces deux solutions sont les meilleures pour intégrer des métadonnées à une page HTML sur l'Internet dans l'optique d'une urbanisation du système d'information avec Zotero. Cependant, unAPI est compatible avec la plupart des logiciels de gestion bibliographique, mais avec une utilisation moins souple qu'avec Zotero.

9.3.5 Avantages et inconvénients du glanage

Dans le cas du glanage, l'usager n'est pas contraint, comme il le serait avec un lien hypertexte classique de se déplacer vers une page de notice documentaire, ou de l'ouvrir dans une autre fenêtre voire un *pop-up*. Il s'épargne donc :

- d'ouvrir une fenêtre javascript de type *pop-up*[1] contenant ladite notice.
- de faire un copier/coller d'une notice formatée dans son LGRB.
- de recopier à la main la notice au travers d'un formulaire de saisie bibliographie (méthode traditionnelle de MS Word).

Cependant le glanage n'est pas une méthode de recherche, il s'agit uniquement d'une méthode de présentation des résultats d'une recherche ou de pages web. L'implémentation de cette méthode est donc un atout majeur pour la communication entre un outil de gestion bibliographique interagissant avec le navigateur (Zotero, Mendeley...) ou intégrant des fonctions de navigation comme BibDesk (voir chapitre 7 page 175). L'autre avantage de l'exposition des métadonnées du document, dans le cas d'un hypertexte statique, est que certains formats sont reconnus par les moteurs de recherches et proposent donc une indexation plus fine. Cependant, la multiplicité de formats autorisant la détection contextuelle de notices documentaires forme une « tour de Babel » des langages de description de données sur Internet. Ce problème est un véritable défi pour les communautés et éditeurs spécialisés dans les LGRB. Heureusement, ce type de problème est en cours de résolution avec les récentes évolutions du web de données impulsées par (Berners-Lee et Hendler, 2001, Berners-Lee *et al.*, 2009) et relayées par les communautés de normalisation (IETF, W3C, NISO, DCMI et OASIS) et les organismes de recherche internationaux comme l'IEEE ou ACM ou français avec l'INRIA et le CNRS).

9.4 Méthodes orientées moissonnage

La technologie de moissonnage, *harvesting* (Dempsey et Heery, 1998), propose un système de recherche d'information dédié (sur une seule base) ou mutualisé (sur de

1. Cette méthode, bien que toujours utilisée par nombre de page web est fortement dépréciée au niveau des usages car elle s'apparente aux publicités intempestives. C'est pourquoi, certains logiciel de navigation hypertexte (et parfois même des logiciels de sécurité) bloquent les « *pop-up* ».

nombreuses bases de connaissances). Les systèmes de moissonnage permettent des requêtes à large spectre, avec des mots clés, ou plus précises, en utilisant une grammaire booléenne et des méta-caractères. Ainsi, un système de glanage possède une interface unique pour effectuer une requête sur une ou plusieurs bases de connaissances. Les résultats seront affichés dans l'interface unique qui peut être, soit un client local à la machine, soit un site Web (*Rich Internet Application*). Plusieurs protocoles de natures différentes coexistent dans le cadre du moissonnage :

- OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting).
- SRW/U (Search and Retrieve Web Service).
- SPARQL et RDF.

Voyons le fonctionnement et le cadre de ces protocoles :

9.4.1 OAI-PMH

OAI-PMH repose sur HTTP et XML et fonctionne en mode asynchrone. Les uns « exposent » leurs métadonnées, les autres les « moissonnent » et leur ajoutent éventuellement de la valeur. Ce protocole donne une visibilité en dehors des bibliothèques.

OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) est un protocole développé par l'Open Archives Initiative. Il est utilisé pour récolter (ou collecter) les métadonnées permettent de décrire des enregistrements dans une archive, afin que les services proposés puissent fonctionner en utilisant des métadonnées à partir de nombreuses sources (archives). Une mise en œuvre de l'OAI-PMH doit supporter l'exposition des métadonnées en Dublin Core, mais peut également prendre en charge des représentations supplémentaires. Les chercheurs et les bibliothécaires du *Los Alamos National Laboratory* (États-Unis) ont proposé une réunion tenue à Santa-Fe, Nouveau-Mexique, en Octobre 1999 pour résoudre les difficultés liées aux problèmes de communication entre les serveurs de documents et les dépôts numériques. Le point clé de la réunion a été la définition d'une interface qui permettrait aux serveurs de documents électroniques d'exposer de façon structurée leurs métadonnées. Pour répondre à ce besoin il a été décidé de mettre en place une Open Archives Initiative (OAI) qui a tenu sa première réunion à l'Université Cornell (New York) en Septembre 2000. Lagoze et Van de Sompel (2001) ont ensuite présenté les principes et spécifications techniques dont voici la substance :

Un fournisseur de service (*service provider*) « moissonne » à intervalle régulier des entrepôts de données, les archives ouvertes (*data provider*), en collectant les notices correspondant aux documents qui y sont déposés (*harvesting*). Le fournisseur de service est donc l'articulation entre l'utilisateur final et l'archive ouverte qui ne possède pas d'interface adaptée aux humains (*human readable*). De plus, le fournisseur de service offre point d'entrée unique pour rechercher de l'information dans un grand nombre d'archives ouvertes. Citons par exemple oaister d'OCLC [1] comme fournisseur de services en ligne, mais qui peut également être implémenté dans un catalogue documentaire (OPAC). Notons également l'applicatif ORI-OAI [2] qui moissonne, indexe et offre une interface de recherche et peut s'intégrer dans un environnement numérique de travail universitaire (ENT). Comme archives ouvertes compatibles avec OAI, citons HAL [3], Isidore, STAR [4] et ArXiV [5]. Un réservoir OAI-PMH, aussi appelé silo, met à disposition les métadonnées descriptives avec ou sans le lien vers le texte intégral de documents possiblement répartis sur plusieurs serveurs.

OAI-PMH vise à demander de l'information à travers quelques requêtes à un serveur d'archives ouvertes. Les échanges se font du client vers le serveur. OAI-PMH n'est pas un protocole de recherche et ne possède pas d'interface *human friendly*. Les opérations sont limitées : on s'intéresse à un différentiel, à ce qui a été ajouté entre telle et telle date par exemple. Le vocabulaire de l'OAI se compose de 6 verbes et des attributs en Dublin Core simple (non qualifié) (Nelson *et al.*, 2002) comme le montre le tableau 9.1. Le fichier au format XML présenté dans la figure 9.5 offre un exemple de retour sur une interrogation avec le verbe *GetRecord* pour un identifiant sous forme d'URL.

L'interrogation de l'entrepôt de données se fait au travers d'une requête au format OAI, encapsulée dans du HTML. Cette requête est visible dans la partie haute de la figure 9.5.

- **request verb**="*GetRecord*"
- **metadataPrefix**="*oai_dcterms*"
- **identifier**="*oai :hal.archives-ouvertes.fr :hal-00628355*".

1. http ://www.oclc.org/oaister/, accédé le 1er août 2012.
2. http ://www.ori-oai.org/, accédé le 1er août 2012.
3. http ://archivesic.ccsd.cnrs.fr/oai/oai.php ?verb=Identify, accédé le 1er août 2012.
4. http ://staroai.theses.fr/OAIHandler ?verb=Identify, accédé le 1er août 2012.
5. http ://export.arxiv.org/oai2 ?verb=Identify, accédé le 1er août 2012.

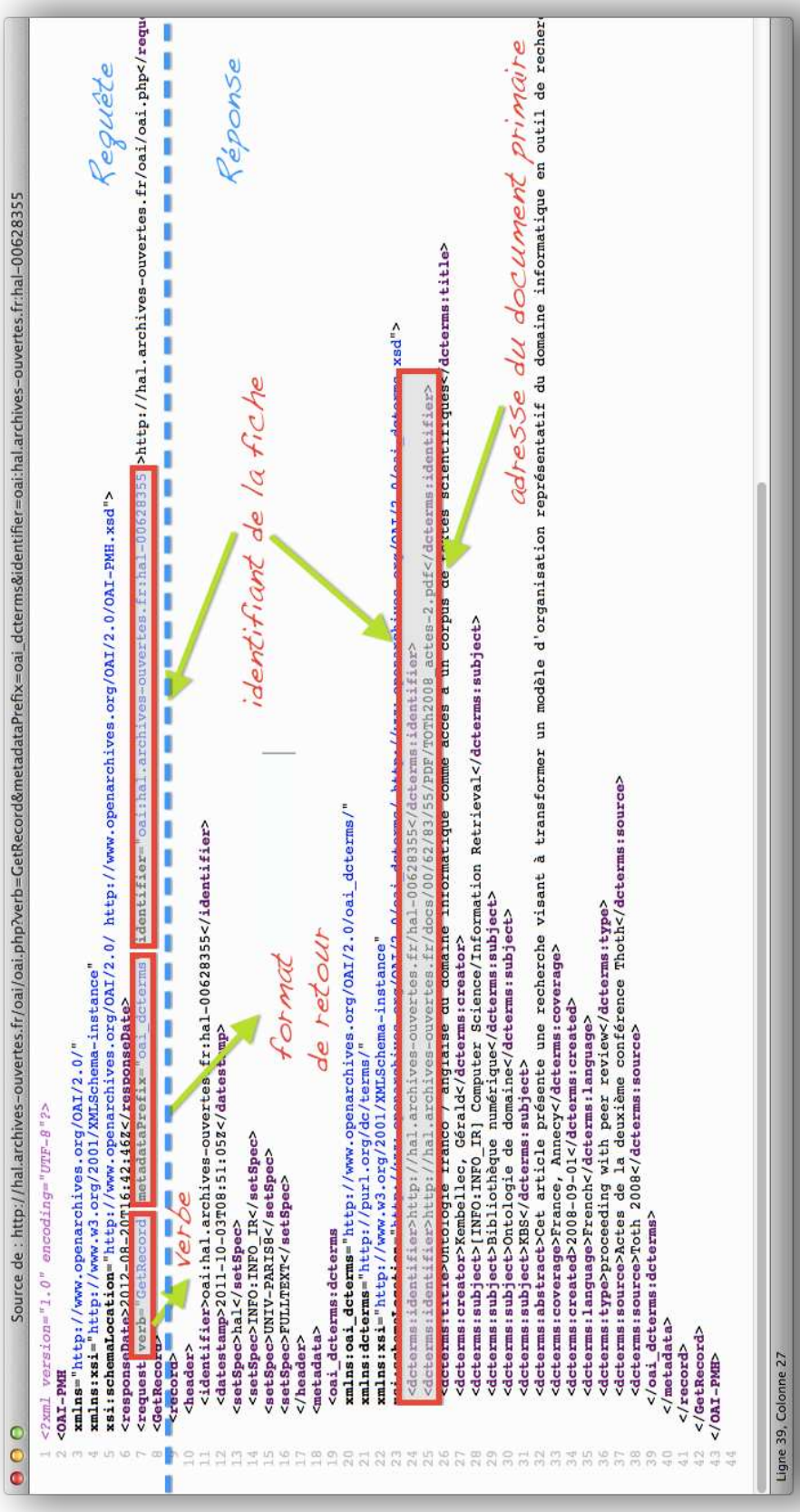


Figure 9.5: Exemple de fiche au format XML dc-oai.

Verbe OAI	Action
Identify	Donne des informations générales sur le serveur
ListMetadataFormat	Donne le ou les formats dans lesquels sont fournies les notices. Au minimum oai_dc pour Dublin Core.
ListSets	Donne la structure de l'Archive Ouverte (nomenclature de classement des notices de l'Archive Ouverte thématique)
ListIdentifiers	Donne les identifiants pour un MetaDataFormat
ListRecords	Donne toutes les notices de l'Archive Ouverte en fonction du MetaDataFormat
GetRecord	Donne l'enregistrement défini par l'identifier

Tableau 9.1: Liste des verbes du protocole OAI-PHM.

Ici, nous réclamons une fiche documentaire, avec le verbe *GetRecord* ayant pour identifiant *oai :hal.archives-ouvertes.fr :hal-00628355*, la réponse étant réclamée au format *oai_dcterms* (dublin core non qualifié).

L'utilisation courante de l'OAI-PMH concerne plutôt la sélection de collections de documents, parfois avec un encadrement temporel. Par exemple on peut souhaiter ne moissonner que les fiches qui ont été ajoutées depuis le dernier moissonnage (méthode incrémentale). On peut également désirer ne pas moissonner toutes les collections, dans le cas d'un serveur généraliste comme HAL ou ArXiV. Il arrive que l'entrepôt ne puisse pas renvoyer l'intégralité d'un moissonnage en une seule fois. Les règles établies par le service informatique peuvent interdire la production de fichiers XML trop volumineux afin de préserver l'accessibilité à l'entrepôt de données pour les autres clients. Dans ce cas, le serveur renvoie une information en ce sens. Le fichier XML généré indiquera le nombre total de notices générées et celui de notices incluses dans le fichier. Un jeton (*token*) sera ajouté au fichier pour que le client puisse moissonner la suite ultérieurement. Il restera à générer une nouvelle requête avec le jeton pour réclamer la suite de sa moisson.

Les clients et silos OAI-PMH sont des solutions complexes à implémenter dont l'usage est coûteux en ressources réseau. On trouve ces solutions principalement au sein des centres documentaires de grands établissements et des universités. L'usage se

démocratise avec ORI-OAI qui peut s'intégrer dans les environnement numériques de travail (ENT) universitaires comme uPortal ou Esup. En effet, le client est composé d'un moteur et d'une interface qui utilisent la même technologie java que Esup. Cet avantage est parfois contrebalancé par la lourdeur des moissonnages. Pour réduire les inconvénients en terme d'utilisation de la bande passante des silos et des établissements clients, les moissonnages de mise à jour sont espacés. Parfois, les informations disponibles pour une collection dans une interface client ne sont pas complètes par rapport à celles disponibles dans l'entrepôt. Ce léger problème est lié à l'essence même des systèmes asynchrones.

9.4.2 Z3950, SRU et SRW

SRU (*Search/Retrieve via UR*L) et SRW (*Search/Retrieve Web service*) sont deux protocoles documentaires qui implémentent et déclinent la norme Z3950 qui est à la fois définie par la norme ISO 23950 :1988[1] de l'*International Organization for Standardization* (ISO) et par l'*American National Standards Institute* (ANSI[2]) Il s'agit d'une technologie orientée « web services » qui autorise l'interrogation simultanée de bases de connaissances réparties au travers d'un portail unique.

Ces protocoles sont gérés par la Bibliothèque du Congrès et sont basés sur le standard de syntaxe de requête, le langage CQL (*Common Query Language*[3]) à travers le protocole réseau hypertexte (HTTP). Ils proposent de normaliser tout à la fois la méthode requête mais aussi le format de réponse.

Les différences entre SRU et SRW résident dans l'encapsulation des requêtes :

- SRU inclut ses requêtes dans les URL via l'architecture REST (*Representational State Transfer*), donc exclusivement en HTTP.
- SRW fait transiter les données au sein d'XML grâce à la technologie SOAP (*Simple Object Access Protocol*[4]).

1. `http://www.iso.org/iso/catalogue_detail?csnumber=27446`, accédé le 1er août 2012.
2. `http://www.ansi.org/`, accédé le 1er août 2012.
3. `http://www.loc.gov/standards/sru/specs/cql.html`, accédé le 1er août 2012.
4. `http://www.w3.org/2002/07/soap-translation/soap12-part0.html`, pour la traduction française, accédée le 1er août 2012.

Les méthodes SRU/SRW utilisent les mêmes instructions, qui permettent l'expression de la requête et de la réponse à cette requête. Les trois opérations principales sont « explain », « scan », et « searchRetrieve »

Il est possible d'envoyer une requête SRU en HTTP soit par la méthode GET [1] ou par POST. Dans le cadre d'une requête GET, une URL est forgée et envoyée au serveur (voir l'exemple en encart page 242). Il y a des contraintes supplémentaires liées à l'encodage des caractères. Le site de la BnF propose la méthode suivante [2] :

1. Convertir la valeur en caractères UTF-8 [3].
2. Utiliser le codage URL qui consiste à remplacer les caractères spéciaux par le caractère « % » suivi du code ASCII du caractère à coder en notation hexadécimale.
3. Construire l'URL avec paramètres et valeurs encodées.

Dans le cadre du POST, les données seront envoyées au serveur dans l'entête de la requête comme dans un formulaire HTML. Les informations demandées sont donc invisibles dans l'URL. La méthode a comme avantage sur GET d'éliminer les problèmes d'encodage d'URL. L'autre avantage est que la longueur d'une URL est limitée de façon variable selon le navigateur alors que la limite des données possibles à recevoir en mode POST dépend du serveur et se trouve être plus importante. Cela permet de créer des requêtes plus complexes, exploitant pleinement les capacités du langage CQL. Dans les deux cas la réponse à une requête SRU/SRW est un fichier XML.
Prenons pour exemple de requête l'URL suivante en s'inspirant du vocabulaire Z3950 SRU/SRW présenté dans le tableau 9.2 :

1. RFC 3986 : `http://www.faqs.org/rfcs/rfc3986.html`, accédé le 1er août 2012.
2. `http://www.bnf.fr/fr/professionnels/proto_sru/s.proto_sru_transport.html?first_Art=non`, accédé le 1er août 2012.
3. Voir RFC 3629 : `http://tools.ietf.org/html/rfc3629`, accédé le 1er août 2012.

Paramètre	Obligatoire	Description
operation	oui	opérateur de recherche contient : 'searchRetrieve', 'scan' ou 'explain'
version	oui	la version du protocole, '1.1' ou '1.2'
query	oui	requête exprimée au format CQL [1]
startRecord	non	index du premier élément à retourner, par défaut : '1'
maximumRecords	non	nombre d'enregistrements maximum, valeur par défaut établie par le serveur
recordPacking	non	détermine la façon dont la réponse doit être retournée soit 'string et 'xml', valeur par défaut : 'xml'
recordSchema	non	format de données dans lequel les enregistrements doivent être retournés, valeur par défaut établie par le serveur [2]

Tableau 9.2: Vocabulaire SRU/SRW (extrait)

Exemple d'URL au format Z3950

1. z3950.loc.gov :7090
2. /voyager ?
3. **version**=*1.1*&
4. **operation**=*searchRetrieve*&
5. **query**=*dc.title%20any%20web%20and%20dc.creator%20any%20berners-lee*&
6. **maximumRecords**=*1*&
7. **recordSchema**=*dc*

Explication : Les caractères & et %20 signifient respectivement ET (séparateur de commande au sein de l'URL) et ESPACE (séparateur de commande au format CQL).

Cela signifie une connexion à l'URL z3950.loc.gov sur le port 7090 (1) à la base voyager (2) pour une requête de type 1.1 (3),celle-c sera une opération de recherche (4). La requête au format CQL signifie que l'on désire les fiches dont l'un des auteurs au moins est Berners-Lee et dont le titre comprend le terme 'web' (5). Le schéma de retour

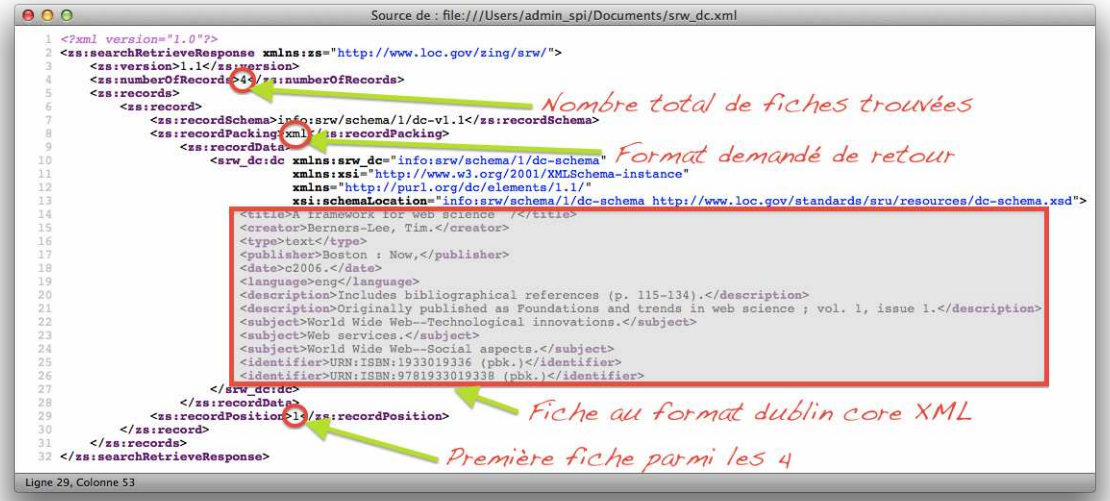


Figure 9.6: Exemple de réponse au format Z3950 SRW.

souhaité est le dublin core (7), par défaut en xml. On désire sélectionner seulement la première réponse parmi celles retournées (6). La réponse fournie par le serveur a été reformatée avec des retours à la ligne pour en faciliter la lecture (voire figure 9.6).

Nous avons explicité l'interrogation d'un serveur Z3950 au travers d'un navigateur web. Il est également possible de récupérer des notices au format XML grâce à un client Z3950. Le logiciel de gestion de références bibliographique BibDesk intègre un client Z3950, ce qui permet d'interroger directement une bibliothèque compatible sans passer par un navigateur web. De plus l'intégration des notices se font ensuite en un clic.

9.4.3 RDF, SPARQL endPoints et triples stores

Le sociologue américain Ted Nelson, pionnier de l'histoire des technologies de l'information et inventeur du concept d'hypertexte voyait dans ce terme « des fichiers structurés à l'information changeante et incrémentale (Nelson, 1965) ». Il proposa alors un format de fichier dont la structure serait évolutive sous forme *Executable And Linking Format*(ELF). Ce système était composé de listes non finies dont les éléments pouvaient être mis en relation entre eux selon des schémas de catégorisation multiples. Ce concept d'hypertexte va plus loin que la navigation entre des éléments textuels réordonnancés pour une navigation intuitive sur un écran. Berners-Lee, pour sortir de la traditionnelle vision en arbre, proposait dès 1989 de voir les systèmes d'informations comme des nœuds

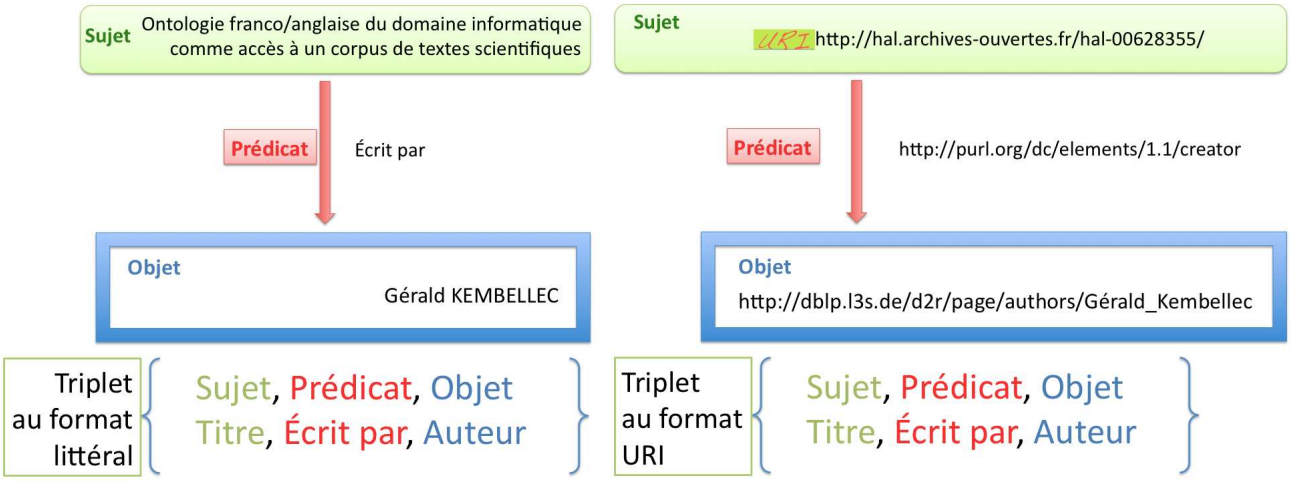


Figure 9.7: Un même triplet RDF présenté en littéral ou avec les URI (de gauche à droite)

de connaissances reliés par des flèches orientées et valuées (Berners-Lee, 1989). Dès la seconde moitié des années 1990 (Gandon *et al.*, 2012, p. 10), plusieurs initiatives sont apparues pour formaliser la description des données sur internet. Citons pour mémoire, le *Metadata Content Framework*, le *XML Data* de Microsoft et *Simple HTML Ontology Extension*. Enfin, comme nous l'avons déjà vu à plusieurs reprises, en 1997, les ateliers du Dublin Core ont enfin dégagé des spécifications viables qui serviront à établir les bases du Web de données. En 1999, le RDF apparaît comme une surcouche au XML qui permet de décrire les ressources de l'Internet et les liens qui les unissent. La typologie des relations s'affine ; si bien que dès 2004, un langage ontologique de description apparaît : le *Web Ontology Language* (OWL) qui étend le RDF. Selon Gandon, Faron-Zucker et Corby, RDF est la première brique des standards du web sémantique (Gandon *et al.*, 2012). Selon eux « tout portail d'information ou site à base de données peut ouvrir ses silos de données décrivant des personnes, des documents, des évènements (...) et en permettre l'utilisation par d'autres ressources. ». Nous apprécions particulièrement le terme de « silo de données » puisqu'il file parfaitement la métaphore agricole initiée avec glanage et moissonnage. La dimension de gigantisme du silo n'est pas usurpée : nous avons moissonné en août 2012 l'ensemble de la base de DBLP au format RDF pour un total de 13,4 giga octets de données liées. Examinons ce qu'est le RDF, ainsi que la manière dont il est possible d'en moissonner les données. Le RDF permet à

la fois de stocker des données, de montrer leurs liens et leurs métadonnées. L'unité de mesure en RDF est le triplet, triple en anglais. La formulation du triplet est de la forme sujet (*subjet*), prédicat (*property* ou *predicate), objet (object)*[1]. La valeur de l'objet est soit une ressource en ligne de type *Uniform Resource Identifier* (URI[2]) ou une chaîne de caractères (un littéral). Une ressource doit disposer soit d'une URI qui est une courte chaîne de caractères identifiant une ressource sur un réseau, ou d'un identifiant électronique *Uniform Resource Name* (URN[3]). Le sujet dénote la ressource, et le prédicat dénote des traits ou des aspects de la ressource et exprime une relation entre le sujet et l'objet. Ici, le verbe dénoter présente le sens fondamental et stable d'une unité lexicale, susceptible d'être utilisé en dehors du discours (par opposition à sa connotation). Nous insistons sur l'adjectif **stable**, car le terme doit être clairement défini pour toujours exprimer le même concept. C'est pourquoi les vocabulaires sont accessibles en ligne et qu'il est conseillé d'y faire référence au sein d'un fichier RDF. Ces vocabulaires sont appelés espaces de nommage (*XML name spaces* ou xmlns).

Les deux cotés de la figure 9.7 illustrent conceptuellement une référence bibliographique en RDF sous forme littérale et sous forme d'URI.

La traduction en RDF donnera le résultat proposé dans le code source de la figure 9.8 en utilisant les espaces de nommage Dublin Core pour décrire le document et FOAF pour l'auteur. Nous aurions également pu lister les mots clés sous forme distincte avec le conteneur RDF <**bag**> (sac en français) qui permet d'offrir une séquence d'éléments identiques. Ces éléments de séquence sont encadrés par des balises RDF <**li**>. Cette méthode serait intéressante pour une utilisation sémantique du RDF, puisqu'elle permettrait des recherches incluant des mots clés communs à plusieurs articles. Nous avons utilisé le validateur officiel RDF du W3C (*Check and visualise your RDF documents*[4]) pour tester la validité de notre exemple mais aussi pour générer le graphe de présentation relatif (cf. figure 9.9). Plusieurs documents présentés bout à bout dans un fichier RDF forment un entrepôt ou silo RDF, appelé *triple store* en anglais. Il est possible d'interroger de plusieurs manières ces entrepôts pour en extraire

1. `http://www.w3.org/TR/rdf-concepts/`, accédé le 1er août 2012.
2. `http://tools.ietf.org/html/rfc3986`, accédé le 1er août 2012.
3. `http://tools.ietf.org/html/rfc2141`, accédé le 1er août 2012.
4. `http://www.w3.org/RDF/Validator/`, accédé le 1er août 2012.

Source de : file:///Users/admin_spi/Documents/RDF_paper.xml

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <rdf:RDF
3       xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4       xmlns:dc="http://purl.org/dc/elements/1.0/"
5       xmlns:foaf="http://xmlns.com/foaf/0.1/"
6       xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
7    <rdf:Description rdf:about="http://hal.archives-ouvertes.fr/hal-00628355/">
8      <dc:title>
9          Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques
10     </dc:title>
11     <dc:creator>
12         <foaf:Person>
13             <foaf:name>Gérald Kembellec</foaf:name>
14             <foaf:surname>Kembellec</foaf:surname>
15             <foaf:givenname>Gérald</foaf:givenname>
16             <foaf:homepage rdf:resource="http://dblp.l3s.de/d2r/resource/authors/Gérald_Kembellec"/>
17         </foaf:Person>
18     </dc:creator>
19     <dc:subject>Ontologie de domaine; CLIR; Bibliothèque numérique; KBS</dc:subject>
20     <dc:publisher>Institut Porphyre</dc:publisher>
21     <dc:date rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">2008</dc:date>
22   </rdf:Description>
23 </rdf:RDF>
```

Figure 9.8: Fichier RDF présenté en XML)

les informations souhaitées. (Gandon *et al.*, 2012) présentent en détail les formats d'interrogation suivants :

- Le XQuery
- Le Xpath
- Le XSLT
- Le SPARQL

Cependant, le format SPARQL (*SPARQL Protocol And RDF Query Language*[1]) est de plus en plus utilisé par de grosses structures de données. Citons entre autres Wordnet (710 000 triplets[2]), DBLP (15 millions de triplets), DBpedia (7 millions de triplets) et Isidore qui possèdent des interfaces d'interrogation de leurs *triples store* en SPARQL. Ces interfaces intégrées sont appelées *endPoints* et autorisent des affichages de résultats dans différents formats RDF et sous forme de graphes. L'autre méthode d'utilisation de SPARQL est l'utilisation d'un moteur de recherche sémantique comme CORESE

1. `http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`, accédé le 1er août 2012.
2. Chiffres proposés par Gandon, Faron-Zucker, & Corby (2012).

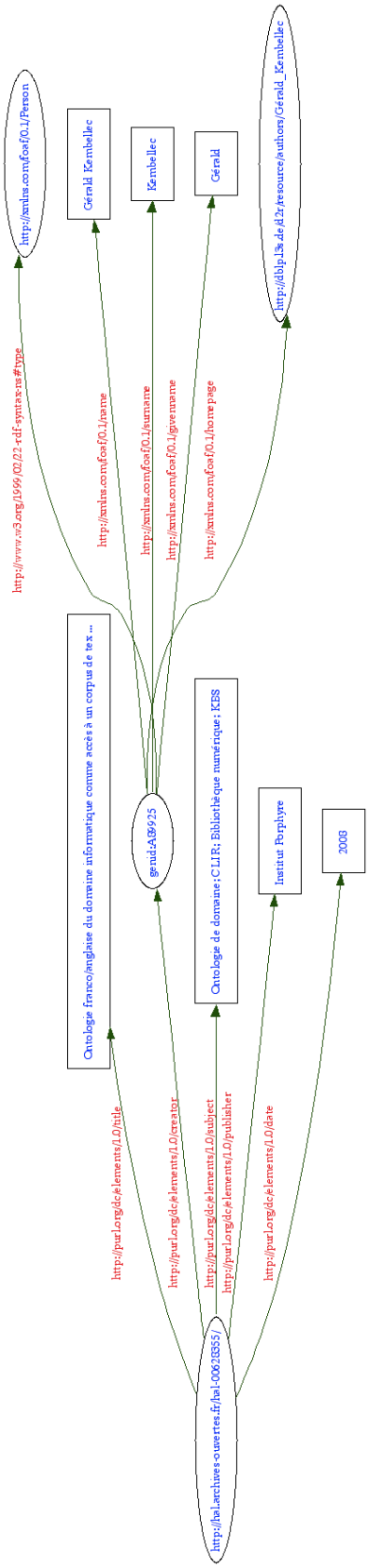


Figure 9.9: Fichier RDF présenté en graphe

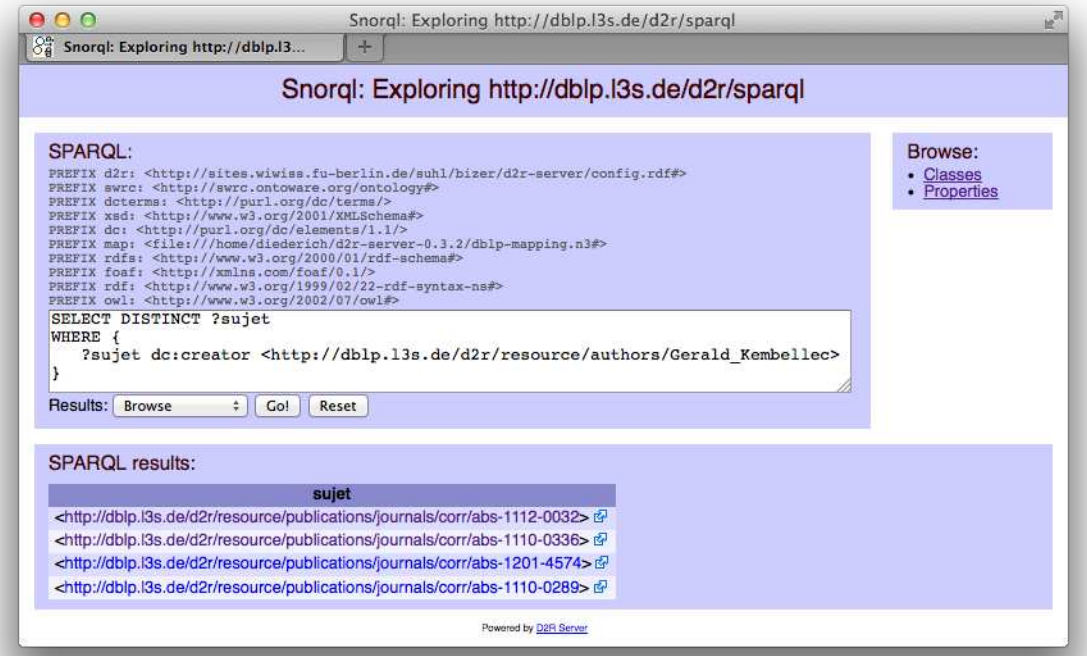


Figure 9.10: Recherche sur le SPARQL *endPoint* de l'entrepôt DBLP

de l'INRIA [1] ou Jena des laboratoires HP en collaboration avec Apache.org [2]. L'outil se chargera d'effectuer les requêtes au format SPARQL sur un fichier RDF local ou distant.

Exemple d'utilisation de SPARQL

Dans la capture d'écran présentée en figure 9.10, nous présentons une requête lancée dans l'interface SNORQL liée à l'entrepôt DBLP. La requête demande de renvoyer les sujets (sous forme d'URI) des articles de la base dont l'auteur (*creator* en dublin core) est Gérald Kembellec représenté sous la forme d'une URI. Le résultat est également présenté sous forme d'une liste d'hyperliens vers des URI représentant les articles.

Si nous prenons l'exemple d'Isidore, la préoccupation d'urbanisation est telle que le site propose une interface classique de recherche dite IHM [3] avec une compatibilité au

1. `http://wimmics.inria.fr/corese`, accédé le 1er août 2012.
2. `http://jena.apache.org/`, accédé le 1er août 2012.
3. `http://rechercheisidore.fr/`, accédé le 1er août 2012.

glanage, mais également des connecteurs pour que les applications distances puissent poster des requêtes et recevoir des réponses dans divers formats [1]. Enfin, les concepteurs d'Isidore ont ouvert un *triple store RDF* pour exposer les fiches aux formats du web sémantique. De plus ils ont même intégré un portail d'interrogation SPARQL pour permettre aux chercheurs d'affiner leurs requêtes [2]. Cet exemple d'ouverture n'est pas une généralité, mais n'est pas non plus unique comme nous avons pu le montrer avec DBpedia, DBLP e WordNet. La tendance en cette deuxième décennie du 21e siècle au cours de laquelle on parle de plus en plus d'*openData* et de web de données consiste donc précisément à ouvrir ses données.

9.5 Conclusion

Nous notons donc deux cadres distincts d'urbanisation de systèmes d'information. Le premier est lié à la navigation d'un hypertexte, c'est dans un contexte de lecture que l'outil de navigation va découvrir et/ou extraire de l'information documentaire vers un logiciel de gestion de ressources bibliographiques. L'autre cadre d'urbanisation d'un SI documentaire (*library mashup*) est le moissonnage. Ce système consiste à fournir des services de notices bibliographiques au travers du protocole HTTP. Plus systématique, le moissonnage permet d'explorer une ou plusieurs bases de connaissances afin de réunir de l'information sur un sujet particulier.

L'urbanisation des systèmes d'information n'est pas une préoccupation récente, mais elle se voit ramenée au goût du jour par les progrès techniques en informatique et la volonté des moteurs de recherche de promouvoir l'information correctement présentée. Mais, surtout les systèmes automatisés de traitement de l'information sont capables d'extraire contextuellement ou systématiquement les métadonnées liées aux objets documentaires de manière automatisée. Un utilisateur de Zotero, Mendeley ou tout autre logiciel capable de glaner des notices sur un système d'information va être aidé dans sa tâche de recherche. En effet, quand on se réfère aux modèles cognitifs de recherche d'information, il est évident que si l'usager n'a pas besoin de rédiger manuellement sa notice bibliographique, il disposera de plus de temps à consacrer au autres étapes

1. `http://rechercheisidore.fr/api`, accédé le 1er août 2012.
2. `http://rechercheisidore.fr/sparql`, accédé le 1er août 2012.

telles que la sélection, la lecture et la compréhension. De plus, au sein du système d'information globalisé, l'urbanisation permet une meilleure indexation par l'agent Panda du moteur de recherche Google.

Pour clore cette longue métaphore filée sur le système urbain appliqué à l'information documentaire, rappelons que selon sa provenance on n'utilise pas la même porte d'entrée dans une ville. Certains systèmes d'information distants utiliseront la contextualisation à travers le HTML, d'autres préférerons la systématisation d'une moisson au travers d'OAI-PMH, de Z3950 ou du RDF. L'essentiel pour offrir un système documentaire fonctionnel est de proposer, en plus de la visualisation des données, une méthode contextuelle d'exposition des métadonnées (glanage) et/ou un service de moissonnage sous forme de silo de données.

Chapitre 10

Panorama de taxonomies en informatique

Si les nuances infinies du langage ne s'accommodent point des classifications rigides qu'on veut faire, tant pis pour les classifications. La science doit s'accommoder à la nature. La nature ne peut s'accommoder à la science..

Ferdinand Brunot

La Pensée et la Langue (1922)

Introduction

Lors de la recherche de documents scientifiques pour la constitution de bibliographie, du bruit et du silence entourent le résultat des requêtes classiques. Cette observation se vérifie sur les moteurs de recherches généralistes, mais aussi sur les bases de connaissances scientifiques. Rappelons que notre objectif final est de proposer une alternative « *look and feel* » aux moteurs de recherche scientifiques qui nécessitent généralement des requêtes à la syntaxe complexe. Selon Roman (2010), les informaticiens peinent à gérer la surabondance des données en ligne. Cette masse de connaissances est écrasante à la fois pour les utilisateurs finaux, comme les chercheurs, ainsi que pour les départements de recherche et développement des sociétés. Nous tentons de fournir un outil valable pour l'ensemble de ces personnes afin de fournir un accès facilité à un maximum d'articles publiés, et ce, à travers une carte du domaine de la recherche concerné, ici l'informatique. Une partie de ce travail a consisté à choisir un système de classification pour établir la cartographie sémantique du domaine de recherche. Classifications, catalogues et thésaurus sont autant de découpages de la réalité qui remportent un succès mitigé auprès des publics (Denecker *et al.*, 2000, p. 46). Toujours selon Denecker, Le « thésaurus d'une bibliothèque (...) ne saurait, tout à la fois, répondre favorablement aux interrogations du néophyte et du spécialiste (Denecker *et al.*, 2000, p. 46) ». Pour assister l'usager, professionnel ou chercheur en informatique, nous orientons notre choix vers un système de classification de l'information qui fasse sens pour le plus grand nombre et qui ait une représentation la plus exacte possible du domaine. Notre réflexion sur le choix d'une taxonomie dans le domaine de l'informatique est décrite dans cette partie du mémoire.

En sciences informatiques, les systèmes de classification les plus courants sont la Computing Classification System (CCS) d'ACM (1998a) (Mirkin *et al.*, 2010) et la taxonomie IEEE (2002) qui étend le modèle d'ACM. Nous allons également examiner un système de classification plus générique, quant aux champs de recherche couverts, mais spécialisé dans la catégorisation des documents : la Classification Décimale de Dewey (DDC). Deux autres classifications pertinentes du domaine informatique sont proposées par le *Journal of Universal Computer Sciences* pour l'indexation de ses propres archives, et l'autre la classification utilisée par la bibliothèque de l'Université Hébraïque de Jérusalem. Commençons par établir une définition de ce que nous entendons par « taxonomie du

domaine informatique », et expliquer notre choix de ce terme pour désigner un système de classification arborescent.

10.1 Définition et usages de Taxonomie

Avant de commencer notre étude comparative, il convient de convenir du sens à donner au terme de taxonomie, qui est en rapport étroit avec les sciences de l'information et de la communication. Ce terme est issu du grec *taxis* « mise en ordre » et *nomos* « loi », ce qu'il est raisonnable de traduire par « loi de structuration ». Nous expliciterons la différence qui existe entre les notions de thésaurus et de taxonomie.

Définition documentaire du thésaurus (ADBS, 2012)

> Liste organisée de termes normalisés servant à l'indexation des documents et des questions dans un système documentaire. Les descripteurs sont reliés par des relations sémantiques (génériques, associatives et d'équivalence) exprimées par des signes conventionnels. Les synonymes (non-descripteurs) sont reliés aux descripteurs par la seule relation d'équivalence. On peut distinguer les thésaurus en fonction :
> – du mode de regroupement des termes (thésaurus à facettes) ;
> – de la variété linguistique des termes (mono- ou multilingue) ;
> – des domaines de connaissances (spécialisé, sectoriel ou encyclopédique).

Définition documentaire (parcélaire) de taxonomie (ADBS, 2012)

> On peut retenir deux concepts distincts de taxonomie selon que l'on met en avant des critères structurels (une structuration hiérarchique) ou des critères fonctionnels (usage pour l'organisation de l'information).
>
> D'un point de vue structurel, on parlera alors de taxonomies (de termes, de classes, de concepts) pour désigner la hiérarchie ou l'arborescence autour de laquelle sont construits différents types d'instruments, comme les thésaurus, les réseaux sémantiques ou les ontologies.
>
> D'un point de vue fonctionnel, une taxonomie est un cadre d'organisation pour des ressources numériques de toute nature (et pas seulement documentaires), destiné à en permettre une présentation ordonnée et y donnant accès par navigation hypertextuelle.

Selon le wiki de l'Université René Descartes [1], en biologie, la science de la systématique utilise les taxonomies, ou plus précisément les taxinomies, pour « établir une classification systématique et hiérarchisée des organismes vivants en les regroupant selon les caractères qu'ils ont en commun, des plus généraux aux plus particuliers, dans diverses catégories emboitées les unes dans les autres, nommées *taxons* ». En informatique, le terme de taxonomie ou taxinomie (*taxonomy* en anglais) désigne une méthode de classification des informations dans une architecture structurée de manière évolutive. Nous conserverons à l'esprit certains aspects de ces différentes définitions. Premièrement, une taxonomie doit structurer de manière fine un domaine de connaissance, en le décrivant le plus fidèlement possible. Deuxièmement, elle est structurée de manière arborescente en partant des éléments les plus génériques vers les plus spécifiques. Troisièmement, une taxonomie doit être évolutive, elle doit suivre les avancées scientifiques, que ce soit en terme de découverte ou d'évolution technique, au même titre que les autres langages documentaires (comme les thésaurus) qui font l'objet d'une maintenance régulière.

Si l'on se replace dans le contexte du web sémantique très lié au web documentaire, il faut faire une différence entre taxonomie, thésaurus et ontologie. Si l'on se réfère à Porquet (2005), il est possible de dégager des définitions fonctionnelles.

Définition fonctionnelle orientée « Web de données » de taxonomie, (Porquet, 2005)

> Dans une taxonomie, le vocabulaire contrôlé est organisé sous forme hiérarchique simple. Cette hiérarchisation correspond souvent à une spécialisation. Il existe donc un lien précis entre un terme du vocabulaire et ses enfants. Ce lien donne un sens supplémentaire, une signification. D'un vocabulaire contrôlé, on passe à un vocabulaire organisé.

Définition fonctionnelle orientée « Web de données » de thésaurus, (Porquet, 2005)

> Un thésaurus est une taxonomie qui fonctionne dans les deux sens. La taxonomie permet d'obtenir une spécialisation des termes employés. Le thésaurus donne de l'information sur les sujets connexes également. On peut donc restreindre ou élargir le champ de connaissance. Cet élargissement se fait en donnant les termes relatifs. Des liens qui permettent la spécialisation, on peut alors dire : c'est une sous-catégorie (spécialisation) ou est « relatif à » ou « voir également » (élargissement).

1. `http://wiki.univ-paris5.fr/wiki/Taxinomie`, accédé le 1er août 2012

Nombre de résultats au 10 janvier 2011 pour les moteurs suivants :	Taxonomie	Taxinomie
Sur Altavista.fr	1 100 000	311 000
Sur Bing.fr	96 900	28 000
Sur Google.fr	145 000	62 400
Sur ASK.fr	35 200	3 700

Tableau 10.1: Usage constaté des termes taxonomie et taxinomie sur Internet

10.1.1 Taxonomie ou taxinomie, historique terminologique

Le terme taxonomie, créé par un botaniste suisse, A. P. de Candolle (1813), à partir de « taxon », s'est implanté dans l'usage comme synonyme du terme taxinomie. Selon la logique de la langue qui est usitée pour réaliser cette thèse, nous devrions choisir l'emploi courant du terme correspondant au français. Mais à l'heure de la rédaction de ce mémoire, il semble que puissions aussi bien employer *taxinomie* que *taxonomie*, l'histoire linguistique n'étant pas à même de déterminer lequel de ces deux mots est d'origine franco-romane ou anglo-saxonne. Cependant, d'après l'usage (cf. Tableau10.1), *taxonomie* étant plus largement employé, nous nous baserons sur ce constat pour utiliser ce dernier terme afin de désigner un système de classification arborescent. De plus, selon Fischer et Rey (1983), si ce terme est « étymologiquement contesté », il n'en est pas moins d'usage courant.

10.2 La Classification Décimale de Dewey

Le système de Classification Décimale de Dewey (CDD ou DDC) est *quasi* universel. Il s'agit d'une méthode de classification générique utilisée par les bibliothèques du monde entier, et depuis 2003 elle s'est dotée d'une partie décrivant assez finement l'informatique.

004.026	Droit de l'informatique	0	00	004	004.0
004.1	Types spécifiques d'ordinateurs : généralités	0	00	004	004.1
004.16	Micro-ordinateurs	0	00	004	004.1
004.165	Micro-ordinateurs (sous-classement alphabétique par nom).	0	00	004	004.1
004.19	Calculatrices électroniques	0	00	004	004.1
004.2	Analyse et conception des systèmes. Architecture des ordinateurs. Évaluation des performances	0	00	004	004.2
004.21	Analyse et conception des systèmes. Conduite de projets	0	00	004	004.2
004.22	Architecture des ordinateurs	0	00	004	004.2
004.24	Évaluation des performances	0	00	004	004.2
004.25	Analyse et conception des systèmes, architecture des ordinateurs, évaluation des performances d'ordinateurs spécifiques	0	00	004	004.2
004.3	Modes de traitement	0	00	004	004.3
004.33	Traitement en temps réel	0	00	004	004.3
004.35	Multitraitement ; traitement en parallèle	0	00	004	004.3
004.36	Traitement réparti. Client-serveur	0	00	004	004.3
004.5	Stockage des données	0	00	004	004.5
004.53	Mémoire interne (dont ROM et RAM)	0	00	004	004.5
004.56	Mémoire externe	0	00	004	004.5
004.563	Disques magnétiques	0	00	004	004.5
004.565	Mémoires optiques (CD-ROM)	0	00	004	004.5
004.6	Interfaçage et communications. Télématique	0	00	004	004.6
004.61	Interfaçage et communications pour des types particuliers d'ordinateurs	0	00	004	004.6
004.62	Protocoles d'interfaçage et communications (p. ex. TCP/IP)	0	00	004	004.6
004.64	Équipements (à bande de base, à bande large, modems, câbles en fibre optique, contrôleurs de périphériques en général)	0	00	004	004.6
004.65	Architecture et conception des réseaux de communication	0	00	004	004.6
004.66	Mode transmission des données, modes de communication des données (par paquets, multiplexage, ..)	0	00	004	004.6
004.67	Réseaux étendus (WAN)	0	00	004	004.6
004.678	Internet	0	00	004	004.6
004.68	Réseaux locaux (LAN). Intranet.	0	00	004	004.6

Figure 10.1: Classification décimale de Dewey (extrait de la classe 000 dédiée à l'informatique)

10.2.1 Historique de la CDD

Melvil Dewey, né Melville Louis Kossuth Dewey (1851-1921), a inventé le système de classification lorsqu'il travaillait comme bibliothécaire assistant à l'Amherst College [1] où il était étudiant en 1873-1874. Dans l'*Encyclopedia of Library and Information Science*, Mitchell et Vizine-Goetz (2009) précisent que la conception de la CDD a eu lieu en 1873 : « The DDC was conceived by Melvil Dewey in 1873 and first published in 1876 ». Depuis 1876, 22 versions de la CDD ont été proposées jusqu'à celle de janvier 2009 [2]. Tous les droits d'auteur du système Dewey Decimal Classification appartiennent actuellement à l'OCLC depuis le rachat en 1988 du nom et de la Fondation *Forest Press*, créée par Dewey lui-même pour gérer la CDD.

1. `https://www.amherst.edu/aboutamherst/history/timeline`, accédé le 1er août 2012
2. `http://www.oclc.org/ca/fr/dewey/updates/enhancements/`, accédé le 1er août 2012

10.3 Le système de classification ACM

La version actuelle du système de classification de l'*Association for Computing Machinery*, qui est encore en vigueur en 2012 (selon l'ACM elle-même dans l'introduction du CCS), date de 1998. Cette dernière version était elle-même basée sur la version de 1991. La première version de cette classification a été proposée en 1964.

10.3.1 Le principal document de l'ACM CCS

L'arbre de classification, ou la taxonomie, se compose de 11 noeuds de premier niveau avec des lettres (de A à K) pour identifiant.

Chaque classe de premier niveau possède une ou deux sous-classes identifiées avec l'identifiant de la classe de premier niveau, auquel on ajoute un numéro (par exemple B.1.).

Les classes de deuxième et troisième niveau se déclinent à travers des thèmes indexés par des identifiants numériques de forme suivante : B.1.1.

Toutes les classes de premier et deuxième niveau intègrent une sous-classe intitulée « Général » dont l'identifiant est « g » et une sous-classe « Divers ». Les sous-classes « Divers » sont utilisées pour classer des documents spécifiques qui ne peuvent être strictement classés dans les autres sous-classes de la catégorie, et leur identifiant se termine toujours par « m » (B.1.m).

Le quatrième et dernier niveau de l'arbre de catégorisation ne possède pas d'identifiant, il est comparable aux feuilles d'une structure arborescente. Les libellés de ce niveau correspondent à un vocabulaire contrôlé (de descripteurs de sujet) relatif au niveau de catégorisation directement supérieur.

La CCS est mise à disposition par l'ACM sous forme de document web hypertexte dans le Portail des ACM. Elle est également disponible dans le format ASCII ou sous forme de fichier XML. Des renvois existent entre catégories pour traiter des sujets connexes (voir figure 10.2). La classification de descripteurs de sujet a été initialement pensée pour être non-exhaustive. Avec les constantes évolutions de la technologie, il était évident que de nouveaux éléments allaient devoir être ajoutés ou retranchés à la taxonomie. Ce raisonnement est également valable pour d'autres éléments de la taxonomie. C'est la raison pour laquelle certains labels sont suivis d'un double astérisque, qui signifie que le terme est devenu obsolète. Cependant, dans la pratique

Figure 10.2: Fichier principal de l'ACM CCS, en ASCII(1), HTML(2) et XML(3)

les descripteurs obsolètes ne sont pas retirés pour permettre de rechercher d'anciens articles. En effet, les articles scientifiques survivent à la technologie, et avec eux, leur indexation.

10.3.2 Les fichiers connexes à l'ACM CCS

Premièrement, la liste des descripteurs de sujet implicite, dans l'ACM CCS, complète la taxonomie. Les « descripteurs implicites de sujet » ou *Implicit Subject* Descriptors (également appelés « noms propres descripteurs ») spécifient les termes génériques de la CCS (voir l'extrait présenté dans le tableau 10.2).

Ces descripteurs de sujet sont des noms de logiciels, des langues et des gens célèbres dans le domaine de l'informatique. Ces noms ne figurent pas dans le fichier d'origine parce qu'ils sont trop nombreux pour être inclus dans un schéma formel sans le rendre illisible.

Deuxièmement, la liste de « termes génériques » ou *General Terms* est une liste

CCS Nodes	*Node label*	*Implicit descriptors*
...	...	...
A.0	*General Literature*	Alan Turing
K.2	*History of computing*	Alan Turing
K.7.2	*Organizations*	ACM
C.5.3	*Microcomputers*	iPhone
I.7.2	*Document preparation*	XML
...	...	...

Tableau 10.2: Extrait de la liste des descripteurs implicites d'ACM

de termes qui peuvent aider à décrire plus précisément un concept que l'intitulé d'une classe de CCS. Ces termes sont des noms obligatoirement communs qui spécifient leur classe (nœud de l'arbre) de rattachement.

Voici quelques exemples de termes génériques : algorithmes, conception, documentation, économie, expérimentation, facteur humain, langage de programmation, aspect juridique, gestion, évaluation, performance, fiabilité, sécurité, normalisation, théorie, vérification.

Ces éléments peuvent être sémantiquement liés à un ou plusieurs noeuds de l'arbre ACM CCS. Les éléments de cette taxonomie, présentés précédemment, seront repris plus en détails dans le chapitre 11 de réalisation d'un outil de recherche d'information.

D'autres éléments qui peuvent être très utiles pour décrire le domaine des connaissances en informatique sont les termes découverts ou « Discovered Terms » qui équivalent à des mots-clés. Ces termes sont présentés par ACM comme « générés » par des algorithmes propriétaires à partir d'un ensemble de classifications natives des sujets ACM et des mots clés fournis par les auteurs. En réalité, les *discovered terms* sont plus que des mots clés, ils sont le résultat d'une synthèse statistique entre la fouille systématique (plein texte) du corpus ACM avec une pondération selon la position des termes dans les textes (titres, résumés, corps du texte), et les mots clés auteurs (cf. figure 10.3). Il est important de noter que la liste de ces termes est dynamique, elle est générée automatiquement par le portail ACM. Notons que les deux autres listes de descripteurs sont des listes statiques, présentées sous forme de fichiers. Ces listes sont

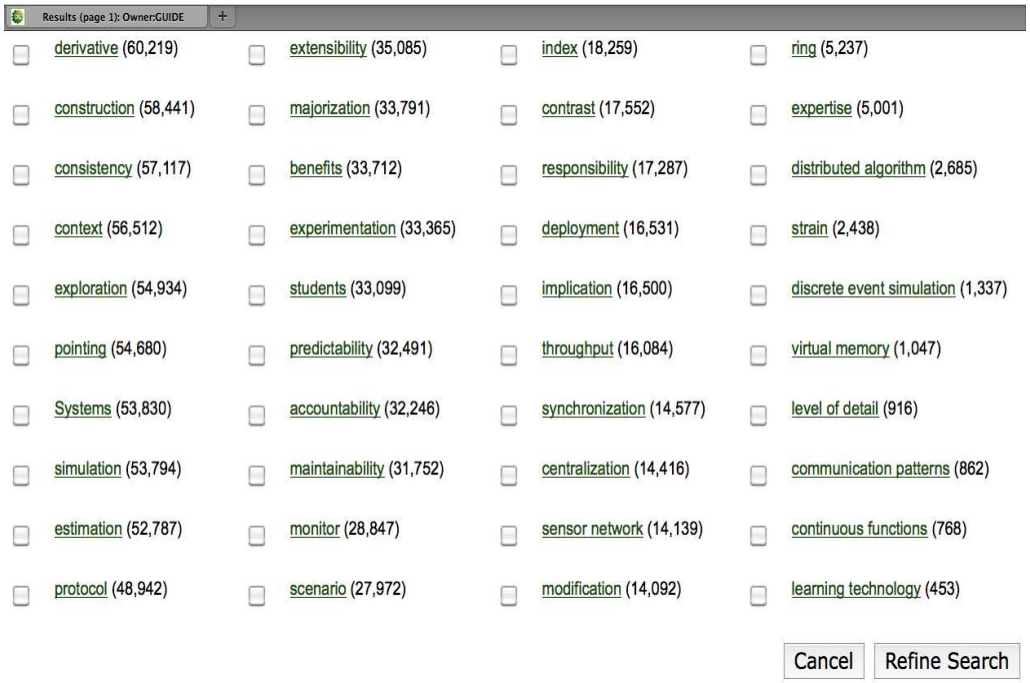


Figure 10.3: ACM CCS Discovered Terms

révisées et augmentées régulièrement, mais ne sont pas dynamiquement générées. Le projet *discovered terms* a été récemment abandonné par ACM car les résultats proposés n'étaient pas jugés significatifs.

10.4 IEEE ACM CCS étendu

L'Institute of Electrical and Electronics Engineers (IEEE) propose une classification [1] qui étend l'ACM, en ajoutant des identifiants aux descripteurs de sujets (les feuilles de la taxonomie) qui correspondent au 4e niveau de la structure arborescente et qui, pour mémoire, n'étaient pas codés dans la CCS. Cette classification est plus utile que celle originellement proposée par ACM dans le cas d'un système de recherche fondé sur une base de données relationnelle.

Dans un travail antérieur d'intégration à une base de données (sous MySQL), nous avons utilisé une version modifiée de celle-ci. Pour ce faire, nous avions décidé de ne pas autoriser l'utilisation des lettres « g » et « m » comme identifiants de classification, car

1. `http://www.ieee.org/documents/2009Taxonomy_v101.pdf`, accédé le 1er août 2012

ACM id	*IEEE id*	*ACM Label*	*IEEE Label*
H.3.3	H.3.3	Information Search & Retrieval	idem
uncoded	H.3.3.a	Clustering	idem
uncoded	H.3.3.b	Information filtering	idem
-	H.3.3.c	-	Internet search
-	H.3.3.d	-	Metadata
uncoded	H.3.3.e	Query formulation	idem
uncoded	H.3.3.f	Relevance feedback	idem
uncoded	H.3.3.g	Retrieval models	idem
uncoded	H.3.3.h	Search process	idem
uncoded	H.3.3.i	Selection process	idem

Tableau 10.3: Comparaison quatre premiers niveaux de l'ACM & IEEE CCS

elles sont réservées pour établir l'identité des sous-classes « général » et « divers ». L'intérêt du système de classification IEEE ne s'arrête pas à une commodité d'*urbanisation de système d'information*, le tableau 10.3 nous montre que si la taxonomie IEEE étend l'ACM CCS avec des identifiants au dernier niveau, elle offre également plus de descripteurs de sujet. C'est la raison pour laquelle cette taxonomie est appelée Système de classification Informatique ACM « étendu ».

10.5 Liste de sujets de l'ACM J. UCS

L'IEEE propose une façon alternative de définir la taxonomie elle-même en fournissant un identifiant à des termes qui n'en avaient pas, voire à ajouter d'autres termes, tous codés. Dans le cadre du *Journal of Universal Computer Sciences* [1], la Graz University of Technology [2] et l'Université de Sarawak en Malaisie [3] proposent une classification de sujets relatifs à l'informatique très pertinente. Pour préciser le contexte de ce journal,

1. `http://www.jucs.org/jucs`, accédé le 1er août 2012
2. `http://portal.tugraz.at/portal/page/portal`, accédé le 1er août 2012
3. `http://www.unimas.my`, accédé le 1er août 2012

notons que le J. UCS paraît mensuellement depuis 1995 et se présente comme une publication électronique qui traite de tous les aspects de l'informatique.

10.5.1 Fichier principal de la taxonomie J. UCS

Le J. UCS utilise le début de la CCS ACM comme système de classification, mais offre un supplément de deux nœuds principaux, juste sous la racine de la taxonomie initiale. Selon le site J. UCS, Les Thèmes de « A » à « K » correspondent à la classification ACM avec ses subdivisions, les sujets de « L » et « M » ont été ajoutés afin de refléter le développement de la discipline informatique. Les libellés de ces deux dernières catégories sont les suivants : « science et technologie de l'apprentissage » et « gestion des connaissances ». De notre point de vue, ces ajouts sont utiles, car ils permettent de traiter spécifiquement des champs d'application pragmatiques des sciences informatiques comme la gestion des connaissances et l'apprentissage. On pourrait en imaginer d'autres comme le traitement automatique du langage ou la recommandation. La nature des nœuds « L » et « M » diffère donc des autres nœuds de la CCS originelle qui sont plus techniques en termes de libellés. L'accent mis sur la partie apprentissage et sur l'aspect communicationnel de l'informatique peut cependant être considéré comme exagéré par des « partisans » la technique en sciences informatiques.

10.5.2 La liste des termes supplémentaires du J.UCS

L'UCS J. propose une liste de termes clés nommée « Mots et phrases ». Cette liste est triée par ordre alphabétique. La liste des termes est tout à fait la même que la liste ACM CCS Mots-clés originelle de 1991. Les termes sont précédés par des symboles qui sont énumérés ci-dessous :

- * : Noeud de premier niveau
- ** : Noeud de second niveau
- *** : Noeud Troisième niveau
- SD : Subject Descriptor (descripteur de sujet)
- GT : General Term (terme général)
- ° : Modified entry (élément mis à jour)

Dans la classification « étendue » (les nœuds supplémentaires de haut niveau et leurs sous-niveaux), on peut remarquer que l'association de ces termes avec les nœuds de la classification n'est pas optimale. Par exemple, le thème « L.1.5-hypertextes et hypermédias » n'est pas lié à l'hypermédia dans la liste des descripteurs de sujet. Le terme « Hypertexte et Hypermédia » est associé (comme descripteur de sujet) au noeud « I.7.2 » labellisé « préparation de documents ». Donc, si le classement étendu existe et offre une granularité plus fine de la recherche, l'association des mots clés n'est pas complètement réalisée. Par conséquent, des améliorations pourraient encore être apportées à cette version de la CCS.

10.6 Le Système de classification HUJI

10.6.1 La classification des sujets HUJI

La bibliothèque de l'Université Hébraïque de Jérusalem [1] possède son propre système de classification de l'informatique. Cette catégorisation à index numérique est basée sur la première taxonomie d'ACM(ACM, 1964) qui est classée avec des identifiants numériques. Le système de classification est uniquement utilisé à la Bibliothèque des sciences informatiques de l'Université de Jérusalem pour répondre aux besoins spécifiques des utilisateurs en informatique et mathématiques. Le système est maintenu et mis à jour par les membres du personnel. Les index de cette classification sont compris entre les nombres 210 et 282. Chaque index numérique représente une réserve documentaire physique, comme une étagère (Library of The Hebrew University of Jerusalem, 2009). Par exemple : Les livres sur la recherche d'information sont classés avec l'indice 237.

10.6.2 Liste alphabétique des mots-clés

Un vocabulaire contrôlé complète la classification HUJI. Dans le tableau 10.4, un échantillon du fichier montre tous les termes associés au nœud identifié par « 241.2 » et labellisé « compilateurs, générateurs de compilateurs ». Dans une optique de réalisation de cartographie sémantique, nous avons porté la taxonomie HUJI au format XML (cf. figure 10.4), ce qui permet également une meilleur visualisation qu'un fichier en texte

1. `http://www.ma.huji.ac.il/~library/`, accédé le 1er août 2012

ids	HUJI keywords
241.2	Code generation
241.2	Code optimization
241.2	Compiler generators
241.2	Compilers
241.2	Lexical analysis
241.2	Parsing
241.2	Semantic analysis

Tableau 10.4: Extrait des mots clés HUJI

brut. Nous n'avons remarqué que peu de liens de proximité sémantique (liens de type *see also*) au sein de la structure de l'arborescence. De plus, les termes descripteurs sont rarement utilisés comme pivot entre deux classes ou sous-classes de la classification. Enfin, s'il existe quelques renvois internes à la taxonomie, on trouve en revanche beaucoup de liens externes vers des identifiants de la classification ACM.

10.7 La classification CORR Subject Areas d'ArXiV

Les articles proposés dans la partie informatique de la plateforme de dépôt ArXiV (CoRR), sont classés de deux façons distinctes :

1. Directement en utilisant la classification ACM .
2. En utilisant les éléments de la taxonomie du *C0mputing Research Repository Subject Areas*[1] (CORR).

Le système de classification ACM (voir partie 10.3.1) fournit un système relativement stable qui couvre l'ensemble des sciences informatiques. De plus, au sein de chaque classe taxonomique principale de l'arborescence CORR, les thèmes ACM traités sont explicitement mentionnés (voir figure 10.5), comme dans le cas de la classification HUJI. Cela nous conforte dans l'idée de la domination de la classification ACM dans la thématique scientifique de l'informatique. Selon les auteurs de CORR, les sous-domaines ne s'excluent pas mutuellement, pas plus qu'ils ne fournissent pour l'instant

1. `http://arxiv.org/corr/subjectclasses`, accédé le 1er août 2012

```
- <node id="235" label="Administrative data processing">
    <isRelatedTo>J.1</isRelatedTo>
  - <isComposedBy>
      <node id="235.1" label="Decision tables"/>
      <node id="235.2" label="Financial"/>
      <node id="235.3" label="Electronic Mail"/>
    - <node id="235.7" label="Transportation, Communication">
        <seeAlso>262.36</seeAlso>
      </node>
      <node id="235.8" label="Management systems, Information systems, EDP"/>
      <node id="235.9" label="Office automation, Word processing"/>
    </isComposedBy>
  </node>
- <node id="236" label="Artificial intelligence, Intelligent agents">
    <seeAlso>258.1</seeAlso>
  - <isComposedBy>
      <node id="236.1" label="Robotics"/>
      <node id="236.2" label="Learning, Adaptive systems"/>
    - <node id="236.3" label="Pattern recognition, Image processing">
      - <isComposedBy>
          <node id="236.31" label="Digital, signal processing"/>
          <node id="236.32" label="Optical, Vision"/>
          <node id="236.33" label="Audio, Speech"/>
          <node id="236.34" label="Image compression"/>
        </isComposedBy>
      </node>
      <node id="236.4" label="Problem solving, Problem planning"/>
      <node id="236.5" label="Simulation of natural systems, Genetic algorithms"/>
      <node id="236.6" label="Heuristic methods, Expert systems, Fuzzy systems"/>
      <node id="236.7" label="Symbol manipulation"/>
      <node id="236.8" label="Knowledge representation, Automated reasoning, Fuzzy Logic"/>
      <node id="236.9" label="Natural language understanding, Natural language generation"/>
```


Figure 10.4: Portage en XML de la classification HUJI (extrait)

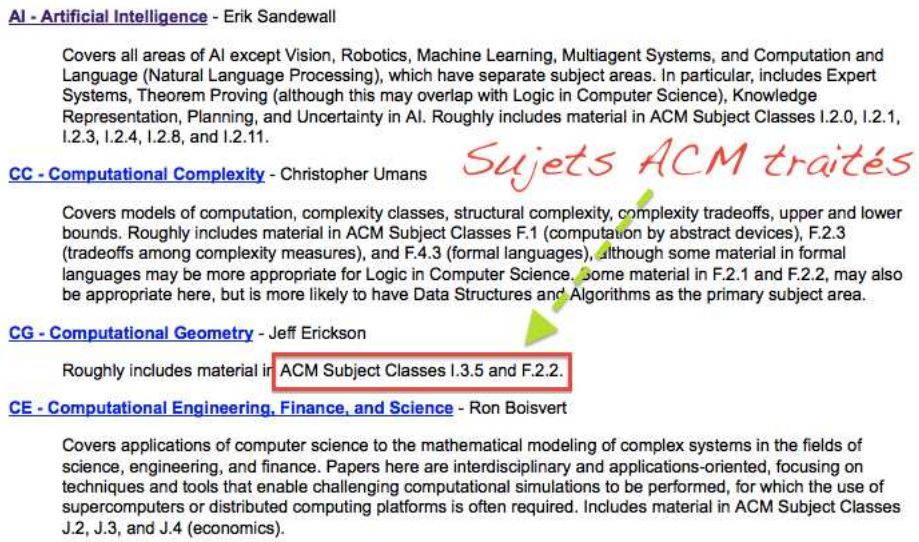


Figure 10.5: Taxonomie CORR (extrait)

une couverture complète du champ. L'objectif revendiqué est de refléter au mieux les thématiques de recherche actives dans le domaine de l'informatique. Dans cette optique, les gestionnaires de la taxonomie se déclarent réactifs aux besoins des chercheurs en terme de classification.

10.8 Conclusion

Dans cette partie, nous avons proposé une définition de ce qu'est une taxonomie et avons également réalisé une étude comparée pragmatique des différentes taxonomies des sciences informatiques. Pour la création de notre outil de recherche d'informations scientifiques et techniques dans le domaine de l'informatique, nous pensions initialement utiliser la classification IEEE étendue, tirée de l'ACM CCS. L'avantage de la version IEEE de la classification ACM est que les termes associés sont présentés comme des concepts, avec un identifiant (ce que ne propose pas l'ACM CCS). Cette pratique donne une forte cohérence à la taxonomie, dans l'optique d'une intégration au sein d'une base de données. Cependant, les identifiants supplémentaires à l'ACM CCS inclus dans la taxonomie IEEE ne sont pas utilisable par le portail ACM pour une recherche (ACM, 1998b). De plus, nous n'avons aucune certitude sur l'alignement de l'IEEE avec les évolutions de l'ACM. Nous allons donc utiliser la taxonomie originale d'ACM en utilisant le principe consistant à fournir un identifiant aux termes descripteurs.

Chapitre 11

OntologyNavigator

Si, le matin, la bibliothèque suggère un reflet de l'ordre sévère et raisonnablement délibéré du monde, la bibliothèque, la nuit, semble se réjouir de son désordre fondamental et joyeux.

Alberto Manguel

La représentation des groupes indiquée dans la figure sur une surface plane est beaucoup trop simple.

Charles Darwin, *De l'origine des espèces*

Introduction

Comme nous l'avons montré dans notre première partie, la recherche d'informations, qu'elle soit d'ordre générique ou scientifique, est un sujet complexe. Une recherche documentaire peut se révéler rapidement déroutante, voire décourageante, plaçant l'usager entre déluge informationnel et absence pure et simple de résultat. Cela est encore plus vrai dans les systèmes de recherche d'informations génériques. C'est pourquoi l'utilisateur de Système de recherche d'information (SRI) doit sélectionner avec soin ses sources d'information pour trouver une information fiable. Même en cas d'utilisation de bases de connaissances spécialisées, l'affichage d'une série de notices bibliographiques exploitables n'est pas une fin en soi. Il faut encore examiner cette liste, sélectionner et retrouver les documents primaires les plus en adéquation avec le besoin informationnel. Ce travail est fastidieux sans assistance technique pour la mise en exergue des métadonnées dans une optique de comparaison. Une aide à la décision sous forme de *Query By Exemple* ou de moteur de recommandations peut être un plus. Le travail de documentation du sujet ne s'arrête pas à cette étape. Il lui reste encore à synthétiser les métadonnées documentaires et le paratexte sous forme de notices bibliographiques et à intégrer les résultats formatés dans une bibliographie compatible avec ses outils de rédaction. Dans notre contexte, le monde de l'enseignement supérieur français, il est courant d'observer des étudiants des 2e et 3e cycles, voire des chercheurs, éprouver de réelles difficultés à rassembler de la documentation sur leur domaine d'études ou de recherches.

La réponse que nous donnons à l'hypothèse principale est qu'il est possible de systématiser le processus de gestion d'une bibliographie. Dans cette partie, nous allons exposer les étapes de la conception d'un artefact ergonomique offrant la possibilité de suppléer à la recherche manuelle traditionnelle par l'outil. Cette interface, que nous désirons appuyer sur une représentation structurelle fiable, permettra d'élargir la perception et la connaissance du domaine de recherche. Pour permettre à l'usager final de se repérer dans l'outil, nous désirons orienter la construction de la base structurelle de connaissance sur une classification qui soit préexistante et qui fasse consensus entre les chercheurs de la discipline.

Dans la première partie sur la RI, nous avons réalisé une étude des bases scientifiques et techniques de connaissance en rapport avec notre domaine de connaissance de prédilection. Cette étude a permis de déterminer les méthodes d'interrogation des bases de données associées. Nous réutiliserons ces technologies pour interfacer notre outil afin de générer des requêtes vers ces bases distantes et intégrerons également un corpus à l'outil.

Ce chapitre commencera par définir notre vision de l'ontologie de domaine sur laquelle s'appuiera notre outil de recherche. Nous poursuivrons par un état de l'art de la recherche d'informations utilisant les ontologies. Nous examinerons également les méthodologies de formalisation d'ontologie, ainsi que les procédures d'interrogation associées. Enfin, nous mettrons en œuvre une interface ergonomique qui reliera tous les éléments du domaine de connaissance, depuis la taxonomie initiale jusqu'à l'interface de navigation dans l'ontologie pour la recherche d'informations scientifiques et techniques. Nous décrirons en particulier les évolutions de notre outil pour mieux répondre aux besoins de l'usager en termes de cognition et d'ergonomie. Ces évolutions nous amèneront à nous interroger sur une méthode efficace de représentation ontologique de domaine.

11.1 Proposition de définition du terme ontologie

Dans le cadre des systèmes de recherche d'informations liés à la connaissance scientifique et technique, nous reprendrons la définition de l'ontologie proposée par Studer *et al.* (1998) dans leur article de référence sur l'ingénierie des connaissances. Ce document propose de définir une ontologie de domaine de connaissance comme une spécification formelle d'une conceptualisation d'un domaine, partagée par un groupe d'experts et acceptée par le plus grand nombre. L'ontologie, toujours selon la définition de Studer *et al.*, se doit d'être formelle dans l'optique d'être lisible par un système d'information. Cela exclut donc *a priori* la langue naturelle. Cette formalisation est établie dans l'optique d'un usage répondant à des contraintes fonctionnelles, telle l'intégration dans une solution logicielle. Si nous examinons l'ontologie d'un point de vue philosophique moderne, imprégné d'une contextualisation technique, nous apercevons un objet de connaissance humaine qu'il est possible d'utiliser pour son bénéfice informationnel. Bachimont, dans son introduction aux ontologies, postule que des catégories formelles

doivent émerger et prendre en compte les nouvelles formes de raisonnement et les nouveaux types de concepts manipulés numériquement (Bachimont, 2007, p. 128).

Ce rapprochement entre formalisme structuro-technique et philosophique des ontologies est présenté par Monnin et Félix comme un apport mutuel pour la représentation des nouveaux paradigmes informationnels sur Internet (Monnin et Félix, 2009) :

> *« L'ontologie des informaticiens est essentielle de ce point de vue [la nécessité d'une réflexion sur les conditions de production d'une ontologie], qui attire notre attention sur son statut d'artefact. Le support écrit en était un lui-même ; non une pure production de signes désincarnée mais un objet technologique complexe répondant à une attente. Mais ce qu'il taisait ou masquait, les artefacts contemporains, numériques, dans lesquels s'incarnent désormais les ontologies, le révèlent. En témoigne la réintégration d'une dimension technique en complément de la dimension sémiotique et langagière, couplage peut-être seul à même de nous ménager un accès renégocié aux catégories ».*

Nous acceptons cette idée que les ontologies de domaines soient une possible évolution naturelle de la structuration numérique des champs de connaissance. Par ontologie de domaine, nous entendons un ensemble de concepts hiérarchisés par un expert au sein d'une structure, liés par des relations de proximité syntaxique ou sémantique.

Dans un domaine donné, une ontologie n'est pas seulement une représentation du champ des connaissances. Une ontologie prétend aussi refléter un consensus sur les interactions entre les éléments du champ de savoir. Toutefois, en pratique, la définition de l'ontologie a été quelque peu diluée, dans le sens où les taxonomies (surtout lorsqu'elles sont enrichies par des thésaurus) sont considérées comme des ontologies à part entière : ce que Monnin et Félix décrivent comme une ontologie générique (Monnin et Félix, 2009). Terminologiquement, les ontologies diffèrent des taxonomies : elles sont structurellement plus riches, ce qui complexifie leur représentation (Studer *et al.*, 1998). La question que soulève cette réflexion est l'influence de la représentation de l'accès à l'information sur la recherche de connaissance. Son impact sur les résultats obtenus par rapport aux recherches plus traditionnelles est-il significatif ? Du point de vue de l'usage, notre solution d'assistance logicielle à la recherche d'information a une forte ambition cognitive. Il est à envisager, espérer même, qu'à terme, la maîtrise de l'outil le rende obsolescent, car source intrinsèque de connaissance (l'usage de l'outil permet la maîtrise du domaine). Dans notre contexte, il s'agit de recherche bibliographique dans le domaine informatique

pour des personnes initiées, mais non expertes, qui de plus sont perdues dans un corpus majoritairement anglophone.

Notre travail commence par la collecte et la sélection d'articles scientifiques relatifs à l'informatique et se poursuit par une intégration à un corpus de documentation. Cette documentation sera représentée sous forme d'arborescence. Cela permettra l'émergence d'une visualisation globale du corpus de textes de la recherche informatique. Cette facilitera également l'accès à l'information recherchée par moteur de recherche en langage naturel, par mots clés, contexte, ou proximité sémantique. Ainsi, et c'est tout l'intérêt du concept, un utilisateur qui ne maîtriserait pas encore l'ensemble du vocabulaire informatique pourrait trouver des articles pertinents en plusieurs langues, articles qu'il n'aurait pas su trouver seul par des méthodes de recherche traditionnelles.

11.2 État de l'art de la recherche d'informations par ontologies de domaine

La démarche classique d'utilisation des ontologies de domaine consiste à hiérarchiser les sou- ensembles du domaine dans une optique de gestion. L'ontologie sert alors le plus souvent à hiérarchiser et classer les éléments composant le domaine, ainsi qu'à décrire leurs relations. Une application courante est l'indexation de corpus spécialisé par ce moyen. Une utilisation plus novatrice de l'ontologie est d'inverser la démarche. Il est possible d'utiliser l'ontologie de domaine comme support de recherche dans un texte, un corpus, une bibliothèque numérique, ou même dans tout Internet. Grâce à une combinaison de différentes technologies sémantiques, Bloehdorn *et al.* (2007) a proposé une méthode intéressante de consultation de bibliothèques numériques. Il a défini une approche par analyse de questions structurées en langage naturel avec une grammaire définie. Il s'agit pour le système de comprendre la manière d'identifier les mots clés, les titres et les auteurs. Par exemple : qui a écrit tel livre ? Quel livre traite d'un sujet défini ? Quel article fait partie de telle conférence et correspond à tels mots clés ? Cette approche traduit le langage naturel en métadonnées, et reformule la question en langage « *SPARQL* ». Comme la réponse se trouve dans un fichier « *Resource Description Framework (RDF)* », les mises à jour en temps réel sont supportées, ainsi

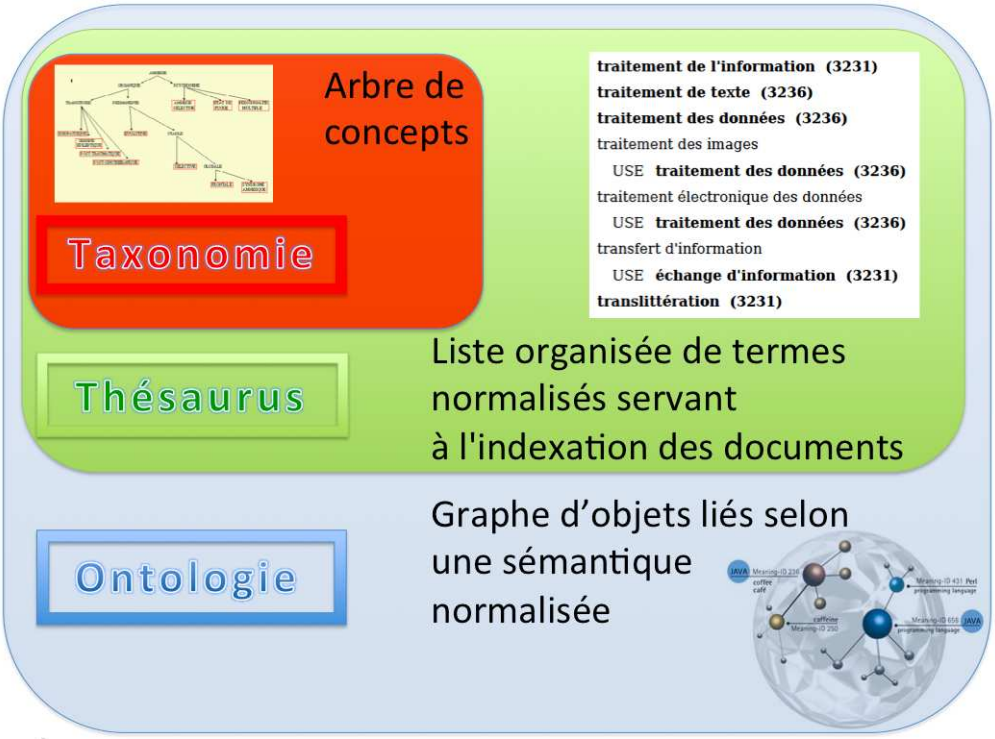


Figure 11.1: Relations entre taxonomie, thésaurus et ontologie de domaine

que l'hétérogénéité des formats et la localisation des ressources. Cette méthode permet de s'abstraire de toute base de données au sens commun du terme.

11.3 Ontologie de domaine informatique, conception d'un modèle exploitable

L'approche de l'accès au savoir sera onomasiologique, ou « *top-down* », c'est-à-dire que le corpus sera représenté dans une structure qui forme un ensemble hiérarchisé, normalisé et fini. Comme démontré par Studer *et al.* (1998), une simple taxonomie ne saurait être considérée comme une ontologie à part entière. Cependant, pour élaborer une ontologie, il faut au départ une structure formelle de classification.

11.3.1 Notion de pertinence utilisateur par désambiguïsation et point de vue

L'informatique est un domaine très vaste, comprenant une multitude de sous-disciplines également en étroite relation avec de nombreux autres domaines scientifiques. Dans le cadre de la recherche d'information (RI), un des problèmes rencontrés dans l'interprétation d'une requête est la possible ambiguïté d'un terme. La parade à ce problème est la désambiguïsation lexicale. Cette technique permet de comprendre le sens d'un terme parmi d'autres sens possibles. Les travaux de Jean Véronis (2003) montrent l'importance du contexte d'un terme, pour en saisir le sens. Pour illustrer son propos, il donne l'exemple du mot cible *station* qui peut être associé à *ski, météo, spatial, travail, radio, Primagaz, eau*, et *ligne* dans son corpus de test. Son algorithme *Hyperlex* permet de désambiguïser les sens de termes par fréquence de co-occurrences. Il faudra donc, autant que possible, s'imprégner de points de vue pour la recherche et saisir le contexte d'étude de l'utilisateur.

Nous pouvons nous inspirer des travaux de Nicolas (2006) basés sur Gillon (2004a) pour fournir un exemple de désambiguïsation, afin de clarifier notre propos. Le terme de *stockage de données* n'aura pas le même sens pour un technicien en assemblage informatique, un ingénieur système ou un documentaliste. Pour le technicien, la représentation qui s'impose du stockage de données est le support physique et sa représentation mentale du concept sera un disque dur ou une clé USB. Le professionnel des systèmes et réseaux aura lui une vision plus large de stockage de données. Il verra les concepts de périphériques, les méthodes de stockages telles les « *NAS* », les redondances de données (niveau de « *RAID* »), la façon dont les informations sont partagées (« *Netbios, NFS, SMB* »...) mais aussi les droits sur les données (lecture, écriture et exécution). Enfin, le documentaliste verra en ce terme principalement un progiciel de SIGB (Système intégré de gestion de bibliothèque) qui gère les prêts, les réservations, le suivi des commandes ou encore l'état des livres. Ces trois professionnels, pointus en leur domaine font un usage différent du terme *stockage de données*, cependant on ne peut pas parler ici de polysémie, mais plutôt de point de vue, terme que nous préférons à *ambiguïté lexicale*.

La question de la pertinence utilisateur se pose donc dans ce cas précis de la RI. Cette observation va grandement influencer l'outil, en l'axant sur l'utilisateur et non

uniquement sur les données. Ce projet doit être une entité à l'utilisation souple, à la portée de l'utilisateur pour l'aider à maîtriser son domaine de connaissance.

11.4 Première ébauche

Nous avons démontré l'intérêt du système de classification du domaine de l'informatique scientifique d'ACM CCS (voir section 10.2). Dans une démarche exploratoire, nous allons tenter de représenter de manière linéaire l'ontologie dégagée de la taxonomie initiale enrichie par deux vocabulaires contrôlés. Cette première heuristique ne prétend pas à l'exploitation ontologique du domaine de connaissance de l'informatique. Il s'agit ici, dans un premier temps, de tester l'évolution de la taxonomie initiale vers un thésaurus comme accès à l'information scientifique.

Objectif

Dans cette partie, nous allons réaliser un hypertexte qui, au sein d'une seule page, va tenter de représenter l'intégralité de la hiérarchie des concepts, mais aussi des relations entre les concepts et les descripteurs. Nous avons pris le parti de présenter la classification sous forme d'arbre afin de symboliser les liens de spécification et de généralisation des concepts. D'autres liens vont permettre de proposer l'accès au vocabulaire de description proposé en annexe par l'ACM, mais jamais encore exploité pour la représentation du domaine de connaissance. Ainsi, nous allons tenter de proposer une visualisation du domaine intégrant les *termes* (vocabulaire d'autorité), et les *noms* liés à la taxonomie (voir l'annexe 10.3.2, page 258). Pour mémoire, rappelons les éléments représentatifs de chaque concept proposés dans la taxonomie (voir partie 10.3.2, page 258) :

1. Les concepts sont des abstractions d'éléments de la classification. Chaque concept possède un identifiant et un littéral, sous forme de chaîne de caractères.
2. Les termes sont des descripteurs génériques spécifiant un concept, ils sont représentés par un littéral, mais n'ont pas d'identifiant unique car un même terme peut être lié à plusieurs concepts.
3. Les descripteurs implicites du sujet (aussi appelés « noms relatifs au sujet ») sont des dénominations de produits, de systèmes, de langage de programmation, et de personnalités dans le domaine de l'informatique. Pour les mêmes raisons que

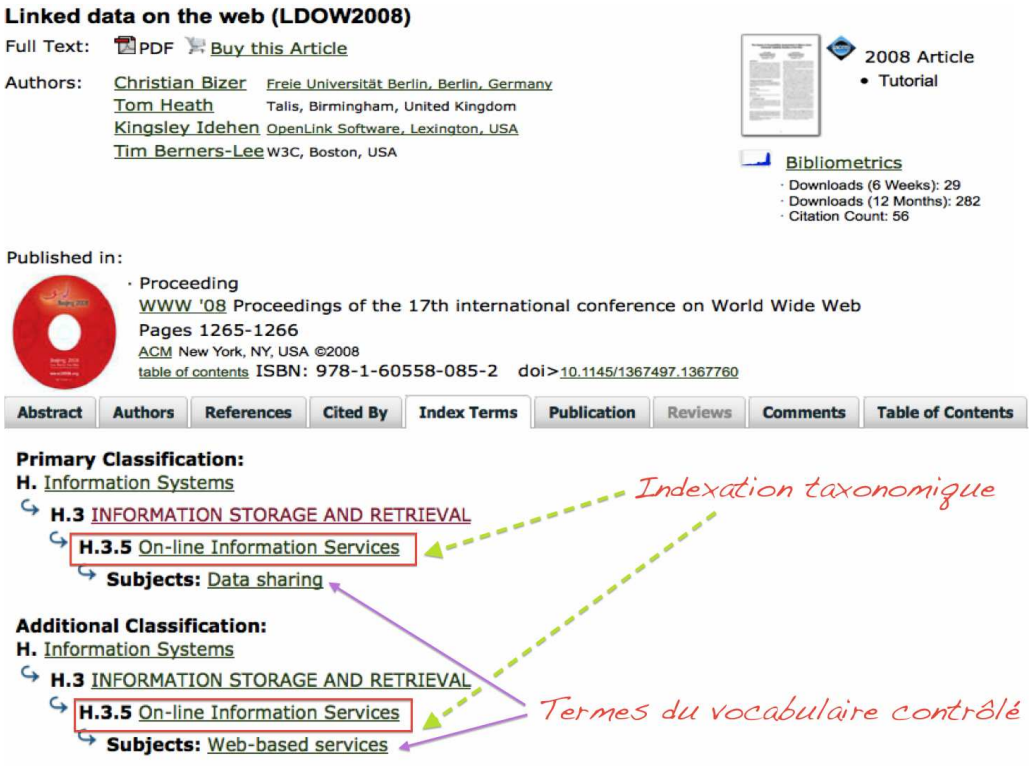


Figure 11.2: Exemple de description d'un document avec la taxonomie ACM

les termes, les descripteurs n'ont pas d'identifiant unique. Il peut s'agir d'acteurs humains emblématiques ou de noms de produits typiquement liés au concept [1]. Cette liste est tenue à jour par les groupes d'experts d'ACM avec plus de régularité que celle des termes, car les noms de produits ou d'acteurs qui la composent sont étroitement liés aux actualités technologique et économique.

Au sein de la classification, tout concept, assimilé à un nœud pour une représentation sous forme d'arbre, possède un certain nombre d'attributs. Chaque concept possède un nœud père et des nœuds fils qui peuvent être des feuilles. Certains concepts sont liés à d'autres par des relations de proximité sémantique. Ce système de classification couvre théoriquement l'ensemble des sciences informatiques et connexes. De manière pragmatique, son usage doit couvrir l'intégralité de la littérature scientifique ayant un rapport avec l'informatique. Chaque production proposée par le portail ACM *Digital*

1. La typicalité est la représentativité d'un élément de langage dans un contexte, voir paragraphe 1.1.2, page 22

Library est décrite par son auteur par les identifiants des concepts traités, ainsi que les termes et descripteurs les plus représentatifs. La figure 11.2 [1] montre que le tutoriel *Linked data on the Web* est classé avec le concept « *On-line Information Services* », ayant pour identifiant H.3.5. Les deux termes associés à cet article sont « *Data sharing* » et « *Web-based services* ».

Ces associations entre les concepts, termes, descripteurs et documents scientifiques devront apparaître sur notre mode de représentation du domaine et seront « navigables », c'est-à-dire hyperliées. Pour accéder à l'information souhaitée, l'utilisateur de notre système devra parcourir l'arborescence de la classification. Le mode de parcours de l'arbre sera théoriquement vertical, sur un modèle de spécification du concept. L'usager devra commencer son processus de recherche par la racine de la classification puis il devra se diriger par spécification.

1. Trouvé sur `http://dl.acm.org/citation.cfm?id=1367760`, accédé le 1er septembre 2012.

A General Literature
B Hardware
C **Computer Systems Organization**
- **C.0** General
- **C.1** Processor architectures
- **C.2** Computer communication networks
- **C.3** Special and application based systems
- **C.4** Performance of systems
- **C.5** **Computer system implementation**
  - **C.5.0** General
  - **C.5.1** Large and Medium (Mainframe) Computers
  - **C.5.2** Minicomputers
  - **C.5.3** **Microcomputers**
    - **Subjects :**
      - Microprocessors
      - Personal computers
      - Workstations
      - **Portable devices (e.g., laptops, personal digital assistants)**
    - **Nouns :**
      - **Apple, IBM, HP, Motorola 6800, Palm, iPad, ...**

Par exemple, si l'on désire retrouver dans l'arbre du domaine de connaissance informatique le concept de tablette tactile, plus précisément celui d'iPad, il faut suivre le cheminement suivant (éléments en rouge). L'encadré expose le cheminement à effectuer pour accéder à un concept, puis aux termes qui le précisent, ainsi qu'aux noms qui le spécifient.

Modélisation et mise en œuvre

Nous avons choisi de convertir toutes les données issues de divers formats (XML, fichier texte) dans un seul format : l'XML. Le fichier des descripteurs implicites accessible en ligne sous forme d'un fichier texte a été traité sous forme de script shell Linux pour être converti. Un processus de journalisation permet d'actualiser ce fichier. Le processus consiste à associer un nom à un ou plusieurs concept(s) à travers d'un lien (voir figure

Source de : http://www.geraldkembellec.fr/ccsnouns.xml

```xml
<?xml version="1.0" encoding="iso-8859-1"?>
<descriptors comment="each implicit descriptor 'implicitdDescritor' is related to one or several CCS node(s)" >
<original file="http://portal.acm.org/lookup/ccsnoun.cfm" author="ACM" />
<translationAuthor
    name="Gerald KEMBELLEC"
    email="gerald.kembellec@paris-sorbonne.fr"
    website="http://www.ontologynavigator.paris-sorbonne.fr"
/>

    <link>
        <node>K.8.1</node>
        <implicitDescriptor>1-2-3/G</implicitDescriptor>
    </link>
    <link>
        <node>D.2.6</node>
        <implicitDescriptor>3DIS</implicitDescriptor>
    </link>
    <link>
        <node>D.3.2</node>
        <implicitDescriptor>8080 assembly language</implicitDescriptor>
    </link>
    <link>
        <node>G.3</node>
        <implicitDescriptor>A-STAT 79</implicitDescriptor>
    </link>
```

Ligne 1, Colonne 1

Figure 11.3: Extrait du fichier XML des descripteurs implicites

11.3). Nous avons traité ensuite l'ensemble du processus grâce à un programme en langage PHP qui charge les deux fichiers XML, celui contenant le fichier taxonomique principal et celui contenant les descripteurs. Les deux fichiers sont entrelacés dynamiquement pour la création d'une vue d'ensemble de la taxonomie. De plus, chaque élément est hyperlié avec le reste du document. D'autre part chaque concept, chaque terme ou chaque nom sera également mis en relation pour offrir un accès direct à la documentation scientifique sur le portail documentaire en ligne d'ACM grâce à la technologie openURL.

Résultats de la première approche

Dans l'optique de permettre à l'usager une meilleure lecture de l'arbre représentant le domaine de connaissance en informatique, nous avons adopté un code couleur dans une feuille de style. Explicitons ce code :

- Noir : Les concepts ;
- Bleu : Les hyperliens vers les articles du portail ACM qui sont en relation avec le concept ou les liens de similarité entre concepts ;
- Vert : Les termes génériques spécifiant le concept associé au nœud supérieur dans

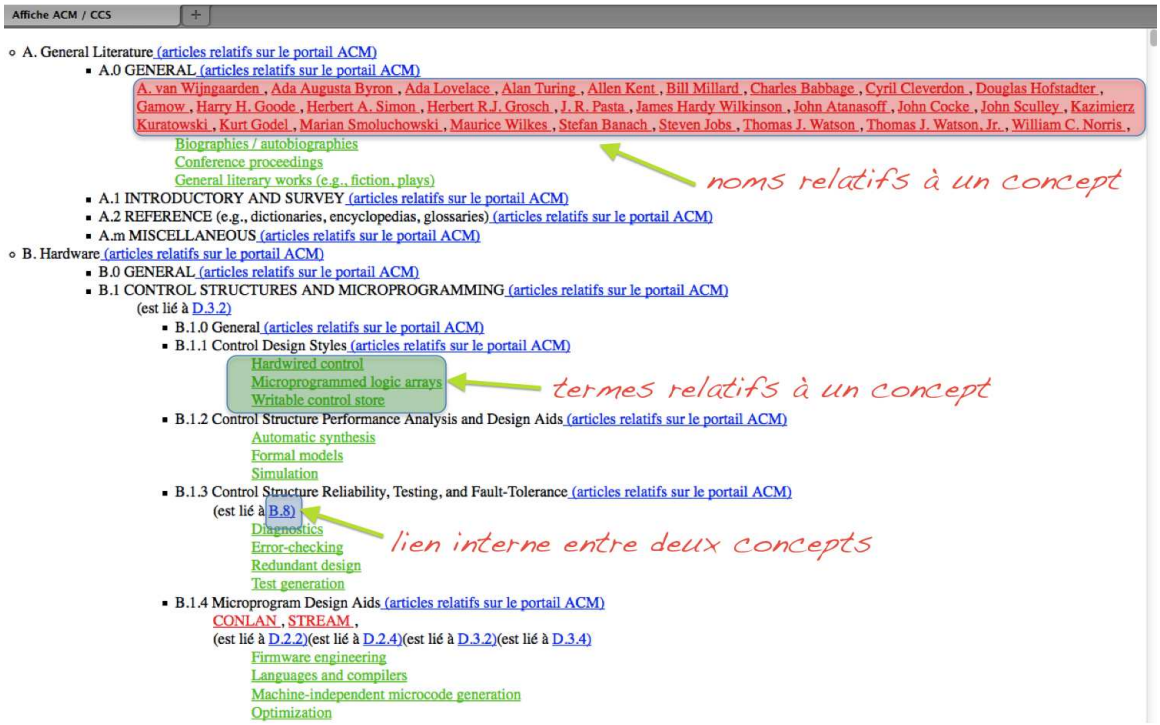


Figure 11.4: Extrait de l'affichage de la première approche

l'arborescence. Ces termes sont utilisés pour créer un hyperlien vers les documents du portail ACM qui combinent dans leur notice descriptive le concept père et le terme ;

– Rouge : Les noms qui précisent le concept supérieur, également mis en relation dans les mêmes conditions que les termes.

La capture d'écran proposée en figure 11.4 présente un extrait du résultat.

Interprétation du résultat

D'un point de vue structurel, l'accès à l'information se fait sur une base qualitative élevée. En effet, le thésaurus a été créé à partir de deux vocabulaires contrôlés et d'une taxonomie. Ces trois éléments sont dédiés au corpus, ce dernier étant parfaitement indexé par l'organisme ACM. L'affichage est clair, intuitif, mais ergonomiquement inutilisable en l'état. Si le thésaurus proposé est exhaustif, sa visualisation implique un affichage sur plusieurs hauteurs d'écran. Cette contrainte implique un usage intensif de la fonction de déroulement du navigateur (*scrolling*) avec la molette de la souris ou tout autre périphérique de pointage. Même si l'usage commun déconseille fortement cette pratique (Nielsen, 1990, 1994a,b, Nielsen et Molich, 1990) les études aussi bien

professionnelles que scientifiques démontrent que la lecture n'est pas réellement gênée par le phénomène de déroulement, dont l'usage est moins disruptif que le changement de page (Boucher, 2011, Lewenstein *et al.*, 2000). Cependant, par cette méthode d'accès linéaire de parcours du domaine d'information, le temps d'accès est excessif (presque une minute pour faire défiler l'ensemble du thésaurus). Nous soutenons que dans ces conditions, un utilisateur n'aura pas la patience de trouver le concept qui l'intéresse.

Il est probable que l'utilisation de blocs de données affichables à la demande permettrait de parcourir plus facilement l'arborescence. Nous pensons à utiliser des méthodes basées sur la manipulation de feuille de style pour ouvrir et fermer à la demande les branches et sous-branches de la taxonomie pour atteindre les termes recherchés [1].

C'est cette méthode qui est proposée sur le site d'ACM pour explorer la taxonomie ACM CCS. Les problèmes majeurs de cette solution sont :

- La perte de la vision globale du domaine.
- Un trop grand nombre d'actions sur le dispositif de pointage, ce qui allonge inutilement le temps d'accès à l'information.

Une présentation structurée des catégories facilite la sélection, mais elle ne dispense pas les utilisateurs de consulter fréquemment la représentation d'ensemble de l'hypertexte (Tricot, 1998). L'expérimentation met en évidence que, si les usagers ont le choix entre naviguer d'un nœud à l'autre par des liens directs (ou menus enchâssés) ou utiliser une table des matières, c'est cette dernière option qui est préférée (Britt *et al.*, 1996). Tricot à démontré la limite cognitive de ce type de représentation de connaissance. Cette limite est fortement liée au niveau de profondeur de l'arborescence, mais également au nombre de liens partant d'un nœud (Tricot, 1995). Or notre première représentation de l'arborescence est non seulement profonde, mais chaque nœud est abondamment hyperlié avec des liens de divers types. Nous comprenons maintenant le parti pris d'ACM de ne pas afficher dans une même page leur taxonomie et les vocabulaires contrôlés pour proposer un thésaurus exhaustif de l'informatique scientifique. Nous abandonnons donc cette méthode d'accès linéaire au domaine d'information. Néanmoins, nous retenons qu'il est nécessaire d'ajouter une dimension graphique à notre interface de recherche pour éventuellement simuler la manipulation d'un graphe. Nous posons

1. Ce dispositif technique implique l'utilisation, au moyen de fonctionnalités javascript, des attributs de visibilité (CSS) des blocs de données du langage HTML

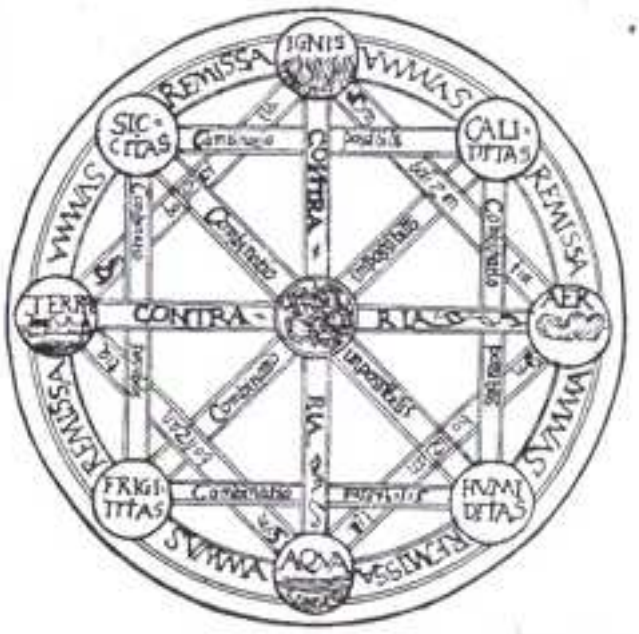


Figure 11.5: Graphe de visualisation de la connaissance universelle de Leibniz (1666)

comme pré-requis la possibilité d'afficher l'intégralité du thésaurus dans une seule page, sans augmenter le nombre de validations à l'aide de la souris ou de l'écran tactile.

11.5 Visualisation d'un domaine de connaissance

Dans son document précurseur de la visualisation moderne d'information *Conceptual Structures – Information processing in mind and machine*, Sowa propose une synthèse entre plusieurs écoles de pensée pour offrir un modèle de représentation de domaine de connaissance (Sowa, 1983). Dans l'analyse qu'il fait de ce document, Clancey propose de résumer la pensée initiale de Sowa de la manière suivante (Clancey, 1985) : Sowa réunit la vision aristotélicienne de la représentation et de la catégorisation de la connaissance (métaphysique) avec celle de Peirce sur la distinction entre concepts et objets – *Types and Tokens* (Wetzel, 2011). Il rajoute une dimension à la fois philosophique et mathématique sous forme de treillis sémantique des domaines de connaissance en ce basant sur le modèle conceptuel proposé et illustré (voir figure 11.5) par Leibniz (Leibniz, 1666) et en propose par la suite un modèle d'implémentation (Sowa, 2000).

Christophe Tricot, dans une approche cognitive de la carte et donc du graphe conceptuel, s'inspire de la méthode du cartographe Jacques Bertin dans l'ouvrage *Sémiologie graphique* (Bertin, 1967, Tricot, 2006). Il se positionne dans une approche cognitive et pragmatique en plaçant l'œil de l'usager comme centre du domaine de connaissance.

La recherche d'informations par visualisation se divise entre la présentation visuelle de l'information et la recherche visuelle d'informations.

11.5.1 La représentation

Selon Card, Mackinlay et Shneiderman (Card *et al.*, 1999) et Tricot (Tricot, 2006, Tricot et Roche, 2006), il y a trois paradigmes visuels de présentation de l'information que l'on nomme paradigmes cartographiques :

- Les paradigmes de représentation. Ils permettent de représenter la structure de l'information. Nous distinguons quatre types de structures d'information qui sont :
  1. la structure tabulaire (Rao et Card, 1994) ;
  2. la structure arborescente (Van Ham et Van Wijk, 2003) ;
  3. la structure par graphe (Cassidy *et al.*, 2006) ;
  4. la structure en spirale (Carlis et Konstan, 1998).
- Les paradigmes de la visualisation. Il s'agit des techniques permettant d'afficher la représentation des informations d'une manière claire et cohérente dans un espace physique limité. Un usager peut ainsi rapidement se faire une représentation de l'information présentée en embrassant du regard l'ensemble du domaine qu'il explore. Les techniques de visualisation sont classées en deux groupes :
  1. Les techniques de visualisation uniformes avec une vue d'ensemble et les détails à la demande (Jerding et Stasko, 1998).
  2. Les techniques de visualisation hétérogènes en focus avec contexte (Card *et al.*, 1999).
- Les paradigmes d'interaction. Ils concernent des techniques permettant aux utilisateurs d'interagir avec la visualisation :
  1. le zoom et le panoramique ;

2. le focus et le contexte – listes hiérarchiques zoomables (Card *et al.*, 1999) ;
3. le filtrage dynamique (Pediotakis et Hascoët-Zizi, 1996) ;
4. le zoom sémantique (Bartram *et al.*, 1995, Hascoët et Beaudouin-Lafon, 2001).

11.5.2 Visualiser pour chercher de l'information

Selon Zhang, il y a trois paradigmes de visualisation en recherche d'information (Zhang, 2008)

- Le paradigme QB (*Query searching and Browsing* / Requête et navigation). Initialement une requête est nécessaire pour limiter le jeu de résultats de recherche. Puis une visualisation de ces résultats est proposée pour que les utilisateurs puissent la parcourir et concentrer leur champ visuel sur des informations spécifiques.
- Le paradigme BQ (*Browsing and Query searching* / Navigation et requête). Une présentation visuelle de l'ensemble du domaine informationnel est préalablement établie pour la navigation. Ensuite, les utilisateurs interrogent le système et les résultats sont mis en évidence.
- Le paradigme BO (*Browsing Only* / Navigation seule). Ce paradigme n'intègre pas de composante de recherche, il s'agit uniquement de navigation.

Nous avons défini ce qu'était une taxonomie, mais la manière de représenter une telle masse de connaissances peut se révéler problématique. Dans son article « Taxonomie », dans le *Dictionnaire de sémantique*, Patrick Saint-Dizier (2006) pose la question suivante : « L'un des grands débats autour des taxonomies est de savoir si elles sont représentées par des arbres ou par des graphes. » Cette question trouve une réponse dans l'explicitation des deux composantes des cartes cognitives.

11.5.3 Cartes cognitives

Il y a deux types de cartes cognitives :

1. Les cartes en forme d'« arbre », plus rapides à établir, mais limitées. Il est pratique de les utiliser pour analyser et mémoriser un texte, clarifier un projet. On parle de carte heuristique, ou *mind map*.

2. les cartes en forme de « graphe » ; ce sont les « vraies » cartes conceptuelles, ou *concept maps*.

11.5.4 Cartes conceptuelles, ou concept maps

Une carte conceptuelle est une représentation graphique d'un champ du savoir, d'un ensemble de connaissances. Elle se compose de concepts et de liens entre concepts. Ils sont inclus dans des cellules de formes géométriques variables reliées par des lignes fléchées et étiquetées. Le(s) mot(s) ou le texte court associés à ces lignes expriment les relations sémantiques entre les concepts. La technique du *concept mapping* trouve son origine dans les recherches menées par Joseph D. Novak à l'université Cornell. Elle repose sur les théories constructivistes selon lesquelles l'image mentale de la réalité est une projection de l'interaction entre l'esprit humain et la réalité, et non le reflet exact de cette dernière. L'équipe de Novak cherchait alors à comprendre les mutations du savoir scientifique des enfants. Les chercheurs constatèrent à quel point il était difficile d'identifier les changements dans l'appréhension par les enfants des concepts scientifiques. Le programme de recherche était basé sur la psychologie cognitive de David Ausubel et la théorie de l'assimilation (Ausubel *et al.*, 1968). Son principe fondamental est que l'apprentissage se produit par assimilation de nouveaux concepts et propositions au sein de structures cognitives préexistantes (Novak et Cañas, 2008). La théorie de l'assimilation établit aussi une distinction essentielle entre apprentissage par cœur (*rote learning*) et apprentissage signifiant (*meaningful learning*). L'apprentissage signifiant requiert trois conditions :

1. Le matériau à apprendre doit être clair et présenté dans un langage et au moyen d'exemples adaptés aux connaissances antérieures de l'individu.
2. L'apprenant doit posséder un savoir préalable approprié.
3. L'apprenant doit choisir de s'impliquer, il doit être réellement motivé.

L'idée de Novak est de représenter sous forme de cartes les connaissances conceptuelles acquises par les enfants. Elles sont conçues pour encourager l'adoption du schéma d'apprentissage signifiant par les élèves en fournissant notamment un outil d'évaluation de leurs connaissances. Les cartes conceptuelles sont en effet très efficaces pour extérioriser le savoir des apprenants dans un domaine particulier (voir figure 11.6).

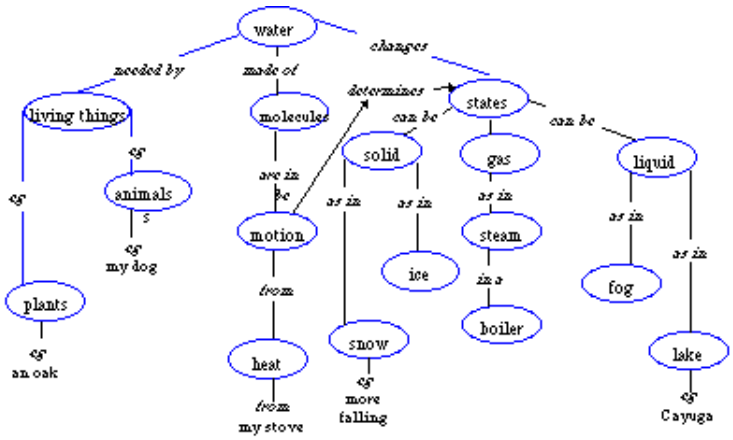


Figure 11.6: Visualisation en graphe de concepts, proposée par Roger *et al.* (2001) d'après Novak et Gowin (1984)

Principales caractéristiques

- les concepts sont représentés comme des nœuds (*nodes*) : le triplet nœud-lien-nœud forme une proposition signifiante, parfois appelée unité sémantique (*semantic unit* ou *unit of meaning*).
- structure semi-hiérarchique : représentation semi-hiérarchique des concepts, du général au spécifique. La lecture d'une carte conceptuelle se fait donc en général de haut en bas.
- représentation en réseau : des liens transversaux (*cross-links*) permettent de mettre en évidence les relations entre plusieurs domaines de la carte conceptuelle.
- contextualisation de la carte : elle cherche à répondre à une question particulière définie au préalable (*focus question*).

11.5.5 Synthèse des cartes cognitives

Pour rendre le corpus plus accessible, nous décidons de faciliter la représentation de l'ontologie de domaine sous la forme d'une carte navigable. L'arborescence doit permettre une mise en exergue, ou *focus*, de la branche contenant une formalisation du concept recherché.

Carte conceptuelle (Concept Map)	*Carte mentale (Mind Map)*
Représentation en graphe	Représentation en arbre
Cartographie de l'univers réel des concepts	Reflet personnel de la pensée
Liens étiquetés : emphase sur les connexions sémantiques entre les concepts	Relations non spécifiées entre les idées
Lecture du haut vers le bas	Lecture du centre vers l'extérieur
Niveau de complexité moyen à élevé	Niveau de complexité faible
Règles de représentation formelles et strictes	Règles moins formalisées, plus flexibles
Mémorisation difficile	Mémorisation plus facile
Compréhension aisée par d'autres personnes	Compréhension problématique : tendance idiosyncrasique (voir encadré 11.2).

Tableau 11.1: Mind Mapping vs Concept Mapping

En linguistique le susbstantif « idiosyncrasie » sert à marquer l'exception dans les idiomes institutionnels : par exemple l'anglais possède deux mots pour désigner le bœuf (animal = *ox* et viande = *beef*) est une idiosyncrasie (Akouaou, 1995, Pagés et Derghal, 1984).

Tableau 11.2: Idiosyncrasie

11.5.6 Pratiques comparées de visualisation de graphes

Il existe un certain nombre de manières de visualiser les ontologies, mais toutes ne sont pas propres à la navigation, en tous cas pas à une navigation intuitive. Dans notre contexte, l'outil de représentation doit se conformer à des règles exposées par Christophe Tricot et Christophe Roche suite à un certain nombre d'observations (Tricot et Roche, 2006). Au minimum, pour être efficace le système de visualisation doit respecter le code de bonnes pratiques suivant :

- Offrir une vue globale de l'ontologie. Cela permettra à l'utilisateur d'embrasser la structure du domaine ;
- Utiliser une approche « focus + contexte » pour permettre à l'utilisateur de se concentrer sur certains éléments tout en ayant accès aux autres.
- Utiliser la géométrie plane, pour éviter de nuire à la perception et à la manipulation.

Le dernier point en particulier semble difficiles à respecter, car considérant la masse de données à afficher et la volonté de respecter les points d'ergonomie précédents, il est complexe de combiner un affichage en arborescence et la géométrie euclidienne.

Du retour d'expérience de Christophe Tricot, nous noterons également qu'il émerge deux types d'utilisateurs : novices et experts. Les novices comprennent le domaine et ses concepts sans pour autant saisir la finesse de l'organisation et les interactions. Les experts quant à eux saisissent parfaitement la globalité du domaine tant du point de vue des concepts que des rapports qui les lient. Dans notre contexte d'utilisation, les utilisateurs ont un profil de connaissance qui peut être celui d'un étudiant en master ou d'un jeune enseignant chercheur qui se renseigne sur un sujet transversal à ses travaux. Il peut également s'agir d'un profil expert pour un spécialiste de domaine comme ce serait le cas pour un chercheur ou un documentaliste spécialisé dans un domaine. Nous essayerons donc de trouver un compromis de représentation du domaine offrant des accès directs au contexte sur l'élément en focus. Dans l'article de Tricot et Roche, il semble qu'un modèle de représentation par *radial tree* soit le plus approprié pour des experts et que l'*eye tree* soit indiqué pour des novices.

La visualisation en *eye tree* permet une vision globale du domaine ainsi que la possibilité d'un grand angle focalisé, « *fisheye polar* », sur un point de détail autour duquel s'articule le domaine. De plus, il s'accorde parfaitement avec la géométrie euclidienne. Le « *radial tree* » est assez similaire à l'« *eye tree* », combinant la vision

globale du domaine et le « *fisheye polar* ». Cependant, une plus grande place est faite au contexte et au focus au sein même du graphe. Il semble que ce qui fasse l'intérêt du « *radial tree* » (le focus + contexte) cause également une perte de contact avec l'objectif premier, qui est de conserver la vue globale. De plus, un *radial tree* décrivant l'ACM serait parfaitement illisible du fait même de la taille de l'ontologie.

Considérant la dimension de l'ontologie, une visualisation par grappe d'informations émerge au moyen de la combinaison d'ontologie et de la technologie dite « *Topic Mapper* ». Cet affichage est rendu possible par l'*applet* « *open source* » *Hypergraph*. Bien que n'étant pas spécialement préconisé pour représenter efficacement une ontologie, le *Topic Mapper* est une représentation de type « *hyperbolic tree* » qui consiste à cartographier l'ontologie afin d'y naviguer à volonté. Nous allons adapter ce procédé pour faire émerger des points de vues, mais aussi des mises en exergue, ou *focus*, avec leurs contextes. Il s'agira ainsi d'une approche hybride entre l'« *eye tree* » et l'« *hyperbolic tree* ».

11.6 Deuxième approche

Dans notre première approche, nous représentions une méthode de visualisation linéaire d'accès à l'information à travers une taxonomie enrichie pour devenir un thésaurus. L'accès au corpus scientifique d'ACM était assuré, mais la méthode de navigation n'était cependant pas cohérente avec la volonté de faciliter la démarche psycho-cognitive.

11.6.1 Objectifs

Après avoir décrit l'état de l'art des méthodes de visualisation de l'information, nous allons proposer, dans cette deuxième approche, une synthèse personnalisée des méthodes de présentation (Tricot, 2006, Tricot et Roche, 2006) et de recherche graphique (Zhang, 2008, Zhang et Marchionini, 2004, Zhang, 2002). De plus, nous allons également faire évoluer notre « structure de taxonomie-thésaurus » vers une ontologie de domaine au sens de Grüber (Grüber, 1993, 1995), ce que Monnin et Félix décrivent comme une ontologie formelle (Monnin et Félix, 2009).

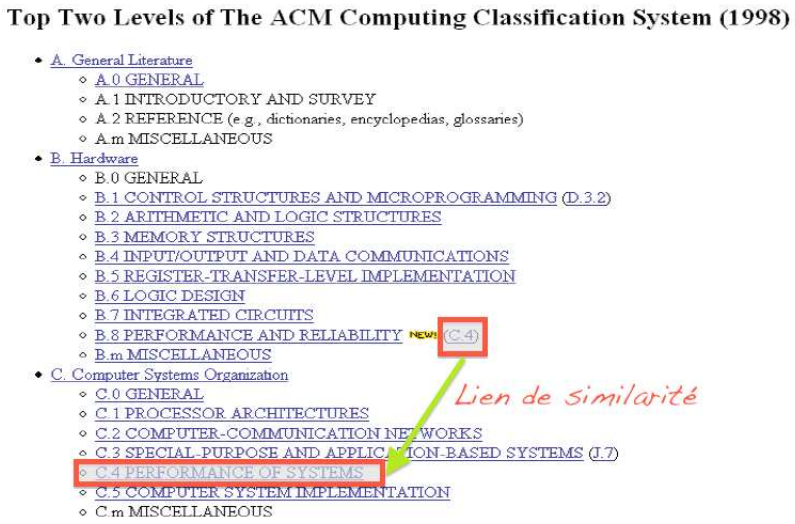


Figure 11.7: Lien transversal au sein de l'ACM CCS

11.6.2 Le modèle

Dans notre contexte, le *Computer Classification System* (*CCS*) n'est pas exploitable en l'état, ni structurellement, ni techniquement (ce que nous avons vu dans notre première approche). Le *CCS* semble a priori être plus une taxonomie qu'une ontologie. Dans une taxonomie, le vocabulaire est organisé sous une forme hiérarchique. Cette hiérarchisation correspond souvent à une spécification. Une taxonomie est une forme d'ontologie dont la grammaire n'a pas été formalisée. Dans le *CCS* cette grammaire a été réduite à des relations de généralisation/spécification et à des liens transverses de proximité sémantique, d'où le manque indubitable d'un thésaurus et de sa grammaire. Ainsi on peut noter un lien de proximité sémantique entre les nœuds B.8 « *Performance and reliability* » et C.4 « *Performance of systems* » (voir figure 11.7).

Si l'on revient sur l'apport du thésaurus comme transition entre taxonomie et ontologie, il est possible d'illustrer simplement cet aspect grâce à la figure 11.8. Cette illustration, tirée du thésaurus MotBis [1] propose quatre types de relations entre des termes et un concept :

TA : Terme(s) associé(s) ;

TG : Terme générique ;

1. Élaboré par le réseau SCÉRÉN et le CNDP, accédé le 1er septembre 2012 à l'URL : `http://www.cndp.fr/thesaurus-motbis/site/`.

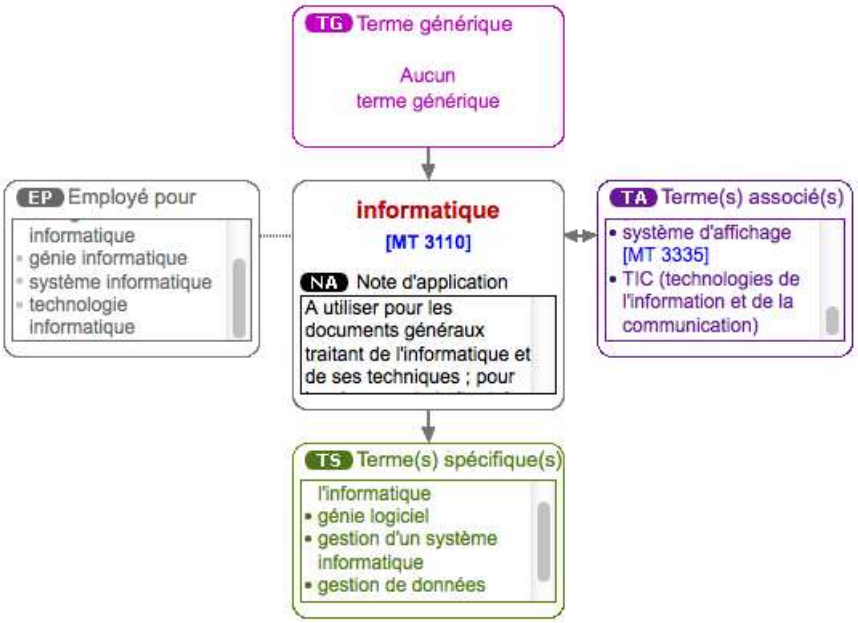


Figure 11.8: Exemples tirés du thésaurus MotBis

TS : Terme(s) spécifique(s) ;

EP : Employé pour.

Cependant, d'après Grüber, un des aspects importants d'une ontologie (en sus de la clarté, de la cohérence) est l'extensibilité (Grüber, 1993, 1995). Il convient donc d'effectuer un traitement pour permettre d'anticiper les évolutions de l'ontologie. En effet le système d'identifiant du CCS ne s'applique qu'aux nœuds conceptuels et non aux feuilles terminologiques. Cela empêche de conserver l'esprit de référencement si pratique proposé par l'*ACM* en cas de spécialisation d'une feuille. Nous ne pouvons en effet pas imaginer structurellement une relation de spécification pour un élément non référencé. C'est pourquoi nous choisissons de donner de manière arbitraire un identifiant aux feuilles pour les transformer en nœuds conceptuels potentiels. Pour distinguer nos évolutions du travail initial, nous avons choisi d'utiliser pour identifier les feuilles l'identifiant de la classe supérieure (nœud père) auquel s'ajoutera une lettre de l'alphabet. Notons que la notation CCS utilisant déjà le « m » pour « divers » (*miscellanous*) et le « g » pour « général » (*general*), nous avons ôté ces deux lettres de notre processus d'identification des nœuds et feuilles.

L'ontologie de domaine est composée d'une arborescence de sujets allant d'une racine générique : le domaine (ici l'informatique), vers des feuilles de connaissance. Les arcs seront des relations de spécification/généralisation, des liens de similarité. L'ontologie ne contient pas les articles, mais des mots clés dont l'héritage se fait de manière « *top-down* » (descendante) et qui permettent de générer une requête qui sera transmise à la bibliothèque scientifique en ligne d'ACM.

Cette arborescence constitue le squelette externe, ou exosquelette, du domaine. Nous reprendrons la terminologie de la première approche. Les nœuds de connaissance sont des concepts (par définition abstraits), les feuilles sont les termes issus du thésaurus. Dans l'optique d'une meilleure appropriation par l'usager, nous envisageons de le laisser ajouter ses « mots clés » pour décrire les concepts de son point de vue. Dans l'ontologie, les termes et noms originaux seront considérés comme des éléments « natifs », par opposition aux mots clés ajoutés *a posteriori*, qui seront dits « ajoutés ». Cette distinction aura une importance si nous exploitons ces apports avec un système de recherche, car il faudra signaler aux usagers qu'ils s'appuient sur un élément folksonomique.

11.6.3 L'implémentation technique

La section 11.5 consacrée à la visualisation d'un domaine de connaissance (et aux apports cognitifs qui en découlent) conclut sur un choix conceptuel et une approche technique associée. Nous avons opté pour une méthode hybride entre l'« *eye tree* » et l'« *hyperbolic tree* ». Ce modèle doit permettre une navigation fluide dans l'ontologie et offrir une option de zoom contextuel sur le concept visualisé. L'applet Java [1] Hypergraph [2] permet d'offrir une vision de l'ensemble de la structure du domaine et d'y naviguer pour trouver le concept intéressant l'usager. Il peut ensuite spécifier son intérêt grâce aux termes et au noms relatifs.

Pour qu'Hypergraph puisse afficher la taxonomie, celle-ci doit être compatible avec le format de description de graphes GraphXML (Angelaccio et Buttarazzi, 2002, Herman et Marshall, 2001).

Pour formater le contenu de la taxonomie ACM au format GraphXML, nous réutilisons le script initial en PHP de la première approche pour générer une base

1. Une applet est une application Java pouvant être intégrée dans une page au format HTML.
2. Proposé sous licence publique OpenSource GNU LGPL à l'url : `http://hypergraph.sourceforge.net`, accédé en ligne le 1er septembre 2012

Source de : file:///Users/admin_spi/Desktop/acm_ccs.xml

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <!DOCTYPE GraphXML SYSTEM "GraphXML.dtd">
3      <GraphXML xmlns:xlink="http://www.w3.org/1999/xlink">
4          <graph id="Ontologie Info">
5              <dataref>
6                  <node name='A.'>
7                      <ref xlink:href = "javascript:afficheNoeud('A.','en');" />
8                      <label>A.general literature</label>
9                  </node>
10             </dataref>
11             <edge source='A.' target='/' />
12             <dataref>
13                 <node name='A.0'>
14                     <ref xlink:href = "javascript:afficheNoeud('A.0','en');" />
15                     <label>A.0.general</label>
16                 </node>
17             </dataref>
18             <edge source='A.0' target='A.' />
19             <dataref>
20                 <node name='A.1'>
21                     <ref xlink:href = "javascript:afficheNoeud('A.1','en');" />
22                     <label>A.1.introductory and survey</label>
23                 </node>
24             </dataref>
25             <edge source='A.1' target='A.' />
26             <dataref>
27                 <node name='A.2'>
28                     <ref xlink:href = "javascript:afficheNoeud('A.2','en');" />
29                     <label>A.2.reference (e.g., dictionaries, encyclopedias, glossaries)</label>
30                 </node>
31             </dataref>
32             <edge source='A.2' target='A.' />
33             <dataref>
34                 <node name='A.m'>
35                     <ref xlink:href = "javascript:afficheNoeud('A.m','en');" />
36                     <label>A.m.miscellaneous</label>
37                 </node>
38             </dataref>
39             <edge source='A.m' target='A.' />
40         </graph>
41     </GraphXML>
```

Grammaire et espace de nommage

Fonction de focus et d'affichage du contexte

lien de l'hypergraphe

Figure 11.9: Branche A. de la Taxonomie ACM au format GraphXML simplifié

de connaissances au format MySQL. L'avantage d'automatiser cette action est de maintenir la taxonomie à jour à chaque évolution. Un script PHP permet d'extraire les informations en temps réel (même en cours de mise à jour) et de les afficher au format GraphXML. Nous pouvons noter sur la figure 11.9 que le code XML ne contient que l'arborescence de la taxonomie. Nous intégrons dans le code des appels à une fonction en langage javascript. Celle-ci prend en paramètre le nœud à mettre en focus (et donc le contexte à afficher) ainsi que la langue souhaitée pour l'affichage. Une simple page HTML peut alors mettre en relation la taxonomie et l'hypergraphe pour afficher la taxonomie. Nous ajoutons à notre interface un onglet de contexte pour intégrer le thésaurus à notre navigateur conceptuel (voir figure 11.10). Cet onglet permettra de proposer un contexte sémantique pour le concept mis en exergue par la navigation. Cela permettra éventuellement à l'usager d'affiner sa recherche, ou de l'ouvrir sur d'autres points de vue.

Un hyperlien permet de générer une requête au format openURL vers la bibliothèque ACM en utilisant l'identifiant du concept en focus et éventuellement les éléments de

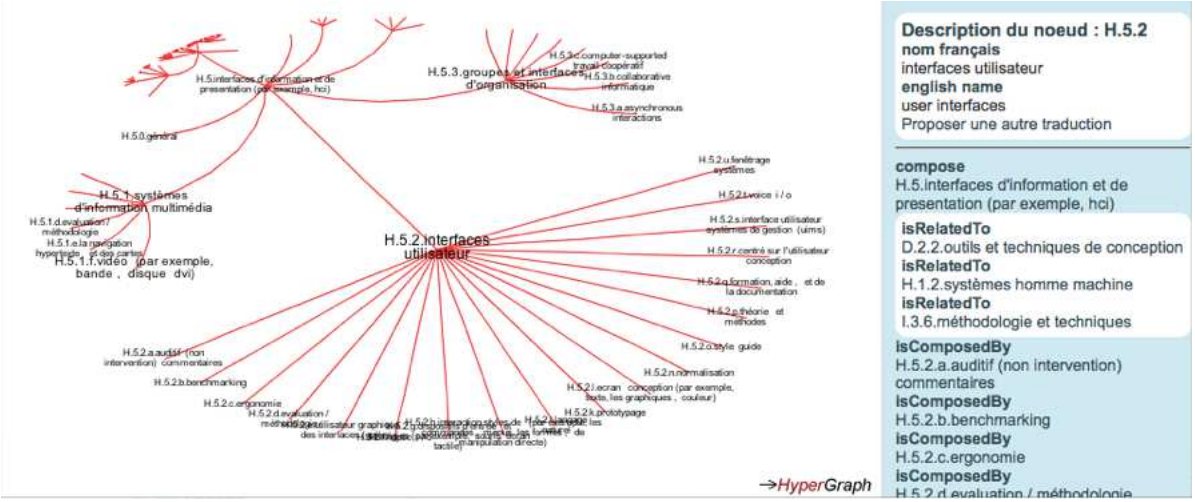


Figure 11.10: Focus + contexte dans notre interface

thésaurus. Une fenêtre s'ouvrira permettant la consultation des résultats sur le site distant.

11.6.4 Évaluation

Dans cette deuxième version, nous avons résolu le problème principal posé par la première version. Nous n'avons donc plus à utiliser un ascenseur interminable pour à accéder aux concepts, termes et noms qui nous intéressent. Nous avons également tenu compte des principes de C. Tricot pour la cartographie et la visualisation d'un domaine de connaissance (Tricot, 2006, Tricot et Roche, 2006). Nous avons repris le paradigme BQ (*Browsing and Query searching* / Navigation et requête) proposé par Zhang (2008). Les résultats des requêtes générées par la navigation répondent parfaitement aux requêtes. En effet, le corpus ACM est indexé grâce à la taxonomie et aux éléments de thésaurus que nous avons utilisés pour construire notre outil.

Cependant, de notre propre avis, il est difficile pour une personne ne maîtrisant pas le domaine informatique de trouver l'emplacement du concept recherché. Cela va à l'encontre de notre principe consistant à faciliter l'accès à l'information.

11.6.5 Conclusion

Cette solution est partiellement satisfaisante puisque, si elle aide à la formalisation des requêtes, elle n'est pas aisée à manipuler par un non-expert du domaine. Un

documentaliste ou un chercheur d'une autre discipline ne pourra pas l'utiliser sans s'être, au préalable, familiarisé avec la taxonomie ACM. De plus, si cette solution ne génère pas de bruit, sans visibilité sur le corpus, nous n'avons pas de moyen de juger du silence (c'est-à-dire des documents pertinents qui ne sont par retournés). Pour améliorer notre interface de recherche, nous proposons de trouver une solution qui permette à l'utilisateur de se positionner rapidement dans l'hypergraphe. Par ailleurs, afin de tester notre outil sur un autre corpus, non indexé, nous allons intégrer une base de connaissances publique.

11.7 Troisième approche

Cette troisième approche est plus exploratoire que les deux premières. Nous allons proposer d'aligner l'interface déjà proposée (focus + contexte) en navigation et requête avec corpus non indexé de notices bibliographiques.

11.7.1 Méthodes de recherche proposées et présomptions de modèles exploitables

Pour résoudre le problème posé par l'accès direct à l'information dans un graphe, il faut rejoindre le paradigme QB : *Query searching and Browsing* de Zhang (2008) (voir les paradigmes de Zhang page 283). Cette solution n'est que partiellement pertinente puisqu'elle annule les bienfaits cognitifs de la navigation. Nous allons donc proposer d'étendre le modèle de Zhang avec la proposition d'un paradigme de recherche supplémentaire :

QBQ-S (*Query, Browsing, Query-Searching*)

Cette heuristique de recherche se décompose en trois parties. Une simple requête (assistée par autocomplétion) dans un champ de formulaire va permettre de découvrir un contexte de recherche. Une fois un concept, un terme ou un nom sélectionné, un focus se crée autour pour permettre d'affiner sa recherche par navigation. Si le résultat conceptuel trouvé par l'utilisateur du système est exactement ce qu'il pensait trouver,

il lui reste donc à lancer la fonction de recherche de documents associés. Dans le cas contraire, il aura découvert des sujets connexes par sérendipité. Si le terme ou le nom qu'il cherchait n'est pas répertorié, libre à lui de l'ajouter. Cet ajout lui sera bénéfique, puisqu'il pourra régulièrement venir effectuer sa veille scientifique sur son domaine de prédilection avec un point d'entrée personnalisé. Les éléments ajoutés, bien que n'ayant pas subi de curation seront mis à disposition des autres usagers [1]. Une mention sera tout de même faite sur le caractère possiblement approximatif de la classification de l'élément ajouté. De plus, pour optimiser l'accès au graphe de connaissance, nous avons choisi de traduire les termes relatifs aux concepts et les lexiques associés en français. Détaillons le modèle fonctionnel de cette version, proposé en figure 11.11 (la numérotation des points suivants est reprise sur la figure).

1. et 1'. Possibilité de se positionner dans l'ontologie par navigation ou par une requête en langage naturel.
2. Le positionnement permet de cerner un point de vue utilisateur et des centres d'intérêt.
3. et 3'. ce qui va dégager des métadonnées et constituer des requêtes vers le RDF interne ou les bibliothèques numériques en ligne.
4. et 4'. Les intitulés des articles correspondant à la requête, et trouvés dans le RDF ou les bases de connaissances scientifiques, sont proposés.
5. Les articles sélectionnés sont cherchés sur l'Internet via *Google scholar* si l' « *Uniform Resource Identifier* » (URI [2]) est absente de la base interne. Si l'on utilise les bibliothèques numériques, l'accès aux documents est direct.

11.7.2 Traduction de l'ontologie en français

D'après la lettre ouverte à l'Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur (AERES) signées par quelques milliers de chercheurs français, il est largement admis que la « *lingua franca* » de la recherche scientifique est aujourd'hui l'anglais.

1. Il y aura inévitablement des inexactitudes dues à la malveillance, la méconnaissance ou des questions de points de vue. C'est pourquoi les propositions d'origines folksonomiques devraient idéalement être soumises à la curation de spécialistes de la catégorisation, à savoir des documentalistes (Earley, 2011).
2. Spécification W3C : `http://www.w3.org/2004/11/uri-iri-pressrelease.html.fr`

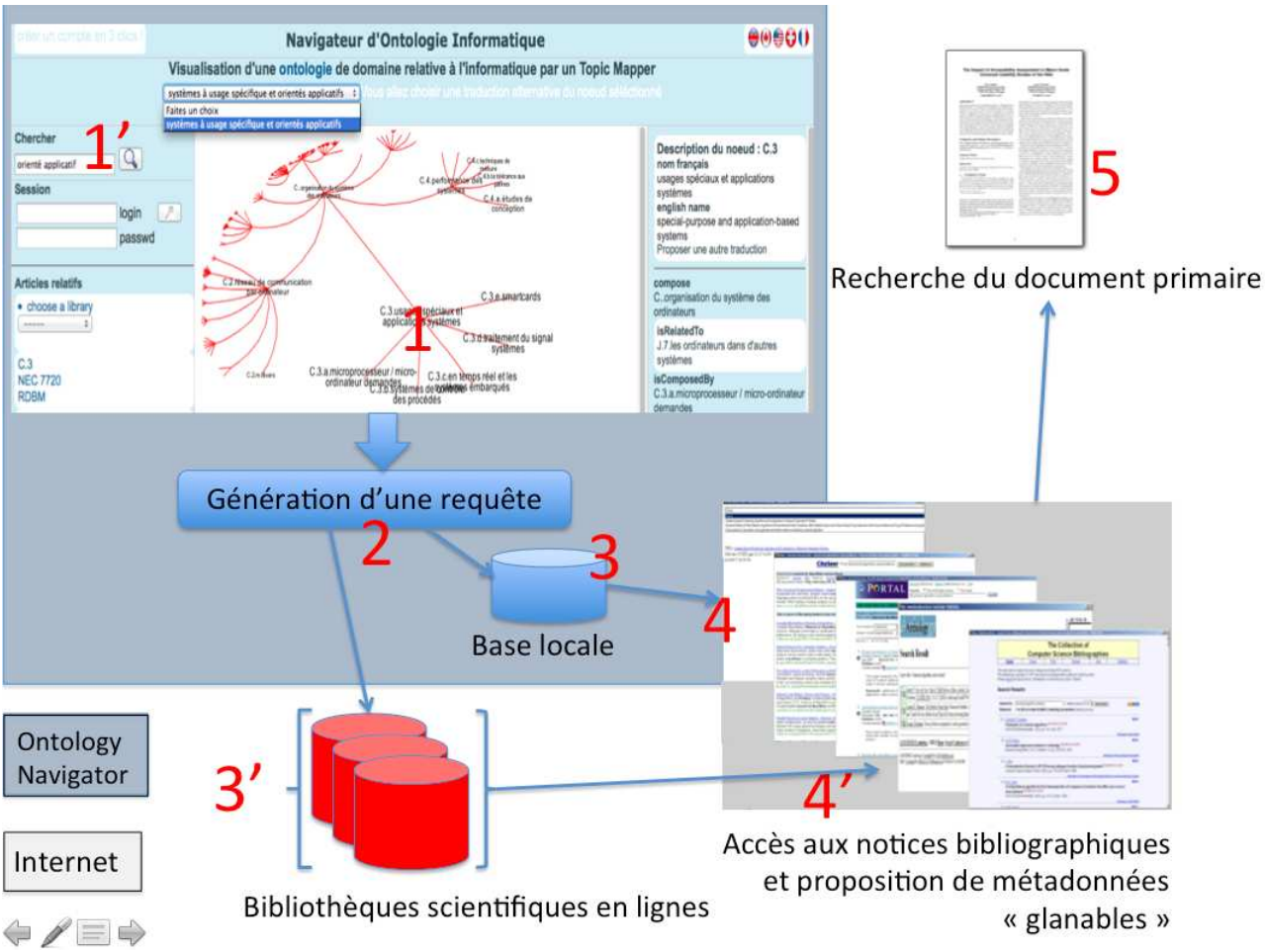


Figure 11.11: Modèle fonctionnel d'OntologyNavigator

Pourquoi traduire les intitulés des branches de l'ontologie en français alors que le corpus est majoritairement en anglais, la langue scientifique ? Nous attirons l'attention sur le fait que si l'utilisateur final maîtrise peut être la lecture de textes techniques et scientifiques, il peut se sentir plus à l'aise en français pour effectuer sa recherche, quitte à lire les articles ultérieurement en anglais avec un bon dictionnaire sous la main. Le choix le plus simple et le plus économique pour automatiser une traduction anglo-française est l'utilisation d'un outil de traduction en ligne (Kembellec, 2009). Les outils qui ont attiré notre attention sont Babelfish de Yahoo ! et Google Translate de la suite Google. Nous avons conçu et utilisé une Interface de Programmation Applicative (API) de « *wrapping* » pour générer une version française de l'ontologie basée sur un de ces outils.

Une fois cette étape terminée, nous avons rapidement compris que rien ne remplace une traduction manuelle, c'est pourquoi nous intégrons une notion de « *folksonomy* » pour l'aide à la curation sociale. Selon Thomas van der Wal, la valeur du marquage extérieur de la *folksonomy* vient des usagers. Ces derniers en utilisant leur propre vocabulaire, ajoutent une dimension d'inférence au concept par spécification.

L'aspect technique de cette démarche devra être simplifié au maximum pour l'utilisateur afin de ne pas le décourager de faire une proposition. L'opération ne doit également pas lui prendre plus de quelques secondes (cf. figure 11.12). L'intitulé français modifié ne sera remplacé qu'après une vérification manuelle, mais la nouvelle traduction sera intégrée dans l'ontologie comme vocabulaire complémentaire avec une relation d'équivalence. Nous ambitionnons ainsi de corriger la partie française de l'ontologie sur une période de temps encore indéterminée.

Nous élargissons ce principe folksonomique à l'ajout de mots clés, proposés contextuellement par les usagers pour spécifier un concept, de leur point de vue dans la langue de leur choix (identifié par le contexte [1]). Le procédé permet aussi de tenir compte des mutations terminologiques inhérentes aux évolutions du domaine *Information Technology* (IT). Ibekwe-SanJuan et SanJuan (2003) a démontré l'importance de la veille terminologique dans le domaine scientifique. L'utilisateur final bénéficie grâce à son interaction avec le système d'un enrichissement de sa connaissance du pôle de connaissances tout en participant à son évolution.

1. Notre outil intègre la gestion des sessions et donc des préférences utilisateur en matière de langue.

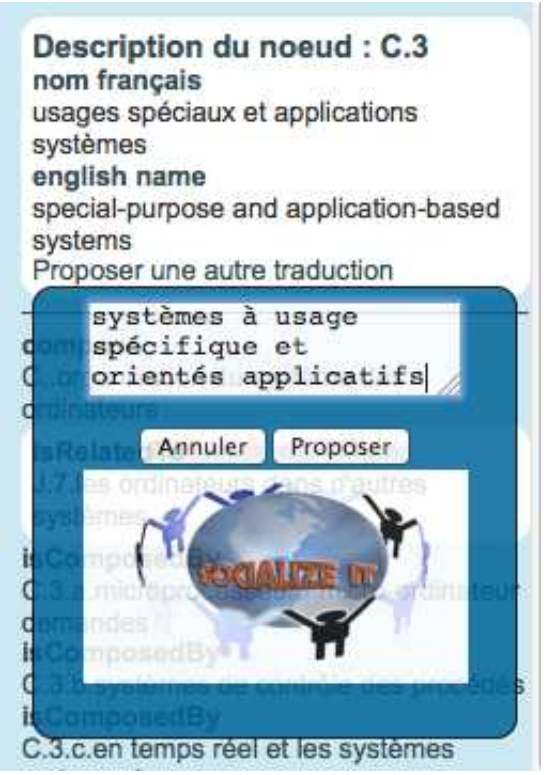


Figure 11.12: Exemple de curation folksonomique

11.7.3 Mise en œuvre du modèle QBQ-S

Comme nous l'avons vu, il existe un certain nombre de bases de connaissances scientifique. Malheureusement, peu exposent l'ensemble de leurs données au moissonnage massif. Cependant, un projet de qualité recense la plupart des publications scientifiques en informatique scientifique. Il s'agit du projet DBLP initié par Michael Ley de l'Université de Trèves (Trier) en Allemagne. Le suivi, le sérieux de ce travail ont fait connaître le projet sur la scène internationale. En 2003, M. Ley a reçu le prix de contribution ACM SIGMOD (groupe d'intérêts en gestion de données[1]) pour ses travaux.

Depuis, la base DBLP est une référence incontournable dans le monde de l'informatique scientifique. L'ACM a montré son adhésion en répliquant le site DBLP. Cette base propose un SRI avec moteur classique et une navigation avec recherche à facettes grâce au projet *Faceted DBLP*[2] de Diederich *et al.* (2007).

Le grand intérêt de cette base est que M. Ley propose l'intégralité des 1 631 850

1. `http://www.sigmod.org/sigmod-awards/award-people/michael-ley`
2. `http://dblp.l3s.de/`

notices (au 1er juillet 2012) dans un seul fichier XML. Ce projet, dont l'expérimentation est étalée sur deux décennies peut être considéré comme stable. Son succès lui a valu d'être plusieurs fois répliqué, et ce dans plusieurs contextes d'usage avec différentes IHM [1]. Le corpus n'est cependant ni annoté, ni indexé, ce qui nous laisse la possibilité de tester notre modèle.

Intégration du corpus DBLP

Nous commençons par choisir le corpus d'articles scientifiques liés à l'informatique scientifique. Le corpus de recherche sera composé des intitulés d'articles parus depuis 1945 et référencés dans la « *DataBase systems and Logic Programming (DBLP)* » par Michael Ley de l'Université allemande de Trier. Il s'agit à l'origine d'un document *XML* d'environ un million d'entrées au format BibTEX(format de description bibliographique de LaTEX). Notons que les articles sont rédigés dans diverses langues. Pour faciliter l'accès du corpus par concept, terme ou nom recherché dans notre outil, nous avons intégré cette base de connaissances sous forme de base de données MySQL. Ce processus informatique est détaillé dans l'annexe F.

Le point d'entrée facilité

Notre heuristique QBQ-S propose de faciliter l'entrée dans le graphe, pour les non-initiés au domaine de connaissance avec une saisie semi-automatisée. Nous avons implémenté cette idée grâce à la technologie JSON (*JavaScript Object Notation*). JSON est un format de données qui permet de représenter de l'information structurée. Créé par Douglas Crockford, le format JSON est décrit (et donc normalisé) par la RFC 4627 [2] de l'IETF (Crockford, 2006a,b). Le principe de cette technique est en mettre en relation des ensembles de paires nom / valeur dans un fichier texte en ligne. Ces listes ordonnées de valeurs permettent à une application distante, typiquement à un moteur de recherche, de corréler des portions de texte à l'intégralité d'un terme ou d'un nom [3].

1. Pour plus d'information sur ce projet, voir l'annexe F
2. Accessible en ligne à l'URL `http://tools.ietf.org/html/rfc4627`, accédé le 1er septembre 2012
3. Pour de plus amples explications conceptuelles et techniques, voir le site officiel de JSON : `http://www.json.org/`, accédé le 1er septembre 2012.

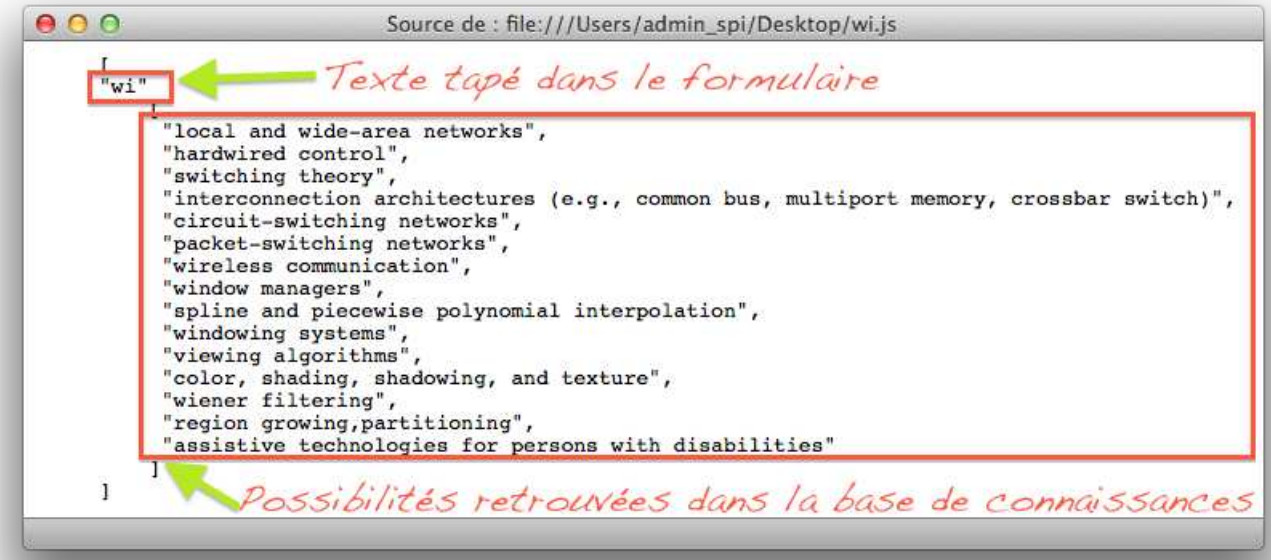


Figure 11.13: Exemple d'utilisation de JSON

Nous avons mis en relation les intitulés des termes et concepts, ainsi que les vocabulaires thésaurisés avec les identifiants de la taxonomie au moyen d'un script PHP qui produit un résultat au format JSON.

Cette ressource est accessible [1] par un champ formulaire de moteur de recherche et permet d'assister à la saisie de l'usager, s'il souhaite utiliser le module de recherche en plus du mode de navigation. À chaque caractère supplémentaire tapé, le champ de recherche rétrécit jusqu'à ce que soit trouvé l'élément recherché.

11.7.4 Explicitation d'usage

Cette interface se manipule de deux manières : par recherche assistée ou par navigation. Au-delà de la préférence d'un usager pour l'une ou l'autre méthode, nous avons conçu un accès pour les personnes à l'aise avec le système de classification et un autre pour aider les nouveaux utilisateurs. Examinons ces deux méthodes d'usage.

1. En ligne a l'URL : `http://www.geraldkembellec.fr/moteur/suggest.php?query=wi`, l'argument à passer est query. L'URL complète correspond donc à l'exemple proposé dans la figure 11.13, accédé le 1er septembre 2012.

Navigation pour les professionnels et initiés

Cette première approche est similaire au 2e modèle présenté précédemment, elle est directement inspirée du paradigme BQ de Zhang (2008). La première étape de la recherche par navigation consiste à descendre dans l'arborescence jusqu'au nœud le plus représentatif du concept recherché.

Ici, la démarche de navigation pour atteindre le nœud « Gestion de base de données » a été de commencer la navigation par la racine « / Informatique » puis la sélection de « H.Systèmes d'information » dans le graphe pour enfin choisir le nœud « H.2.Gestion de base de données », comme le montre la figure 11.14.

Accès facilité par champ de saisie pour les novices

Dans le cas d'une requête passée par le champ de saisie, le système d'autocomplétion n'est pas parfois directement exploitable car un traitement du langage naturel doit être préalablement effectué. Un processus traditionnel de recherche de traitement de la requête commence alors. En premier lieu, un filtrage du bruit est effectué sur la requête grâce à un dictionnaire de mots vides ou « *stop-lists* » (voir paragraphe sur le traitement automatisé de la langue 3.2.2, page 63). Cette première étape va éliminer les articles, pronoms, ainsi que les substantifs trop communs pour avoir un sens significatif et positionner l'utilisateur dans le navigateur de l'ontologie. La deuxième étape consiste à une racinisation des mots avec la méthode heuristique d'Enguehard (Enguehard, 1992). Les racines collectées sont alors comparées avec l'ensemble des mots dégagés avec la grappe de mots clés d'une des branches de l'ontologie, qu'il s'agisse de termes originaux, de mots clés proposés par les usagers ou de curation folksonomique sur la traduction. Les résultats les plus proches au sein de l'ontologie sont alors proposés sous forme de menu déroulant. Dans les cas de propositions qui ne sont pas issues de la taxonomie originale, un message informe du caractère alternatif de l'offre avec le message : « Vous allez choisir une traduction/un descripteur alternatif du nœud sélectionné (voir le haut de la figure 11.14) ».

11.7.5 Exemple de recherche contextuelle d'articles

Le bloc rouge du contexte propose un accès direct aux articles des bibliothèques numériques en ligne comme CSBIB, DBLP, ou ACM en générant des requêtes contextuelles

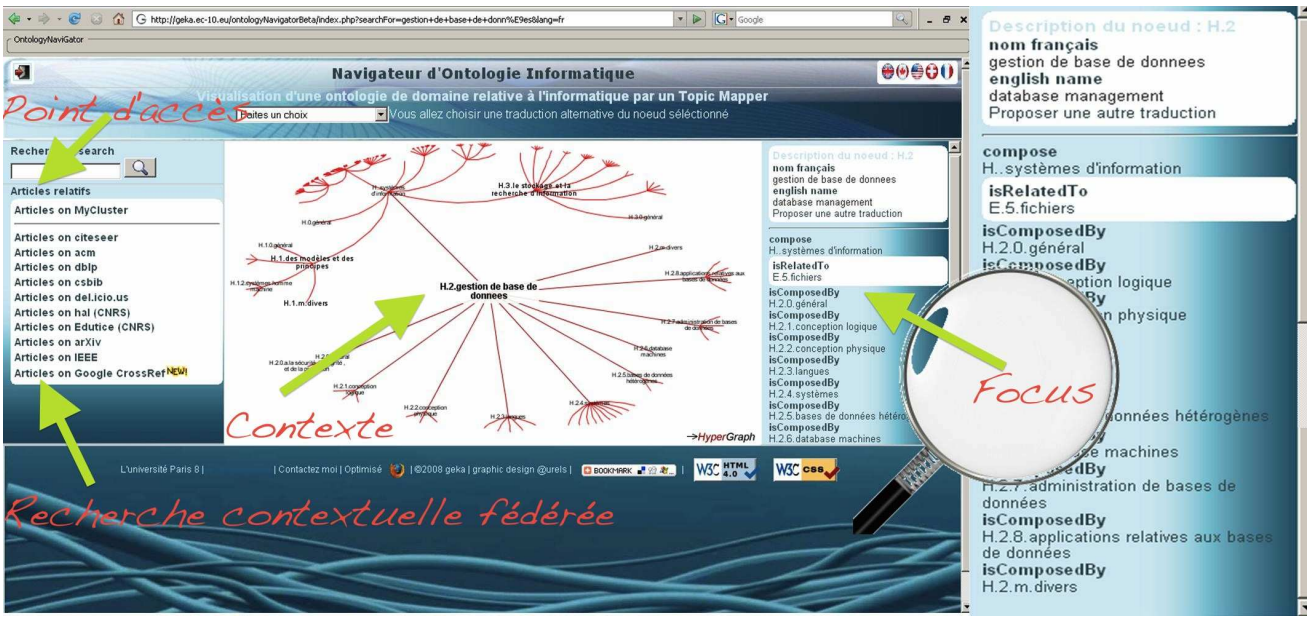


Figure 11.14: Exemple d'utilisation de cette version d'OntologyNavigator

vers ces sites. Mais l'outil propose également d'interroger la base interne d'intitulés d'articles. Dans l'exemple présenté sur la figure 11.15, page 303, une recherche sur « gestion de base de données » est générée et propose quelques dizaines de résultats. Nous choisissons le document « *Managing document taxonomies in relational databases* », la base nous donne l'auteur principal. L'outil vérifie la présence d'une *URI* relative à l'article dans la base, et en l'absence de celle-ci une requête au format openURL est proposée vers Google Scholar sous forme d'hyperlien. Ce lien offre un accès direct à la fiche de l'article. Les métadonnées bibliographiques sont exposées au moissonnage contextuel (pour Mendeley, Zotero ou tout autre LGRB compatible). De plus, un lien d'accès vers le document primaire est également proposé, ce qui complète la chaîne d'automatisation du processus de recherche documentaire. Cependant, l'accès au document primaire n'est pas toujours offert par la base DBLP ou par Google Scholar. Dans ce cas, nous avons prévu d'exposer le contenu de la notice bibliographique grâce à une bibliothèque (au sens logiciel du terme) qui charge la notice depuis la base et l'intègre à la page d'affichage du résultat sous les formes COinS et « Dublin-Core intégré aux métadonnées HTML » (voir le chapitre 9, page 223 et suivantes sur l'urbanisation de systèmes d'information).

Si l'usager ne trouve pas ce qu'il cherche par autocomplétion dans le formulaire de recherche, il reste la solution de lancer tout de même la requête : il est tout à fait

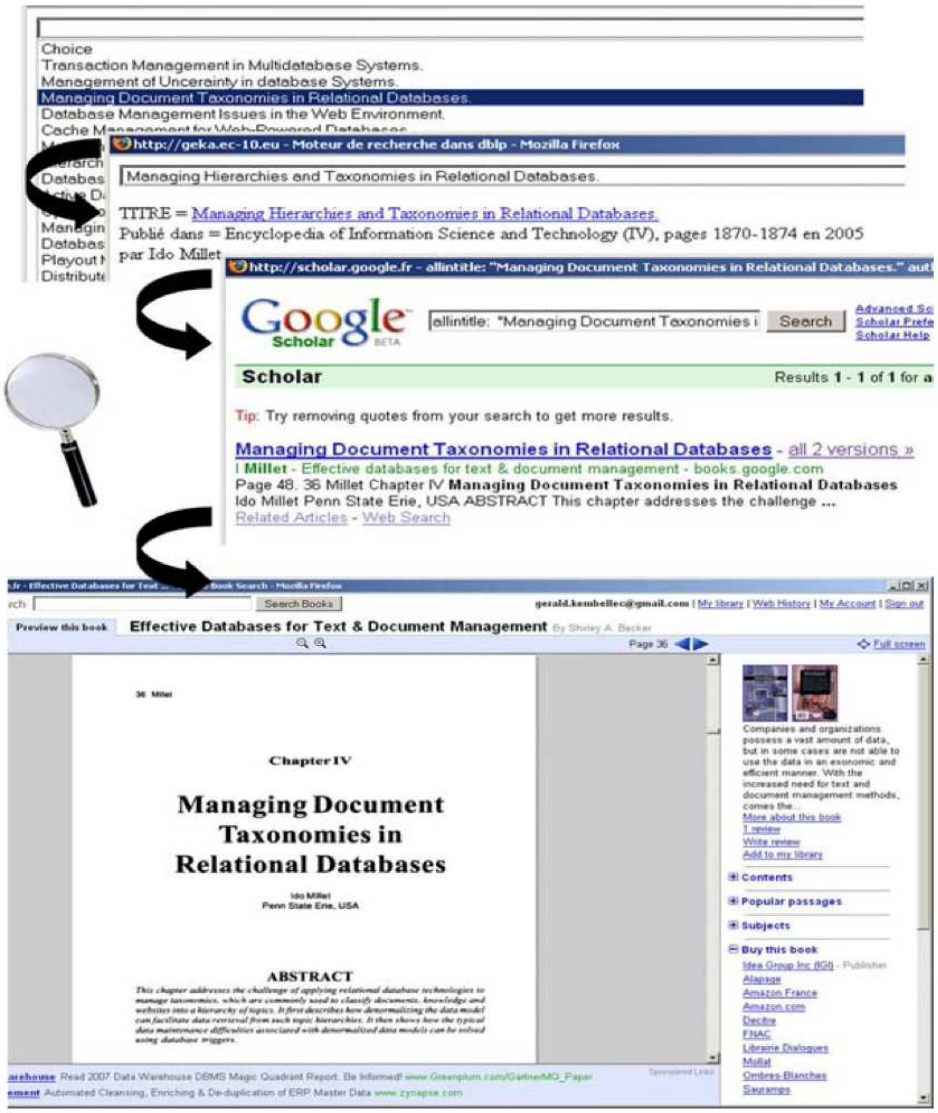


Figure 11.15: Processus de sélection et d'accès au document

possible qu'un autre utilisateur ait ajouté l'élément cherché (cf. figure 11.12, page 298).

11.7.6 Évaluation du modèle

Étude théorique d'OntologyNavigator

Dans un premier temps, avons évalué, avec des collègues issus des sciences informatiques et des sciences humaines, la pertinence des notices bibliographiques retournées par notre outil depuis la base interne. Le groupe d'étude était composé, en plus de l'auteur, d'un enseignant chercheur à la double compétence en sciences de l'information et de la communication et de l'informatique ainsi que d'une chercheuse en informatique. Les résultats furent présentés à la conférence IEEE Multimedia Computing and Systems par Kembellec, Saleh, et Sauvaget Kembellec *et al.* Ces résultats sont résumés dans le tableau 11.3.

Avec 71 % de résultats retournés en rapport avec le sujet choisi, les premiers résultats étaient encourageants, mais ils ont fait apparaître les limites de l'usage d'un corpus non indexé.

Identifiant	Valeur textuelle anglaise	Valeur textuelle française	Pertinence
H.3	information storage and retrieval	stockage et recherche d'information	100 %
G.3	probability and statistics	probabilités et statistiques	90 %
J.6	computer-aided engineering	ingénierie assistée par ordinateur	80 %
F.1	computation by abstract devices	calcul sur système virtuel	100 %
F.2.1	numerical algorithms and problems	algorithmes numériques et problèmes	90 %
B.5	register-transfer-level implementation	mise en œuvre d'un niveau de registre de transfert	70 %
B.5.2	design aids	aide à la modélisation	20 %
B.3	memory structures	structures de mémoire	100 %
I.2.7	natural language processing	traitement du langage naturel	100 %
C.5	computer system implementation	implantation de systèmes informatiques	80 %
D.3.2	language classifications	classification des langages	60 %
E.1	data structures	structures des données	100 %
K.2	history of computing	histoire de l'informatique	20 %
A.2	reference (e.g., dictionaries, encyclopedias, glossaries)	référence (par exemple, dictionnaires, encyclopédies et glossaires)	0 %
I.3.3	picture/image generation	génération d'images et de photos	60 %
Moyenne			71 %

Tableau 11.3: Résultat de l'évaluation de l'outil (Kembellec *et al.*, 2009)

Évaluation d'OntologyNavigator : étude de terrain

Dans un deuxième temps, pour valider l'adéquation de notre modèle avec les besoins du terrain, nous avons confronté OntologyNavigator à une population d'utilisateurs ciblés sur le domaine de connaissance que nous avions choisi. L'outil est donc mis à disposition des élèves ingénieurs en informatique de l'EiCnam [1] dont douze ont participé à l'évaluation sur la base du bénévolat. L'étude a eu lieu dans le prolongement d'un travail dirigé de trois heures sur la recherche d'information scientifique et technique du domaine informatique [2]. Les principales informations fournies par l'étude sont : Seuls 17 % des membres du groupe connaissaient la taxonomie ACM avant le cours.

Un tiers des élèves ingénieurs pensent qu'une telle classification aide réellement l'utilisateur final à classer et à retrouver de l'information, 47 % n'en sont pas sûrs et 27 % n'ont pas d'avis tranché. Pour connaître la première impression de ces usagers face à la taxonomie ACM sur le portail du même nom, nous leur avons posé la question ouverte (et optionnelle) suivante : « Quel est votre état d'esprit face à la taille et la complexité de la classification initiale ACM ? ». Voici les réponses des participants :

- *Les thèmes transverses, comme la sécurité, la gestion des identités sont difficiles à trouver. Il faut se familiariser avec pour l'appréhender.*
- *Cette classification peut être utile dans une bibliothèque physique.*
- *Cette classification est complète et bien faite.*
- *Je suis intéressé par la classification malgré un premier abord austère et très complexe. Je pense que cela m'aidera à trouver des informations dans mon domaine.*
- *Je suis quelque peu dubitatif... et il y a du boulot pour tout analyser.*
- *Je suis confiant pour obtenir des informations pertinentes, mais un peu dérouté par son utilisation.*

Avant de procéder à l'essai de notre outil (dans ses première et dernière versions), nous avons demandé aux participants quel mode de représentation de la taxonomie ACM leur semblait le plus approprié en se basant uniquement sur des captures d'écran. Les résultats sont présentés dans le tableau 11.4. Il est intéressant de constater que, *a priori*, ce n'est ni le site original d'ACM avec 25 %, ni OntologyNavigator avec seulement 18 %

1. École d'ingénieurs du CNAM, voir l'url : `http://ecole-ingenieur.cnam.fr/`, accédé en ligne le 1er septembre 2012.
2. Vous trouverez le questionnaire en annexe D.

Mode de présentation	choix *a priori*
Le mode site ACM simple	0 %
Le mode ACM avancé	25 %
Le mode HTML avec coloration (premier modèle)	50 %
L'hypergraphe, focus et contexte (troisième modèle)	17 %
Ne se prononcent pas	8 %

Tableau 11.4: Choix *a priori* d'un modèle de représentation / accès

Mode de présentation de la taxonomie	choix *a posteriori*
Le formulaire ACM	0 %
Le formulaire avancé de l'ACM CCS	17 %
Le mode HTML avec coloration (premier modèle)	0 %
L'hypergraphe, focus et contexte (troisième modèle)	58 %
Ne se prononcent pas	25 %

Tableau 11.5: Choix *a posteriori* d'un modèle de présentation / accès.

qui remportent le plus de suffrages mais notre premier modèle HTML. Nous expliquons ce choix très simplement par la simplicité de visualisation en arbre et l'identification par couleurs des termes et noms relatifs aux concepts du domaine (voir la capture d'écran en figure 11.4, page 279). Cette question a été reposée après le test des différents modèles, les réponses sont présentées dans le tableau 11.5. La première remarque sur ces résultats porte sur l'évolution de l'opinion sur notre premier modèle. Après utilisation, plus personne ne le choisit comme modèle favori de navigation/recherche. La forte contrainte liée au temps d'accès à l'information du à l'usage de l'ascenseur *scrolling* a eu raison de l'attrait provoqué par la simplicité du système.

Nous constatons également que notre troisième modèle est préféré par une large majorité du groupe étudié (58 %). Cependant, ce bon résultat est nuancé par le fait que 25 % des membres du groupe ne se prononcent pas.

Pour clore cette étude nous avons essayé d'évaluer l'opinion du groupe sur OntologyNavigator en terme de qualité des résultats retournés. Nous avons donc posé la question suivante : « Après avoir effectué une recherche sur OntologyNavigator, évaluez l'adéquation des résultats proposés avec votre besoin d'information de 1 (pas du tout

Note donnée	Ventilation des notes
1-2	0 %
3-4	25 %
5-6	0 %
7-8	50 %
9-10	25 %

Tableau 11.6: Adéquation des résultats proposés avec le besoin d'information.

en adéquation) à 10 (parfaitement en adéquation) ». Ce que nous notons en premier est que les trois quarts des usagers ont évalué l'adéquation entre les résultats et leur besoin par une note supérieure ou égale à 7/10. Nous observons également qu'un quart du groupe trouve le résultat en inadéquation avec leur besoin informationnel avec une note comprise entre 3 et 4.

Ces résultats, bien que satisfaisants, viennent corroborer ceux du groupe d'experts proposés dans la partie 11.7.6 (page 304) et publiés précédemment (Kembellec *et al.*, 2009) sur la qualité des résultats retournés par OntologyNavigator.

Voici les remarques libres retranscrites *in extenso* :

- *« C'est organisé comme les mind mappers, ça permet de suivre les idées, de naviguer plus facilement. »*
- *« L'interface graphique sous forme d'arbre est très lourde visuellement. »*
- *« Plus fluide et synthétique*[1]*. »*
- *« L'hypergraphe apporte une alternative intéressante et plus interactive que les nuages de mots par ex. Cependant, d'un point de vue ergonomique de la page en général, je trouve que ce serait bien de rendre le site plus attrayant. D'un point de vue plus technique, mon choix ne se serait pas porté sur une applet Java, car trop lourd. De plus, le rechargement incessant de l'applet finit par être lassant. Peut-être (il me semble avoir vu ça quelque part) qu'une approche HTML5 avec support 3D ou openGL, voired même graphML serait peut-être à explorer. Mais projet prometteur. »*

1. La personne interrogée fait ici la comparaison entre le premier modèle et OntologyNavigator.

Conclusion sur les études

Ces deux évaluations de l'outil, l'une portant uniquement sur la qualité des résultats (évaluation expert), l'autre sur la navigabilité et les résultats mettent en exergue les points suivants :

1. La qualité de l'information retrouvée est satisfaisante mais perfectible.
2. L'interface d'OntologyNavigator est appréciée, mais, jugée d'un abord complexe, elle ne fait pas consensus.

Nous avons noté de la part des élèves ingénieurs en informatique une piste technique intéressante avec la proposition de mise en page HTML 5, éventuellement en 3D.

Réflexions sur l'évolution du corpus DBLP

Cette réflexion sur notre approche du contenu de DBLP aurait pu s'intituler *DBLP-some lessons learned*, du nom de l'article concluant les travaux de Ley sur son corpus documentaire (Ley, 2009b). Pour citer l'auteur : « *DBLP is a (very imperfect authority) file for computer science researchers [...] Incomplete and inconsistent information, imperfect software, lack of time, and our own inability are the limiting constraints for this task*[1] ». Ce travail est remarquable, la base de connaissances est effectivement considérée dans le milieu de l'informatique scientifique comme une référence. Cependant la surabondance d'informations pose le problème de la qualité des sources. D'une source unique et fiable, en l'occurrence les DVD de données fournis par l'ACM (Ley, 2009a), DBLP est passé à un moissonnage systématique des grosses bases de connaissances, qui elles-mêmes moissonnent l'information d'autres entrepôts.

Nous avons suivi, par exemple, des documents[2] issus d'entrepôts locaux, eux-mêmes moissonnés par HAL. Cet entrepôt libre est ensuite moissonné par ArXiV, dont les notices seront recueillies et traitées par DBLP. Nous avons constaté une importante « perte de signal » entre l'intégration du document primaire et des métadonnées associées par l'auteur ou l'éditeur sur l'entrepôt de données et les fiches stockées dans DBLP.

1. Proposition de traduction : « DBLP est un fichier d'autorité très imparfait à destination des chercheurs en sciences informatiques [...]. Des informations incomplètes et incohérentes, les logiciels imparfaits, le manque de temps, et nos propres limites sont les contraintes qui brident l'accomplissement de cette tâche. ».
2. Les références bibliographiques que nous mentionnons sont extraites de notre bibliographie.

Nous assistons dans ce cadre, de manière sporadique, à des altérations significatives des notices bibliographiques entre l'émetteur initial (l'auteur) et le récepteur (le scientifique en recherche d'information via DBLP). Ce phénomène s'apparente à ce que l'on appelle trivialement, si l'on nous permet l'expression, le jeu du « téléphone arabe » (*Chinese whispers* en anglais). Ce jeu consiste à répéter une phrase dans une chaîne humaine pour vérifier que le message émis en début de chaîne est bien celui reçu à l'autre bout. Évidement, plus la chaîne est longue, plus la phrase initiale est altérée, parfois jusqu'à l'absurde. Cette problématique a été étudiée et modélisée dans le cadre du langage naturel afin de démontrer que la qualité de l'information décroît avec la quantité d'intermédiaires entre émetteur et récepteurs (Blackmore, 2000, p. 59). Notre expérience semble aussi montrer que le concept s'applique aux opérations de communication électronique. Cette altération du signal informationnel intervient en dépit des normalisations pour les échanges, qui sont précisément censées régler ce problème. Pour améliorer notre outil, nous avons alors eu le choix d'utiliser notre modèle original pour fédérer des requêtes vers des portails de connaissances pré-indexées, ou d'intégrer un corpus pré-indexé.

11.7.7 Corpus ACM ré-indexé et enrichissement ontologique

Nous avons noté lors de notre étude du portail ACM que la présentation d'une notice bibliographique offrait un large panel de métadonnées. Nous avons contacté dans un premier temps le portail ACM pour demander une version de l'index de leur corpus, car ce dernier était parfaitement compatible avec l'ontologie que nous avons modélisée. Sans réponse de l'organisme, nous avons créé un outil de moissonnage massif des données publiques en langage PHP-DOM [1]. Le principe de cet ensemble de scripts consiste, dans une démarche de traitement par lots, à extraire les notices bibliographiques complètes, avec les métadonnées, dans une démarche incrémentale depuis les pages de présentation du portail ACM. Le résultat est formaté et indexé dans un fichier XML de 14 giga-octets (soit plus de dix fois le volume de DBLP, qui n'est ni indexé, ni annoté et ne contient pas les résumés). Ce document de description de données est parfaitement aligné avec la taxonomie ACM, ce qui permet, de plus, d'enrichir l'ontologie du domaine. La figure 11.16 illustre l'apport à l'ontologie des éléments de la collection. Chaque document

1. Un script PHP manipulant des objets.

Source de : file:///Users/admin_spi/Desktop/acm-ccs.xml

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<collection>
    <paper id='1721833'>
    <title>Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks</title>
    <date>03/01/2010</date>
    <authors>
        <author>Kostakos, Vassilis</author>
        <author>O'Neill, Eamonn</author>
        <author>Penn, Alan</author>
        <author>Roussos, George</author>
        <author>Papadongonas, Dikaios</author>
    </authors>
    <issn>1073-0516</issn>
    <journal>ACM Transactions on Computer-Human Interaction (TOCHI)</journal>
    <pages>2</pages>
    <doi>10.1145/1721831.1721833</doi>
    <publisher>ACM</publisher>
    <keywords>
        <keyword>Bluetooth</keyword><keyword>Pervasive</keyword><keyword>brief encounter</keyword>
        <keyword>complex network</keyword><keyword>copresence</keyword><keyword>encounter</keyword>
        <keyword>epidemic</keyword><keyword>information diffusion</keyword><keyword>mobile interaction</keyword>
        <keyword>mobility</keyword><keyword>sensing</keyword><keyword>social network</keyword><keyword>trail</keyword>
        <keyword>ubiquitous</keyword><keyword>urban computing</keyword><keyword>virus</keyword><keyword>visualisation</keyword>
    </keywords>
    <abstract>
        Moving human-computer interaction off the desktop and into our cities requires new approaches to understanding people and technologies
        in the built environment.
        We approach the city as a system, with human, physical and digital components and behaviours.
        In creating effective and usable urban pervasive computing systems, we need to take into account the patterns of movement
        and encounter amongst people, locations, and mobile and fixed devices in the city.
        Advances in mobile and wireless communications have enabled us to detect and record the presence and movement of devices through cities.
        This article makes a number of methodological and empirical contributions.
        We present a toolkit of algorithms and visualization techniques that we have developed to model and make sense of spatial
        and temporal patterns of mobility, presence, and encounter.
        Applying this toolkit, we provide an analysis of urban Bluetooth data based on a longitudinal dataset containing millions of records
        associated with more than 70000 unique devices in the city of Bath, UK.
        Through a novel application of established complex network analysis techniques, we demonstrate a significant finding on the relationship
        between temporal factors and network structure.
        Finally, we suggest how our understanding and exploitation of these data may begin to inform the design and use of urban pervasive
        systems.
    </abstract>
    <ccsnodes>
        <ccsnode>G.4</ccsnode>
        <ccsnode>H.5.2</ccsnode>
    </ccsnodes>
</paper>
</collection>
```

Enrichissement sémantique de l'ontologie

Position dans l'ontologie

Figure 11.16: Corpus bibliographique indexé et annoté (extrait)

indexé offre d'enrichir l'ontologie, soit de mots clés (ceux choisis par l'auteur), soit de liens entre les termes et noms des vocabulaires contrôlés originaux.

Du point de vue technico-sémantique, cette démarche était parfaitement viable avec pour résultat un corpus parfaitement aligné avec notre modèle de visualisation du domaine de connaissance. Malheureusement, ACM ne nous a pas accordé le droit d'utilisation, y compris dans un cadre de recherche, des données publiques que nous avions récoltées. Face à cette logique financière, tout à faire compréhensible, au vu de la valeur d'un tel index, nous avons donc dû choisir entre indexer nous-même un autre corpus et explorer la piste de la recherche fédérée.

À ce stade de nos recherches, trois ans s'étaient écoulés depuis la présentation initiale de notre modèle de navigation par graphe d'une ontologie de domaine pour l'accès à l'information scientifique contenue dans le corpus DBLP (Kembellec, 2009). Lorsque nous avons décidé, dans la continuité directe de nos travaux, d'indexer un corpus de connaissance scientifique en informatique pour le relier au graphe que nous avons modélisé à partir de la taxonomie ACM, l'idée venait d'être implémentée par

Kboubi, Chaibi, et BenAhmed (Chabi *et al.*, 2011, Kboubi *et al.*, 2011, 2012). Nous avons donc choisi de finaliser la modélisation de notre outil avec un accès fédéré aux bases de connaissances scientifiques. Cette option se concrétise par une liste déroulante de propositions d'accès aux bases de connaissances scientifiques au moyen de requêtes au format OpenURL.

Idéalement, nous aurions aimé proposer une moisson des différentes sources d'informations documentaires. Cela nous aurait permis une présentation incluant l'affichage à facettes et une assistance accrue à la décision. Cette approche n'est applicable actuellement que sur peu de sources. Citons par exemple Isidore et DBLP, qui sont accessibles en SPARQL. Isidore ajoute l'accès oai-pmh à ses méthodes d'accès. La librairie du Congrès, l'OCLC et la BnF sont moissonnables grâce à des requêtes au format SRU-W.

11.8 Limites et perspectives de notre modèle

Nous l'avons constaté lors de l'essai de l'outil, les résultats proposés en réponse à l'usage actuel des méta-requêtes générées par la navigation sont parfois approximatifs. Il devrait cependant apparaître que, plus l'ontologie s'étoffera de données, plus la recherche et la navigation contextualisée seront précises.

Une autre limite est l'accès physique aux articles sur les bases externes qui est souvent soumis à un abonnement payant, quand ce n'est pas un payement à l'unité. C'est pourquoi cette solution trouvera plus facilement une place dans les locaux d'un laboratoire universitaire ou une bibliothèque. Cependant, l'utilisation d'un *proxy*[1] devrait permettre d'étendre l'accès aux abonnements des bibliothèques numériques à tout un campus.

Dans un avenir proche nous envisageons d'étendre l'ontologie aux acteurs scientifiques grâce au format « *friend of a friend* » (FOAF). Cela devrait permettre de mieux cerner les groupes de travail, équipes, et laboratoires ainsi que les liens de transversalité disciplinaire (Brickley et Miller, 2007). Éventuellement, le système de navigation de type « *hyperbolic tree /eye tree* » sera délaissé si un autre type de visualisation plus navigable ou ergonomique émerge.

1. Un serveur mandataire (*proxy*) est un serveur informatique dont le rôle est de servir de relais entre un client et un serveur Web. Ce service peut être associé à l'utilisation d'un identifiant, ou être transparent.

11.9 Conclusion

Selon Dinet, il existe un certain nombre règles à connaître pour l'usabilité d'une interface de recherche (Dinet, 2009). Nous avons tenté de réaliser une interface capable de répondre aux besoins de l'usager sous deux axes :

1. Un premier axe d'ergonomie de l'accès à l'information ;
2. Un deuxième axe d'inter-médiation technique pour l'interopérabilité avec les outils choisis par l'usager ;

Au cours de ce travail, nous avons réalisé un outil de recherche pour les chercheurs dont les travaux sont liés à l'informatique. Cette interface consultable en ligne permet de lier un contexte de recherche ontologique à des bibliothèques scientifiques en ligne. Cette ontologie basée sur la CCS de l'ACM a été traduite en français de manière automatique pour proposer aux chercheurs francophones un outil en langue maternelle. Notre solution propose aux chercheurs de trouver des articles relatifs à un contexte d'étude gravitant autour d'un concept de l'ontologie du domaine informatique. Cette requête est générée par navigation graphique du domaine ou par le langage naturel. Une fois le contexte de recherche dégagé, un travail automatisé permet de trouver des articles en relation dans la base de données interne ou de proposer des méta-requêtes vers les bibliothèques numériques scientifiques en ligne.

11.10 Généralisation du modèle de recherche

Dans la partie précédente, nous avons créé un système d'information capable de fournir un accès facilité à l'information scientifique dans le domaine de l'informatique. Nous souhaitons élargir le principe à un modèle plus générique, utilisable dans n'importe quel domaine de connaissance, en intégrant tous les principes de bonnes pratiques exposés tout au long de cette thèse. La construction de notre modèle passe par cinq points fondamentaux :

1. Le respect des besoins cognitifs des usagers.
2. La navigation dans une représentation du domaine avec contextualisation des métadonnées.

3. La possibilité d'enrichir l'ontologie.
4. La fédération des ressources documentaires de qualité.
5. Une interopérabilité des données fournies.

Ce modèle théorique se traduit concrètement lors de la création d'un portail de recherche d'information scientifique de la manière suivante :

1. Le choix d'une taxonomie du domaine de connaissance.
2. L'enrichissement constant par thésaurisation et folksonomie vers une ontologie de domaine.
3. L'implémentation de la structure au sein d'une base de données (MySQL, RDF, XML) pour un accès souple et rapide.
4. Le choix d'une possibilité de navigation et de recherche graphique dans le domaine.
5. L'accompagnement de l'usager dans sa recherche avec un système d'aide à la décision.
6. L'offre d'une méthode d'interrogation fédérée des principales bases de connaissances du domaine.
7. L'exposition de métadonnées pour l'interaction avec les logiciels de gestion de références bibliographiques.

Cette proposition de bonne pratique doit permettre de modéliser un système de recherche d'information scientifique et technique d'une utilisation souple pour l'usager. La considération de l'ensemble du processus garantit au minimum le respect des capacités cognitives de l'usager, particulièrement du non spécialiste du domaine de connaissances.

Chapitre 12

Conclusion générale

Dans les pages qui précèdent, nous avons présenté l'information scientifique dans son contexte et les canaux de diffusion qui y sont associés. Après un bref rappel des méthodes de recherche d'information scientifique traditionnelles, nous avons présenté quelques approches de recherche avancée, notamment en termes ergonomiques et qualitatifs avec la réduction du bruit et du silence lors de l'affichage des résultats. Cette démarche nous a amenés à passer en revue les techniques d'aide à la décision liées au processus de sélection de l'information.

Grâce aux études présentées, nous avons constaté un recul dans l'usage des bibliothèques universitaires traditionnelles et de leurs versions électroniques dans le contexte scientifique. Après avoir étudié la littérature à propos des méthodes classiques des processus documentaires, nous avons réfléchi aux aspects psycho-cognitifs qui leur sont liés. Nous avons tenté d'expliquer la baisse de fréquentation des OPAC par une inadéquation entre leur offre et les usages des étudiants et des chercheurs ?

Pour étayer cette idée, nous avons organisé une enquête auprès des enseignants, chercheurs et documentalistes universitaires. Cette étude exploratoire a permis de préciser le contexte d'usage des bibliographies scientifiques, que cela soit dans le cadre d'un cursus d'enseignement ou de recherche.

De plus, cette étude a offert une réponse à la première hypothèse secondaire : les populations cibles de la recherche documentaire scientifique (étudiants, enseignants-chercheurs et documentalistes) sont effectivement prêtes pour l'usage de l'automatisation du processus documentaire.

Nous avons ensuite précisé les éléments techniques, relatifs à la création des bibliographies scientifiques, abordés dans l'enquête. Dans un premier temps, les modèles et styles bibliographiques ont été présentés et commentés. L'enjeu connexe aux bibliographies et à la citation de références est le risque de plagiat par défaut de mention à l'auteur. Dans un deuxième temps, nous avons présenté les solutions logicielles susceptibles de s'intégrer dans un processus d'automatisation de la gestion bibliographique.

Notre contribution

À travers de l'étude des outils de gestion de bibliographie et des méthodes d'accès à l'information scientifique pour les étudiants, enseignants-chercheurs et documentalistes, nous avons exposé les paradigmes de glanage et de moissonnage d'information ainsi que les interactions techniques entre systèmes informatiques dans le web de données.

Nous avons démontré, comme pressenti dans notre hypothèse principale, que l'automatisation des phases techniques de la création de bibliographies scientifiques est actuellement déjà possible. Cela représente un atout précieux en termes de gain de temps et d'allègement de la charge psycho-cognitive pour les chercheurs soumis à de fortes contraintes de délais.

Il est donc d'ores et déjà possible d'automatiser le processus complet de documentation scientifique grâce à des outils d'intermédiation dédiés à des domaines scientifiques spécifiques et par l'urbanisation des différents systèmes d'information. Ce processus commence par la compréhension du domaine, la recherche documentaire assistée, l'aide à la sélection et se poursuit par l'intégration des sources bibliographiques dans un outil de gestion documentaire, ainsi que l'intégration de références aux documents dans le traitement de texte, pour s'achever par la génération de la bibliographie dans le format souhaité. Toutes les actions techniques du processus de documentation peuvent donc être effectuées par l'utilisateur sans qu'il ait à recopier et formater les références aux documents qu'il a sélectionnés. Pour les tâches de compréhension du domaine de connaissances et l'aide à la décision, nous rendons le processus aussi souple que possible par une visualisation adaptée au moyen d'une vue globale du domaine concerné sous forme d'une carte conceptuelle.

Notre étude exploratoire sur les usages et pratiques de recherche et de gestion de la documentation scientifique en France a permis de répondre à quelques interrogations :

Premièrement, les pratiques bibliographiques évoluent vers une automatisation progressivement adoptée par les utilisateurs. Deuxièmement, des utilisateurs commencent à être prêts pour cette évolution des usages, mais de manière sensiblement différente selon le profil professionnel (scientifique ou documentaliste), le domaine de prédilection et l'expérience. Ainsi, pour répondre à la première hypothèse secondaire, les étudiants, jeunes chercheurs, enseignants chercheurs et documentalistes sont intéressés par le type d'outil que nous présentons, sous réserve de formation. Il existe donc une niche écologique favorable à notre modèle et à son implémentation. Nous nuancerons cependant ce propos en rappelant que les usages des technologies présentées restent émergents.

Nous avons la conviction qu'il doit y avoir un élargissement des services proposés par les interfaces de documentation, notamment les OPAC. Ces derniers gèrent l'accès aux références bibliographiques depuis une base de données, mais doivent également proposer des services pour en faciliter l'utilisation. Nous pensons qu'une intermédiation entre les interfaces documentaires et les outils de gestion bibliographiques utilisés par les utilisateurs doit être systématique. C'est ce que nous présentons sous le terme d'*urbanisation de systèmes d'information documentaires*. Cette interopérabilité entre la base de connaissances et l'usager serait un grand pas vers la résolution de son besoin d'automatisation des tâches techniques dans le cadre de la recherche documentaire. L'utilisateur serait alors dégagé des tâches répétitives et aurait plus de temps à consacrer à la lecture et à l'évaluation des documents nécessaires à sa production scientifique.

En réponse à la deuxième hypothèse secondaire, nous avons modélisé et mis en place une solution logicielle adaptée graphiquement pour répondre aux besoins psychocognitifs, mais aussi techniques des populations étudiées. Enfin, nous avons décliné cette solution sous forme d'un modèle abstrait afin qu'il puisse être implémenté dans d'autres domaines scientifiques.

Enfin, pour ce qui est de la troisième hypothèse, nous n'avons pas pu vérifier finement l'impact cognitif de notre modèle sur les usagers. Nous devons, pour obtenir des résultats mesurables de manière effective, étudier l'impact à court, moyen et long terme de l'usage de notre modèle sur une population donnée.

Discussion sur l'accès aux données scientifiques

Les discussions autour du Web de données, de l'OpenData et du BigData sont au cœur de l'actualité et forment un sujet intrinsèque de recherche en documentologie. Nous reprenons le propos d Vajou *et al.* (2009) relatif à ce sujet : « l'un des principaux défis qui se pose alors aux responsables de la recherche publique est d'organiser ces bibliothèques [...] pour les rendre pérennes, directement accessibles et (ré)exploitables ». En corollaire à ce paradigme d'ouverture et d'universalité des données, l'interopérabilité des données dans un système d'information globalisé vient immédiatement à l'esprit [1] avec les groupes de discussion autour du FRBR [2]. Il semble que des efforts soient effectués par les communautés pour aller en ce sens. L'édition numérique scientifique devra composer avec cette dynamique. Le concept d'interopérabilité des données, introduit en 2003 dans la stratégie de soutien financier de la *National Science Foundation* des États-Unis vise à mettre en place un espace de communication scientifique sans barrière, reliant données, publications et autres matériaux utiles à la recherche (Laakso *et al.*, 2011, Van de Sompel et Lagoze, 2007).

Toutefois, l'édition scientifique des résultats de la recherche et leur accès sont fortement liés à une logique financière, celle des « cyberinfrastructures » privées de la recherche, comme les éditeurs scientifiques (Chartron, 2007). La publication scientifique est également soutenue par les initiatives libres financées par la recherche institutionnelle sous forme de bases de dépôt HAL ou ArXiV et de revues libres (PloS).

Le droit d'accès payant aux données issues de la recherche, financée par des fonds majoritairement publics, ou l'intégration du modèle « auteur-payeur (Chartron, 2007) », sont autant de polémiques actuelles qui ont lieu dans un cadre économique complexe. Il est en effet discutable qu'un auteur doive payer une somme forfaitaire « sans rapport avec le prix de revient [3] » à un éditeur à « forte notoriété » pour avoir accès à la version

1. Le rapport final du groupe d'incubation du W3C « Bibliothèques et web de données » accessible à l'adresse : `http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025`, accédé le 1er octobre 2012.

2. Le rapport de l'IFLA est accessible en français à l'adresse suivante : `http://www.ifla.org/files/cataloguing/frbr/frbr-fr.pdf` pour la version originale de 2001 et à l'url `http://www.ifla.org/files/cataloguing/frbr/frbr-fr_2012.pdf` pour la version amendée en 2012, accédés le 1er octobre 2012.

3. Les termes entre guillemets sont tirés de l'article de Vajou et Al. reprenant l'étude menée par *Welcome Trust* (Trust, 2003, Vajou *et al.*, 2009).

« officielle » de son propre travail sur un portail. Cela est d'autant plus vrai, s'il a dû financer son travail de recherche pour partie et que son laboratoire a payé des frais relatifs à une conférence, comme le déplacement et les droits d'inscription. Ces pratiques sont en mutation (Björk *et al.*, 2010, Laakso *et al.*, 2011), mais pour qu'un modèle en libre accès, scientifiquement reconnu et économiquement viable se généralise, un financement en amont sera nécessaire, éventuellement encadré par les organismes de recherche pour pouvoir prendre en considération des préoccupations éthiques (par exemple pour la recherche dans les domaines du médical, de l'agro-alimentaire, de l'énergie ou encore de la chimie).

Perspectives

Nous avons constaté le 22 septembre 2012 que l'ACM a révolutionné sa méthode d'affichage de la taxonomie informatique. Depuis cette date, une représentation graphique tabulaire (Rao et Card, 1994) propose l'accès au corpus par un système de spécification conceptuelle (voir en Annexe E.1 page 405). La nouvelle interface offre en outre la possibilité d'afficher les documents les plus récents du concept visité. Si cette nouvelle interface ne respecte ni la règle des trois clics d'accès à l'information, ni la visualisation contextuelle du concept visité, elle n'en est pas moins, de notre point de vue, réellement efficace pour un spécialiste du domaine. Les usagers apprécieront certainement l'effort entrepris pour faciliter l'accès à la connaissance, ce qui va dans le sens du modèle que nous présentons.

La suite de nos recherches s'orientera autour de trois axes principaux : la mesure effective des gains cognitifs liés à l'utilisation d'OntologyNavigator, l'amélioration de l'ontologie avec un système de curation folksonomique et enfin un système d'assistance à la décision lors de la sélection en contexte des documents scientifiques. Si nous parvenons, dans le cadre d'un projet encadré institutionnellement, à obtenir un partenariat avec un éditeur scientifique, nous pensons pouvoir offrir une méthode bénéfique sur le plan cognitif, avec une dimension possiblement pédagogique, mais surtout réduisant de manière significative la surcharge informationnelle.

Bibliographie

ACM (1964). The 1964 Computing Reviews Classification System - *Obsolete* (consulté le 01/07/12). http://www.acm.org/about/class/cr64, (consulté le 01/07/12). 263

ACM (1998a). The 1998 ACM Computing Classification System. http://www.acm.org/about/class/1998 (consulté le 01/07/12). 252

ACM (1998b). How to classify works using ACM Computing Classification System. http://www.acm.org/about/class/how-to-use (consulté le 01/07/12). 266

ADBS (2012). Vocabulaire de la documentation - l'association des professionnels de l'information et de la documentation. http://www.adbs.fr/vocabulaire-de-la-documentation-41820.htm, (consulté le 01/07/12). 56, 57, 114, 253

Adomavicius, G. et Tuzhilin, A. (2005). Towards the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 734–749. 95, 97, 98

Agosto, D., Rozaklis, L., MacDonald, C. et Abels, E. (2011). A model of the reference and information service process. *Reference & User Services Quarterly*, **50**, 235–244. 84

Akouaou, A. (1995). Derivation et idiosyncrasie : Le point de vue pédagogique. In *La langue franccaise au Maghreb.*, 7–18. 286

Alpert, J. et Hajaj, N. (2008). We knew the web was big... `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`, (consulté le 01/07/12). 36

American Psychological Association (2010). *Publication manual of the American Psychological Association*. American Psychological Association, Washington DC, 6th edn. 205

Andreu, R., Ricart, J., Valor, J. et Centre, N.C. (1992). *Information Systems Strategic Planning : A Source of Competitive Advantage*. NCC Blackwell. 223

Angelaccio, M. et Buttarazzi, B. (2002). A GraphXML description of query maps. In *Information Visualisation, 2002. Proceedings. Sixth International Conference on*, 339–343. 291

Antelman, K. (2004). Do open-access articles have a greater research impact ? *College & research libraries*, **65**, 372. 48

Ausubel, D., Novak, J., Hanesian, H. *et al.* (1968). Educational psychology : A cognitive view. 284

Bach, P.L. (2010). *Redesigning academic library websites in small, medium and large scale institutions : reasons and solutions*. Master's thesis, Oslo University College, Oslo, Norway ; University of Parma, Parma, Italy and Tallinn University, Tallinn, Estonia, Oslo. 227

Bachimont, B. (2007). *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents.*, vol. 67 of *Science informatique et SHS*. Lavoisier. 270

Baeza-Yates, R. et Ribeiro-Neto, B. (2008). *Modern Information Retrieval*. Addison-Wesley Publishing Company. 135

Balabanović, M. et Shoham, Y. (1997). Fab : content-based, collaborative recommendation. *Communications of the ACM*, **40**, 66–72. 95

Bartlett, J. et Toms, E. (2004). Validating qualitative research : Determining the generalizability of qualitative findings. In *ASIST 2004 Annual Meeting ; "Managing and Enhancing Information : Cultures and Conflicts"*, Providence, Rhode Island, USA. 111

Bartram, L., Ho, A., Dill, J. et Henigman, F. (1995). The continuous zoom : A constrained fisheye technique for viewing and navigating large information spaces. In *Proceedings of the 8th annual ACM symposium on User interface and software technology*, 207–215. 283

Bates, M. (1993). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, **13**, 407–424. 57, 128

Baudoin, L., Haeffner-Cavaillon, N., Pinhas, N., Mouchet, S. et Kordon, C. (2004). Indicateurs bibliométriques. *médecine/sciences*, **20**, 909–915. 29

Baudrillard, J. (1973). *Le miroir de la production : ou, l'illusion critique du matérialisme historique*, vol. 27. Casterman. 114

Beel, B., Joeran ; Gipp (2010). Academic search engine spam and google scholar's resilience against it. *Journal of Electronic Publishing*, **13**. 40

Belkin, N. et Marchetti, P. (1989). Determining the functionality features of an intelligent interface to an information retrieval system. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, 151–177, ACM. 72, 74

Benali, K., Rieu, D. et Soule-Dupuy, C. (2009). *Documents annotés et langages d'indexation*, vol. 12 of *Document numérique-RSTI*. Hermes-Lavoisier, Paris, hermes-lavoisier edn. 225

Bérard, R. et Gibert, J. (2008). Le sudoc dans google scholar. *Bulletin des Bibibliothèques de France*, 64–66. 40

Berizzi, L. et Zweifel, C. (2005). Le pingouin bibliothécaire : les logiciels libres de gestion de bibliothèque. *Revue Electronique Suisse de Science de l'information*, **35**, –. 193

Berners-Lee, T. (1989). Information management : A proposal. 244

Berners-Lee, T. et Hendler, J. (2001). Scientific publishing on the semantic web. *Nature*, **410**, 1023–1024. 225, 235

Berners-Lee, T., Bizer, C. et Heath, T. (2009). Linked data. *International Journal on Semantic Web and Information Systems*, **5**, 33. 225, 235

Bertin, J. (1967). *Sémiologie graphique*. Mouton/Gauthier-Villars, Paris, France. 282

Bitouzé, D. (2012). Cours latex. Tech. rep., Université du Littoral Côte d'Opale, `http://gte.univlittoral.fr/members/dbitouze/pub/latex`, (consulté le 01/07/12). 153

Bitouzé, D. et Charpentier, J. (2010). *LaTEX, l'essentiel.*. Pearson Education Paris. 153

Björk, B.C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. et Guðnason, G. (2010). Open access to the scientific journal literature : Situation 2009. *PLoS ONE*, **5**. 319

Blackmore, S. (2000). *The Meme Machine*. Oxford Paperbacks, new ed edn. 310

Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I. et Volker, J. (2007). Ontology-based question answering for digital libraries. *Research and Advanced Technology for Digital Libraries*, 14–25. 271

Boole, G. (1854). *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*. Original Edition Macmillan. 67

Borlund, P. (2003). The concept of relevance in ir. *Journal of the American Society for information Science and Technology*, **54**, 913–925. 102

Bornmann, L. et Daniel, H. (2005). Does the h-index for ranking of scientists really work ? *Scientometrics*, **65**, 391–392. 31

Bosc, H. (2005). *Les Archives Ouvertes : enjeux et pratiques. Guide à l'usage des professionnels de l'information*, 27–54. C. Aubry and J. Janik, Paris, France, adbs edn. 48

Boubée, N., Tricot, A., Couzinet, V. et Equipe, M. (2005). L'invention de savoirs documentaires : Les activités de recherche d'information d'usagers dits «novices». In *Colloque Enjeux et usages des TIC : aspects sociaux et culturels, Bordeaux*, 22–24. 58

Boucher, A. (2011). *Ergonomie web illustrée : 60 sites à la loupe*. Eyrolles. 280

Boulogne, A. (2004). Vocabulaire de la documentation. *Sciences et techniques de l'information*, 333–334. 141

Boulogne, A. (2006). *Comment rédiger une bibliographie*. Armand Colin. 143, 144

Boutell, M. et Luo, J. (2004). Bayesian fusion of camera metadata cues in semantic scene classification. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, 623–630, IEEE. 94

Boutin, É. (2008). *La recherche d'information sur Internet au prisme de la théorie des facettes*. Ph.D. thesis, Toulouse. 72, 75

Breese, J.S., Heckerman, D. et Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *UAÍ98 Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, 43–52, Morgan Kaufmann Publishers Inc., San Francisco. 92

Brickley, D. et Miller, L. (2007). Foaf vocabulary specification. Namespace Document 0.91, ILRT Bristol. 312

Brien, R. (1994). *Science cognitive et formation.*. Presses de l'université du Québec, 3rd edn. 117

Britt, M., Rouet, J.F. et Perfetti, C.A. (1996). Using hypertext to study and reason about historical evidence. In J.F. Rouet, J.J. Levonen, A. Dillon et R.J. Spiro, eds., *Hypertext and cognition*, 43–72, Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, England. 280

Brookes, B. (1980). The foundations of information science. *Journal of information science*, **2**, 125–133. 115

Brown, C. (2011). The poehlman case : understanding and indexing ethical problems in scientific journals. *The Indexer*, **29**, 179–184. 34

Bryson, T. (2010). Making library data available without API's. 227

Burke, R. (2000). Knowledge-based recommender systems. 87, 96

Burke, R. (2002). Hybrid recommender systems : Survey and experiments. *User modeling and user-adapted interaction*, **12**, 331–370. 93, 94, 97

Calenge, B. et di Pietro, C. (2005). Le guichet du savoir®. *Bulletin des Bibliothèques de France*, **50**, 38–42. 82

Caplan, P. et Arms, W. (1999). Reference linking for journal articles. *D-Lib magazine*, **5**, 1082–9873. 231

Card, S.K., Mackinlay, J. et Shneiderman, B., eds. (1999). *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufmann, 1st edn. 282, 283

Carlis, J. et Konstan, J. (1998). Interactive visualization of serial periodic data. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*, 29–38. 282

Cassidy, K., Walsh, A. et Coghlan, B. (2006). Using hyperbolic geometry for visualization of concept spaces for adaptive eLearning. In *Proceedings of the 1st International Workshop on Authoring of Adaptive & Adaptable Hypermedia*, Dublin, Ireland. 282

Chabi, A.H., Kboubi, F. et Ahmed, M.B. (2011). Thematic analysis and visualization of textual corpus. *International Journal of Computer Science & Engineering Survey (IJCSES)*, **2**. 312

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R. *et al.* (2002). Budapest open access initiative. *ARL Bimonthly*. 48

Chartron, G. (2007). Evolution de l'édition scientifique, 15 ans après. In *Actes EUTIC 2007*. 36, 318

Chaudiron, S. et Ihadjadene, M. (2002). Quelle place pour l'usager dans l'évaluation des SRI ? In U. de Toulouse, ed., *Recherches récentes en sciences de l'information :*

Convergences et dynamiques, actes du colloque MICS-LERASS, 211–231, ADBS Editions, Toulouse, France. 226

Chen, Y.C., Shang, R.A. et Kao, C.Y. (2009). The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment. *Electron. Commer. Rec. Appl.*, **8**, 48–58. 86, 88

Chinchor, N. et Robinson, P. (1997). Muc-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference*. 65

Chudnov, D., Binkley, P., Summers, E., Frumkin, J., Giarlo, M.J., Rylander, M. et Singer, R. (2006). Introducing unapi. *Ariadne, Web Magazine for Information Professionals*. 233

Clancey, W.J. (1985). Review of sowa's "Conceptual structures". Tech. Rep. STAN-CS-85 1065, Stanford University, Department of Computer Science, Stanford, CA, USA. 281

Claypool, M., Miranda, T., Gokhale, A., Murnikov, P., Netes, D. et Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of Recommender Systems Workshop at ACM SIGIR*, June, 40–48, ACM. 93, 94, 97

Connolly, D. (2007). Gleaning resource descriptions from dialects of languages (GRDDL). *World Wide Web Consortium, Recommendation REC-grddl-20070911*. 230

Cooper, W. (1971). A definition of relevance for information retrieval. *Information storage and retrieval*, **7**, 19–37. 102

Coutrot, L. (2008). Sur l'usage récent des indicateurs bibliométriques comme outil d'évaluation de la recherche scientifique. *Bulletin de méthodologie sociologique*, 45–50, pour le meilleur et pour le pire, les grandes manœuvres sont lancées dans la recherche française sur le front de la bibliométrie : celle-ci cesse désormais d'être un outil réservé aux documentalistes et spécialistes de l'information. Les princes qui nous gouvernent s'en sont emparé à des fins d'évaluation des disciplines, des laboratoires,

des individus. Les personnels de recherche doivent pouvoir s'informer, réfléchir et critiquer, et, éventuellement inventer des usages de la bibliométrie qui e (...). 36

Crockford, D. (2006a). The application/json media type for javascript object notation (json). Tech. rep., IETF. 299

Crockford, D. (2006b). Json : The fat-free alternative to xml. In *Proceedings of XML conference.*, vol. 2006. 299

Daphy, E. et Ha-Duong, M. (2010). Archives ouvertes. le savoir scientifique est-il en accès libre ? j'ai vu que t'étais connu de hal (conte à rire). 217

de Candolle, A.P. (1813). *Théorie élémentaire de la Botanique*, chap. Théorie des classifications appliquée au règne végétal, 19. Déterville. 255

de Kaenel, I. et Iriarte, P. (2007). Les catalogues des bibliothèques : du web invisible au web social. *Revue électronique de science de l'information*, **1**. 3, 192

De Rosa, C. (2006). *College Students Perceptions of Libraries and Information Resources. A Report to the OCLC Membership*. OCLC. 2, 3

de Saxcé, A. (2006). Les étudiants et la documentation électronique. *Bulletin des Bibliothèques de France*, **51**, 67–68. 84

de Saxcé, A. (2010). Internet en bibliothèques. *Bulletin des Bibliothèques de France*, **55**, 83–84. 84

de Solla Price, D. (1969). The structure of publication in science and technology. *Factors in the Transfer of Technology. MIT Press, Cambridge, MA*, 91–104. 24

Delgado-López-Cózar, E., Robinson-García, N. et Torres-Salinas, D. (2012). Manipulating google scholar citations and google scholar metrics : simple, easy and tempting. 33

Dempsey, L. et Heery, R. (1998). Metadata : a current view of practice and issues. *Journal of Documentation*, **54**, 145–172. 235

Denecker, C., Elisabeth, N. et Thirion, P. (2000). *Les compétences documentaires : des processus mentaux à l'utilisation de l'information*. Presses de l'ENSSIB. 20, 21, 22, 36, 114, 115, 116, 117, 252

Dennis, S., Bruza, P. et McArthur, R. (2002). Web searching : A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, **53**, 120–133. 59

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302. 64

Diederich, J., Balke, W. et Thaden, U. (2007). Demonstrating the semantic growbag : automatically creating topic facets for faceteddblp. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 505–505, ACM. 298, 407

Dillon, A. (1996). TIMS : A framework for the design of usable electronic text. *Cognitive Aspects of Electronic Text Processing*, 99–120. 125

Dinet, J. (2009). Pour une conception centrée utilisateurs des bibliothèques numériques. *Communiquations et images*, 59–74. 313

Dinet, J. et Rouet, J. (2002). La recherche d'information : processus cognitifs, facteurs de difficultés et dimension de l'expertise. In C. Paganelli, ed., *Interaction homme-machine et recherche d'information*, 133–161, Hermès. 55, 122

Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. et Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems CHI 88*, 281–285. 96

Earley, S. (2011). Content curation : Contributing to improved findability. *Information Outlook*, **15**, 14–16. 295

Eisermann, M. (2009). Comment Google classe les pages web. *Images des Mathématiques*, http ://images.math.cnrs.fr/Comment–Google–classe–les–pages.html. 109

Eissen, S. et Stein, B. (2002). Analysis of clustering algorithms for web-based search. *Practical Aspects of Knowledge Management*, **1**, 168–178. 60, 100

Enguehard, C. (1992). *ANA : Acquisition Naturelle Automatique d'un réseau sémantique*. Ph.D. thesis, Université de Compiègne. 66, 301

Fichter, D. (2009). What is a mashup ? In N.C. Engard, ed., *Library Mashups : Exploring New Ways to Deliver Library Data*, XVII, Facet Publishing. 227

Filliatreau, G. (2009). Bibliométrie et évaluation en sciences humaines et sociales : une brève introduction. *Revue d'histoire moderne et contemporaine*, **n° 55-4bis**, 61–66. 36

Fischer, J.L. et Rey, R. (1983). De l'origine et de l'usage des termes taxinomie-taxonomie. *Documents pour l'histoire du vocabulaire scientifique*, **V**, 97–113. 255

Florczak, K. (2005). Formation latex pour windows, mac & linux. 152

Fondin, H. (2001). La science de l'information : posture épistémologique et spécificité disciplinaire. *Documentaliste-Sciences de l'Information*, **Vol. 38**, 112–122. 7

Franklin, C. (2000). How internet search engines work. http ://computer.howstuffworks.com/internet/basics/search-engine.htm. 61

Fritch, J.W. et Mandernack, S.B. (2001). The emerging reference paradigm : A vision of reference services in a complex information environment. 84

Galvin, J. (2005). Alternative strategies for promoting information literacy. *The Journal of academic librarianship*, **31**, 352–357. 84

Gandon, F., Faron-Zucker, C. et Corby, O. (2012). *Le web sémantique, comment lier les données et les schémas sur le web ?*. Infopro, Dunod, dunod edn. 244, 246

Garfield, E. et Small, H. (1985). The geography of sciences : disciplinary and national mappings. *Journal of information science*, **11**, 147–159. 25

Gemmell, J., Schimoler, T., Mobasher, B. et Burke, R. (2010). Hybrid tag recommendation for social annotation systems. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, 829–838, ACM, New York, NY, USA. 99

Genette, G. et McIntosh, A. (1988). The proustian paratexte in reading in and around. *Sub-stance*, **17**, 63–77. 129

Gillon, B. (2004a). Ambiguity, indeterminacy, deixis and vagueness : evidence and theory. *Semantics A reader, ed. S. Davis and Brendan S. Gillon*, 157–187. 273

Gillon, G. (2004b). *Phonological awareness : From research to practice*. The Guilford Press. 64

Grüber, T. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, **5**, 199–220. 288, 290

Grüber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies*, **43**, 907–928. 288, 290

Grün, C. (2006). Pushing XML main memory databases to their limits. In *Proc. of the 18th GI-Workshop on the Foundations of Databases*, 60–64. 410

Guthrie, J. (1988). Locating information in documents : examination of a cognitive model. *Reading Research Quarterly*, **23**, 178–199. 119

Guyot, B. (2004). sciences de l'information et activité professionnelle. *Hermès*, **38**, 38–44. 7

Guyot, B. (2011). *Introduction aux sciences de l'information*, vol. 1. INTD, cours du cycle « ingénierie documentaire » de l'INTD, tome 1, p57. 2

Hascoët, M. et Beaudouin-Lafon, M. (2001). Visualisation interactive d'information. *Revue I3*, **1**, 77–108. 283

Herlocker, J., Konstan, J., Borchers, A. et Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 230–237, ACM. 87

Herman, I. et Marshall, M. (2001). GraphXML—an XML-based graph description format. In *Graph Drawing*, 33–66. 291

Heurgon, E. (1990). Urbanisme et architecture des systèmes d'information. *Technologies de l'Informatique et Societé*, **2**, 39–55. 224

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, **102**, 165–169. 30

Howard, E. et Jankowski, T. (1986). Reference services via electronic mail. *Bulletin of the Medical Library Association*, **74**, 41. 81

Huyghe, G. (2010a). Les services de renseignement virtuel. *Bulletin des Bibliothèques de France*, **55**, 81–82. 82

Huyghe, G. (2010b). Les services de renseignement virtuel. *Bulletin des Bibliothèques de France*, **55**, 81–82. 83

Ibekwe-SanJuan, F. (2007). *Représentation numérique des textes*, 43–58. Hermes science publ., numérique, texte edn. 67

Ibekwe-SanJuan, F. et SanJuan, E. (2003). Termwatch : variations terminologiques et veille scientifique. *Actes du congrés International Society for Knowledge Organization, ISKO 2003*, 1–11. 297

IEEE (2002). ACM Taxonomy, an extended version of the ACM Computing Classification System. `http://www.computer.org/portal/web/publications/acmgeneral`, (consulté le 01/07/12). 252

Ihadjadene, M. et Martins, D. (2004). Experts dans le domaine, experts en Internet : Les effets sur la recherche d'information. *Hermès*. 134

Ippolita (2008). *La face cachée de Google*. Payot. 109

Jackson, C. (2009). Le service de réponses à distance de l'ENSSIB. *Bulletin des Bibliothèques de France*, **54**, 65–68. 83

Jacsó, P. (2010). Metadata mega mess in google scholar. *Online Information Review*, **34**, 175–191. 31, 40

Jakobson, R. et Ruwet, N. (1963). *Essais de linguistique générale*, vol. 1973. Minuit Paris. 225

Jamieson, S. et Howard, R.M. (2011). Unraveling the citation trail. In *Smart Talks*, 8, Project Information Literacy. 10

Jarillon, Pierre (2010). interopérabilité. `http://www.larousse.fr/encyclopedie/article/Laroussefr_-_Article/11007936`. 224

Jepson, B., Rothman, E. et Rosen, R. (2008). *Mac OS X for Unix geeks*. O'Reilly Media. 197

Jerding, D. et Stasko, J. (1998). The information mural : A technique for displaying and navigating large information spaces. *Visualization and Computer Graphics, IEEE Transactions on*, **4**, 257–271. 282

Johnston, P. et Powell, A. (2010). Expressing dublin core metadata using HTML/XHTML meta and link elements. `http://dublincore.org/documents/2008/08/04/dc-html/`, (consulté le 01/07/12). 230

Jones, K.S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**, 11–21. 95

Jouguelet, S. et Vayssade, C. (2010). Comparaison internationale de bibliothèques universitaires : étude de cas. rapport à madame la ministre de l'enseignement supérieur et la la recherche. `http://media.enseignementsup-recherche.gouv.fr/file/2010/78/0/Rapport_etude_comparative_18_fevrier_2010_definitif_137780.pdf`, (consulté le 01/07/12). 3, 40

Kan, M.Y., Klavans, J.L. et Mckeown, K.R. (1998). Linear segmentation and segment significance. In *In Proceedings of the 6th International Workshop of Very Large Corpora*, 197–205. 63

Kapoor, N., Chen, J., Butler, J.T., Fouty, G.C., Stemper, J.A., Riedl, J. et Konstan, J.A. (2007). Techlens : a researcher's desktop. In *Proceedings of the 2007 ACM conference on Recommender systems*, RecSys '07, 183–184, ACM, New York, NY, USA. 88

Kboubi, F., Chaibi, A.H. et BenAhmed, M. (2011). Plateforme de recherche exploratoire et thématique dans un corpus textuel. In ISKO, ed., *Actes de la conférence ISKO, Chapitre Mahgreb, Concepts et outils pour le management de la connaissance (KM)*, ISKO, ISKO, Hammamet, Tunisie. 312

Kboubi, F., Chaibi, A.H. et BenAhmed, M. (2012). Semantic visualization and navigation in textual corpus. *International Journal of Information Sciences and Techniques, IJIST*, **2**, international Journal of Information Sciences and Techniques (IJIST). 312

Kéfi, H. et Kalika, M. (2004). Le cadre analytique structurationniste de l'évaluation des SI. In *Evaluation des systèmes d'information : une perspective organisationnelle*, Col. Gestion, 18, Economica, Paris, economica edn. 224

Kembellec, G. (2009). Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques. In I. Porphyre, ed., *Actes de la deuxième conférence Toth*, 213–231, Annecy, France. 297, 311

Kembellec, G. (2011). Représentation de données et métadonnées dans une bibliothèque virtuelle pour une adéuation avec l'usager et les outils de glanage ou moissonnage scientifique. In ISKO, ed., *Colloque international ISKO-Maghreb*, ISKO, ISKO, Hammamet, Maroc. 226

Kembellec, G., Saleh, I. et Sauvaget, C. (2009). A model of cross language retrieval for IT domain papers trough a map of ACM computing classification system. *IEEE International Conference on Multimedia Computing And Systems Proceedings*, 162–168. xvii, 304, 305, 308

Kuhlthau, C.C. (1991). Inside the search process : Information seeking from the user's perspective. *Journal of the American Society for Information Science*, **42**, 361–371. 123

Kuhlthau, C.C. (2003). *Seeking Meaning*. Libraries Unlimited, 2nd edn. 122

Kuhlthau, C.C. (2005). Information search process. *Hong Kong, China*, **7**. 226

Kunze, J.A. (1999). Encoding dublin core metadata in HTML. http://www.ietf.org/rfc/rfc2731.txt, (consulté le 01/07/12). 228

Kunze, J.A. et Reschke, J.F. (2010). RFC 5791 : RFC 2731 ("Encoding dublin core metadata in HTML") is obsolete. *Internet Engineering Task Force (IETF)*. 230

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.C. et Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLoS ONE*, **6**. 318, 319

Labbe, C. (2010). Ike antkare one of the great stars in the scientific firmament. *ISSI Newsletter*, **6**, 48–52. 33, 34

Lagoze, C. et Van de Sompel, H. (2001). The open archives initiative : building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, JCDL'01, 54–62, ACM, New York, NY, USA. 236

Lam, S.K. et Riedl, J. (2004). Shilling recommender systems for fun and profit. In *WWW 04 Proceedings of the 13th international conference on World Wide Web*, 393–402, ACM Press. 94

Lardy, J.P. (2009). multiplication des modèles économiques d'accès aux publications scientifiques. *Archimag. Hors-série*, 12–15. 36

Lardy, J.P. (2011). Google scholar - articles de revues scientifiques. http://urfist.univ-lyon1.fr/google-scholar-articles-de-revues-scientifiques-537256.kjsp?RH=1215024899213, (consulté le 01/07/12). 40, 41, 217

Le Coadic, Y.F. (1994). *La science de l'information*. 2873, Presses universitaires de France, Paris, France. 6

Le Coadic, Y.F. (1997). *Usages et usagers de l'information*. Armand Colin. 114

Le Coadic, Y.F. (2002). N (ombre) ou lumiére, usage des xmétries en science de l'information et en science de la communication. *Enjeux de la scientométrie et de la bibliométrie*. 27, 30

Le Coadic, Y.F. (2005). Mathématique et statistique en science de l'information et en science de la communication : infométrie mathématique et infométrie statistique des revues scientifiques. *Ci. Inf*, **34**, 15–22. 25

Le Coadic, Y.F. (2008). *Le besoin d'information. Formulation, négociation, diagnostic.* ADBS. 3, 58, 114, 115, 116

Le Coadic, Y.F. (2010). Défense et illustration de la bibliométrie. *Bulletin des Bibliothèques de France*, **55**, 48–51. 24

Le Hégaret, P., Whitmer, R. et Wood, L. (2005). Document Object Model (DOM). `http://www.w3.org/DOM/`, (consulté le 01/07/12). 410

Lee, B.N., Chen, W.Y. et Chang, E.Y. (2006). A scalable service for photo annotation, sharing, and search. In *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, 699, ACM Press, New York, New York, USA. 94

Leibniz, G. (1666). Dissertatio de arte combinatoria. *Sämtliche Schriften und Briefe*, **IV**, 163. xiv, 281

Leresche, F. (2004). Les formats MARC. In *Ècole thématique Documentation en mathématiques*, Lumigny. 141

Letrouit, C. (2005). Des métadonnées pour bien utiliser les ressources électroniques. *Bulletin des Bibliothèques de France*, **50**, 114–115. 224

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, vol. 10, 707–710. 64

Lévy, P. (1998). Sur les chemins du virtuel. *La découverte Poche.* 36, 104

Lewenstein, M., Edwards, G., Tatar, D. et DeVigal, A. (2000). Eye-tracking online news. *Stanford-Poytner Project. Retrieved September*, **15**, 2000. 280

Ley, M. (2001). Dblp dtd. 2001. `http://dblp.uni\discretionary{-}{}{}trier.de/xml/dblp.dtd`, (consulté le 01/07/12). 408

Ley, M. (2009a). Dblp : How the data gets in. *Talk at the University of Manchester*. 309, 408

Ley, M. (2009b). Dblp : some lessons learned. *Proceedings of the VLDB Endowment*, **2**, 1493–1500. 309, 408

Library of The Hebrew University of Jerusalem, C.S. (2009). Computer science - Subject Classification System. `http://www.ma.huji.ac.il/~library/classc.htm`, (consulté le 01/07/12). 263

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, **37**, 30–40. 103

Manning, D., Raghavan, P., C et Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge University Press. 55, 102

Marchionini, G. (1997). *Information seeking in electronic environments*. Cambridge University Press. 130

Marijuan, P., del Moral, R. et Navarro, J. (2012). Scientomics : an emergent perspective in knowledge organization. *Knowledge Organization (ISKO)*, **39**, 153–164. 135

Markey, K. (2007). The online library catalog. *D-Lib Magazine*, **13**. 2, 214

Markey, N. (2009). *Tame the BeaST The B to X of BibTEX*. ENS Cachan, 4th edn. 153

Martin, F. et Bermès, E. (2010). Le concept de collection numérique. *Bulletin des Bibliothèques de France*, **55**, 13–17. 193, 223

Masur, R. (2009). *Etude sur la valorisation des logiciels de gestion de références bibliographiques (LGRB) dans le milieu des bibliothèques universitaires romandes, et sur les avantages et limites du plug-in Zotero*. Ph.D. thesis, Haute école de gestion de Genève (HEG-GE), Genève. 194

Megginson, D. (1998). Simple API for XML (SAX). `http://www.megginson.com/downloads/SAX/`, (consulté le 01/07/12). 410

Merton, R.K. et Barber, E. (2003). *The Travels and Adventures of Serendipity : A Study in Sociological Semantics and the Sociology of Science*. Princeton University Press. 128

Meyriat, J. (1981). Document, documentation, documentologie. l'écrit et le document. *Schéma et Schématisation. 14.*, **51**, 63. 6

Mirkin, B., Nascimento, S., Fenner, T. et Pereira, L.M. (2010). Constructing and mapping fuzzy thematic clusters to higher ranks in a taxonomy. In *Proceedings of the 4th international conference on Knowledge science, engineering and management*, KSEM'10, 329–340, Springer-Verlag, Berlin, Heidelberg. 252

Mitchell, J.S. et Vizine-Goetz, D. (2009). *Encyclopedia of Library and Information Science, third edition*. Marcia J. Bates and Mary Niles Maack, CRC Pres edn. 256

Mkadmi, A. et Saleh, I. (2008). *Bibliothèque numérique et recherche d'informations*. Collection information, hypermédias et communicatio, Hermès Lavoisier. 227

Monday, I. (1996). Les processus cognitifs et la rédaction des résumés. *Documentation et bibliothèques*, **42**, 57–58. 118

Monnin, A. et Félix, E. (2009). Essai de comparaison des ontologies informatiques et philosophiques : entre être et artefacts. In *XVIe Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels de Rochebrune*, Megève, France. 270, 288

Mooers, C. (1948). APPLICATION OF RANDOM CODES, THE GATHERING OF STATISTICAL INFORMATION. 55

Moreau, F. et Claveau, V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In Cépaduès, ed., *Information interaction intelligence*, vol. 6 of *2*, 31–50, INIST-CNRS. 65

Morin, E. (1986). *La méthode : Connaissance de la connaissance, Tome 3*. Seuil. 21

Morin, N. (2003). Contenus et services des sites web des bibliothèques. *Bulletin des Bibliothèques de France*, **48**, 9–13. 81

Morville, P. et Rosenfeld, L. (2006). *Information architecture for the World Wide Web*. O'Reilly Media, Inc. xii, 56

Nelson, M.L., Sompel, H.V.d. et Warner, S. (2002). Advanced overview of version 2.0 of the open archives initiative protocol for metadata harvesting. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '02, 418–418, ACM, New York, NY, USA. 237

Nelson, T. (1965). Complex information processing : a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference*, ACM '65, 84–100, ACM, ACM, New York, NY, USA. 243

Nguyen, C. (2006). Les services de référence virtuels en bibliothèque universitaire. *Bulletin des bibliothèques de France*, **51**, 54–57. 83

Nguyen, C. (2012). Services de questions-réponses en ligne et médiation documentaire numérique : des outils de médiation documentaire à plusieurs facettes. In *Développer la médiation documentaire numérique*, no. 25 in La boîte à outils, 73–80, sd Xavier Galaup, presses de l'ENSIB edn. 85, 86

Nguyen-Xuan, A. (1995). Les mécanismes cognitifs d'apprentissage. *Revue française de pédagogie*, 60. 118

Nicolas, D. (2006). Ambiguité. In D. Godard, L. Roussarie et F. Corblin, eds., *Sémanticlopédie : dictionnaire de sémantique*, GDR Sémantique & Modélisation, CNRS, `http://www.semantique-gdr.net/dico/`, (consulté le 01/07/12). 273

Nielsen, J. (1990). *Hypertext and hypermedia*. Book News, Inc., Portland, Or, USA. 279

Nielsen, J. (1994a). *Usability engineering*. Morgan Kaufmann. 279

Nielsen, J. (1994b). Usability inspection methods. In *Conference companion on Human factors in computing systems*, 413–414. 279

Nielsen, J. et Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems : Empowering people*, 249–256. 279

Nieuwenhuysen, P., Alewaeters, G. et Renard, S. (2005). A new role of libraries and information centers : integrating access to distributed electronic publications. In I. Dobreva et Jan, eds., *From Author to Reader : Challenges for the Digital Content Chain, Proceedings of the 9th ICCC International Conference on Electronic Publishing, ELPUB9, Leuven, June 8-10 2005.*, 13–18, Peeters Publishing, Louvain, Belgique. 224

NISO-PRESS (2004). Understanding metadata. `http://www.niso.org/publications/press/UnderstandingMetadata.pdf`, (consulté le 01/07/12). 142

Novak, J. et Cañas, A. (2008). The theory underlying concept maps and how to construct and use them. *Florida Institute for Human and Machine Cognition Pensacola Fl, www. ihmc. us.[http ://cmap. ihmc. us/Publications/ResearchPapers/TheoryCmaps/TheoryUnderlyingConceptMaps. htm]*. 284

Novak, J. et Gowin, D. (1984). *Learning how to learn*. Cambridge University Press. xiv, 285

Odlyzko, A. (2002). The rapid evolution of scholarly communication. *Learned publishing*, **15**, 7–19. 48

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C. et Johnson, D. (2005). Terrier information retrieval platform. *...in Information Retrieval*, **3408**, 517–519. 96

Page, L. (2001). United states patent : 6285999 - method for node ranking in a linked database. 109

Pagés, R. et Derghal, M. (1984). Réduire ou faciliter l'expression de l'idiosyncrasie individuelle : concepts et esquisse expérimentale. *Bulletin de Psychologie*, **37**, 11–14. 286

Pan, Y., Lu, W., Zhang, Y. et Chili, K. (2007). A static load-balancing scheme for parallel XML parsing on multicore cpus. In *Cluster Computing and the Grid, 2007. CCGRID 2007. Seventh IEEE International Symposium on*, 351–362, IEEE. 411

PATASHNIK, O. (1988). Bibtexing. In *Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77*, 257–286, IEEE. 153

PECCATTE, P. (2007). Métadonnées : une initiation, dublin core, iptc, exif, rdf, xmp, etc. http ://peccatte.karefil.com/Software/Metadata.htm, (consulté le 01/07/12). 142

PÉDAUQUE, R.T. (2006). *Le document à la lumière du numérique*. C&F Editions, Caen, France. 6

PEDIOTAKIS, N. ET HASCOËT-ZIZI, M. (1996). Visual relevance analysis. In *Proceedings of the first ACM international conference on Digital libraries*, DL '96, 54–62, ACM, New York, NY, USA. 283

PÉTROFF, A.J. (1984). Sémiologie de la reformulation dans le discours scientifique et technique. *Langue française*, **64**, 53–67. 23

PETROPOULOS, M., DEUTSCH, A. ET PAPAKONSTANTINOU, Y. (2007a). CLIDE : interactive query formulation for service oriented architectures. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 1119–1121. 90

PETROPOULOS, M., DEUTSCH, A., PAPAKONSTANTINOU, Y. ET KATSIS, Y. (2007b). Exporting and interactively querying web service-accessed sources : The CLIDE system. *ACM Transactions on Database Systems (TODS)*, **32**, 22–78. 90

PEYRE, R. (2007). *Utiliser BibTEX*. École normale supérieure. 153

PINHAS, N. ET KORDON, C. (1997). Du bon usage du facteur d'impact. *Inserm Actualités*, **154**, 7–10. 26

POCHET, B. ET THIRION, P. (1999). Formation documentaire et projets pédagogiques. `http://bbf.enssib.fr/consulter/bbf-1999-01-0016-002`, (consulté le 01/07/12). 36

POIRIER, D., FESSANT, F. ET TELLIER, I. (2010). De la Classification d'Opinion à la Recommandation : l'Apport des Textes Communautaires. *TAL : traitement automatique des langues : revue semestrielle de l'ATALA*, **51**, 19–46. 93, 97

Polanco, X. (1995). Aux sources de la scientométrie. *Solaris*, **2**. 24

Porquet, C. (2005). Une introduction au web sémantique. support de cours. 254

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of documentation*, **25**, 348. 24

Ranganathan, S. (1963). The colon classification. *Rutgers Series on Systems for the Intellectual Organization of Information*, **4**. 72

Ranjard, S. (2000). Pratiques et attentes des publics des médiathèques. *Bulletin des bibliothèques de France*, **45**, 102–107. 195

Rao, R. et Card, S. (1994). The table lens : merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems : celebrating interdependence*, 318–322, ACM. 282, 319

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. et Riedl, J. (1994). GroupLens. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*, 175–186, ACM Press, New York, New York, USA. 92

Richard, J.F. (1990). *Les activités mentales : comprendre, raisonner, trouver des solutions*. Armand Collin, Paris, France. 118

Richard, J.F. et Bonnet, C. (1990). Le traitement de l'information symbolique. In Dunod, ed., *Traité de psychologie cognitive, tome 2*, Dunod, Paris, France. 22

Ripon, R. et Evans, C. (2011). La mise en oeuvre d'une étude quantitative par questionnaire : vices et vertus du chiffre. In *Mener l'enquête. Guide des études de publics en bibliothèque*, La Boîte à outils, 62–79, Presses de l'enssib, Villeurbanne. 195

Robertson, S.E. (1977). Theories and models in information retrieval. *Journal of Documentation*, **33**, 126–148. 57

Roger, D., Lavandier, J. et Kolmayer, E. (2001). Navigation et interfaces : cartes conceptuelles et autres outils. analyse bibliographique, Programme d'Aide à la Recherche en INFOrmation. xiv, 285

Roman, D. (2010). Looking for control. *Communications of the ACM*, **53**, 12–12. 252

Romanello, M. (2008). A semantic linking framework to provide critical value-added services for e-journals on classics. In *Open ScholarProceedings of the 12th International Conference on Electronic Publishing*, 401–414, Leslie Chan and Susanna Mornati, Toronto, Canada. 231

Rouet, J., Coutelet, B. et Dinet, J. (2004). La recherche d'informations dans les documents complexes : processus cognitifs, apprentissage et développement. In *Conférence invitée à la Journée d'études sur le traitement cognitif des systèmes d'informations complexes*. 121

Rouse, W. et Rouse, S. (1984). Human information seeking and design of information systems. *Information Processing & Management*, **20**, 129–138. 134

Ruiz, C. et Noy, J. (2007). Vers une conception globalisée des systémes d'information intégrant tous leurs usages. *La Revue des Sciences de Gestion*, **223**, 87–97. 192

Saint-Dizier, P. (2006). Taxonomie. In D. Godard, L. Roussarie et F. Corblin, eds., *Sémanticlopédie : dictionnaire de sémantique*, GDR Sémantique & Modélisation, CNRS, `http://www.semantique-gdr.net/dico/index.php/Taxonomie`, (consulté le 01/07/12). 283

Salaün, J.M. (2012). *Vu, lu, su : les architectes de l'information face à l'oligopole du Web*. Cahiers libres, La Découverte, Paris, 2nd edn. 6

Salton, G. et Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 355–364. 116

Salton, G. et McGill, M. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc. 116

Salton, G. et Waldstein, R. (1978). Term relevance weights in on-line information retrieval. *Information Processing & Management*, **14**, 29–35. 95

Saracevic, T. (1975). Relevance : A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, **26**, 321–343. 102

Saxproject (1998, 2001). Simple API for XML (SAX 1 et 2). `http://www.saxproject.org/`, (consulté le 01/07/12). 410

Saygin, P., Cicekli, I. et Akman, V. (2000). Turing test : 50 years later. *Minds and Machines*, **10**, 463–518. 63

Scapin, D. et Bastien, J. (1997). Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour and Information Technology*, **4-5**, 220–231. 60

Schafer, J.B., Konstan, J. et Riedi, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce - EC99*, 158–166, ACM Press, New York, New York, USA. 92

Schick, K. (2011). Citation obsession ? get over it ! *The Chronicle of Higher Education*. 8, 9, 163

Shah, B., Rao, P., Moon, B. et Rajagopalan, M. (2009). A data parallel algorithm for xml dom parsing. *Database and XML Technologies*, 75–90. 411

Shannon, C.E. et Weaver, W. (1948). The mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423. 225

Shardanand, U. et Maes, P. (1995). Social Information Filtering : Algorithms for Automating Word of Mouth. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '95*, 210 – 217, ACM Press, New York, New York, USA. 92

Sherman, C. et Price, G. (2002). *The invisible web*. Cyber Age Books. 38

Sieg, A., Mobasher, B. et Burke, R. (2010). Improving the effectiveness of collaborative recommendation with ontology-based user profiles. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 39–46, ACM. xii, 89, 90

Simonnot, B. et Gallezot, G. (2009). *L'entonnoir : Google sous la loupe des sciences de l'information et de la communication*. C&F Editions. 225

Simonsson, M. et Johnson, P. (2006). Defining it governance-a consolidation of literature. In *the 18th Conference on Advanced Information Systems Engineering*, 6, `http://www.ics.kth.se/Publikationer/Working%20Papers/EARP-WP-2005-MS-04.pdf`, (consulté le 01/07/12). 4

Singer, R. (2009). Making your data available to be mashed up. In N.C. Engard, ed., *Library Mashups : Exploring New Ways to Deliver Library Data*, Facet Publishing. 227

Sitbon, L. et Bellot, P. (2005). Segmentation thématique par chaînes lexicales pondérées. *TALN 2005*, **1**, 505–510. 64, 65

Sitbon, L., Bellot, P. et Blache, P. (2007). Traitements phrastiques phonétiques pour la réécriture de phrases dysorthographiées. *Actes de TALN*, **7**. 64

Sitbon, L., Bellot, P. et Blache, P. (2008). Eléments pour adapter les systèmes de recherche d'information aux dyslexiques. *Traitement Automatique des Langues*, **48**, 123–147. 64

Sottet, J., Calvary, G. et Favre, J. (2005). Ingénierie de l'interaction homme-machine dirigée par les modèles. *IDM'05 Premières Journées sur l'Ingénierie Dirigée par les Modèles*. 60

Sowa, J. (1983). *Conceptual structures : information processing in mind and machine*. Addison-Wesley Pub., Reading, MA. 281

Sowa, J.F. (2000). *Knowledge representation logical, philosophical, and computational foundations*, vol. 38. MIT Press. 281

Spink, A., Wolfram, D., Jansen, M. et Saracevic, T. (2001). Searching the web : The public and their queries. *Journal of the American Society for Information Science and Technology*, **52**, 226–234. 67

Studer, R., Benjamins, V. et Fensel, D. (1998). Knowledge engineering : principles and methods. *Data & knowledge engineering*, **25**, 161–197. 269, 270, 272

Swan, A. et Brown, S. (2005). Open access self-archiving an author study. Tech. rep., Key Perspectives Limited, http://cogprints.org/4385/01/jisc2.pdf, (consulté le 01/07/12). 2

Tardif, J. (1997). *Pour un enseignement stratégique : l'apport de la psychologie cognitive*. Editions Logiques, Montréal. 118

Tardif, J. (1998). La construction des connaissances. *Pédagogie collégiale*, **11**, 4–9. 22, 118

Thomson-Reuteurs (2008). *RIS Format Documentation*. ResearchSoft, http://www.adeptscience.co.uk/kb/article/FE26, (consulté le 01/07/12). 157

Toulin, S., Chazelas, M. et Palencia, F. (2009). JabRef : outil de gestion de données bibliographiques Utilisation, avantages et inconvénients. *e-migrinter*, 58–64. 170

Tricot, A. (1995). Un point sur l'ergonomie des interfaces hypermédia. *Le travail humain*, **58**, 17–45. 280

Tricot, A. (1998). Chercher de l'information dans un hypertexte : vers un modèle des processus cognitifs. In J. Rouet et A. Tricot, eds., *Les hypermédias, approches cognitives et ergonomiques*, 57–74, Hermès, Paris. 280

Tricot, A. et Rouet, J. (2004). Activités de navigation dans les systèmes d'information. *Psychologie ergonomique : tendances actuelles. PUF, Paris*, 71–95. 115

Tricot, C. (2006). *La cartographie sémantique, des connaissances à la carte*. Thèse de doctorat en informatique, Université de Savoie. 282, 288, 293

Tricot, C. et Roche, C. (2006). Visualisation of ontology : a focus and context approach. In *International Conference on Multidisciplinary Information Sciences and Technologies*. 282, 287, 288, 293

Trust (2003). *Economic analysis of scientific research publishing, a report commissioned by the Wellcome Trust*. Wellcome Trust. 318

Turing, A. (1950). Computing machinery and intelligence. *Mind*, **59**, 433–460. 63

Tyckoson, D. (2003). On the desirableness of personal relations between librarians and readers : the past and future of reference service. *Reference services review*, **31**, 12–16. 86

Vajou, M., Martinez, R. et Chaudiron, S. (2009). Les enjeux économiques de l'édition scientifique, technique et médicale. *Les Cahiers du numérique*, **Vol. 5**, 143–172. 318

van Andel, P. et Bourcier, D. (2008). *De la sérendipité dans la science, la technique, l'art et le droit : Lecons de l'inattendu*. L'Act Mem. 128

van Andel, P. et Jacquemin, C. (2005). Sérendipité, ou de l'art de faire des trouvailles. *Les automates intelligents robotique, vie artificielle, réalité virtuelle*. 128

Van de Sompel, H. et Lagoze, C. (2007). Interoperability for the discovery, use, and re-use of units of scholarly communication. *CyberInfrastruture Technology Watch Quarterly*, **3**. 318

Van der Velde, W. (2012). Smart searching on academic content. In G. Chartron, I. Saleh et G. Kembellec, eds., *Journée d'études sur les systèmes de recommandation*, CNAM, Paris, France. 44

Van Ham, F. et Van Wijk, J.J. (2003). Beamtrees : Compact visualization of large hierarchies. *Information Visualization*, **2**, 31–39. 282

Van Rijsbergen, C. et Lalmas, M. (1996). Information calculus for information retrieval. *Journal of the American Society for Information Science*, **47**, 385–398. 106

Vassiliadis, P. (2009). A survey of Extract–transform–Load technology. *IJDWM*, **5**, 1–27. 224

Véronis, J. (2003). Hyperlex : cartographie lexicale pour la recherche d'informations. *Actes de TALN*, 265–274. 273

Véronis, J. (2006). Etude comparative de six moteurs de recherche. *Université de Provence*. xv, 77, 78

Vivarès, D. (2009). Refworks. http://urfist.u-strasbg.fr/uploads/support_cours/bddbi_fi/refworks_02_09.pdf, (consulté le 01/07/12). 183, 189

Vourc'h, R. (2010). Les étudiants, le livre et les bibliothèques universitaires. *Bulletin des Bibibliothèques de France*, **55**. 3, 219

Wager, E. (2011). Coping with scientific misconduct. *BMJ*, **343**. 34

Wahnich, S. (2006). Enqutes quantitatives et qualitatives, observation ethnographique. *Bulletin des Bibliothèques de France*, **51**, 8–12. 195

Wal, T.V. (2006). Understanding folksonomy : Tagging that works. In *d.Construct 2006 : Web Application and Web 2.0 Conference*, dConstruct, http://2006.dconstruct.org, (consulté le 01/07/12). 297

Walpole, H. (1754). Three princes of serendip. 128

Weibel, S., Kunze, J., Lagoze, C. et Wolf, M. (1998). Dublin core metadata for resource discovery. http://www.ietf.org/rfc/rfc2413.txt, (consulté le 01/07/12). 146

Weil-Barais, A., Pedinielli, J.L., Streri, A. et Dubois, D. (2011). *L'homme cognitif*. Quadrige Manuels, Presses Universitaires de France - PUF, Paris, France, 5th edn. 22, 117

Wetzel, L. (2011). Types and tokens. *The Stanford Encyclopedia of Philosophy*. 281

Williams, C., Mobasher, B., Burke, R., Sandvig, J.J. et Bhaumik, R. (2006). Detection of Obfuscated Attacks in Collaborative Recommender Systems. In *Proceedings of the ECAI'06 Workshop on Recommender Systems*. 94

Wilson, P. (1973). Situational relevance. *Information storage and retrieval*, **9**, 457–471. 102

Wright, K. et McDaid, C. (2011). Reporting of article retractions in bibliographic databases and online journals. *Journal of the Medical Library Association : JMLA*, **99**, 164. 34

Zhang, J. (2008). *Visualization for information retrieval*, vol. 23. Springer. 283, 288, 293, 294, 301

Zhang, J. et Marchionini, G. (2004). Coupling browse and search in highly interactive user interfaces : a study of the relation browser++. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, 384. 288

Zhang, W. (2002). Developing web-enhanced learning for information fluency : A liberal arts college. *Reference & User Services Quarterly*, **41**, 356–63. 288

Zloof, M. (1977). Query-by-example : A data base language. *IBM systems Journal*, **16**, 324–343. 88

Zweifel, C. (2008). Logiciels de gestion de références bibliographiques : citons le libre ! *Ressi : Revue Electronique Suisse de Science de l'information*, **7**, –. 194

Index

H-index, 30

A

ABES, 2
ACM, 60
ACM, Association for Computing Machinery, 53, 254, 259
ACM CCS, 276
ADBS, Association des professionnels de l'information et de la documentation, 56
AJAX, 300
Amazon, 89
Ambiguïté lexicale, 275
Annuaire, 60
Antidictionnaire, 67
API, 80, 298
API *Application programming interface*, interface de programmation applicative, 80
ASCII, American Standard Code for Information Interchange, 259
ASKAL, 82

B

BATES, Marcia J., 57, 129
Berrypicking, 129
Bibliométrie, 24
Bibliothèque du Congrès, 151
BibTeX, 154
Booléen, 67
Bruit, 82
Burke, 94

C

Carte cognitive, 285
Carte conceptuelle, 285
Carte heuristique, 285
Catégorisation, 22
Catalogue, 92
CCS, 259
CCS, *ACM*-Computing Classification System, 254

CiteSeer, 31
Concept map, 285
Copernic, 81
Couperin, 49
Cross-genre niche, 95

D

Déluge informationnel, 36, 67
Désambiguïsation, 275
Del.icio.us, 100
Dillon, Andrew, 127
Distance, 70
Dmoz, 60

E

Ebsco, 49
ECMA, 157
Enguehard, Chantal, 66
ENSSIB, 84
EST, Évaluation, Sélection, Traitement., 123
Exalead, 78
EXIF, 95

F

Facebook, 83
Facette, 72, 82
Facteur d'impact, 29
Feuille de style, 162
Flexion, 66
Folksonomy, 100

G

Garfield, Eugène, 25
Gmail, 88
Google, 60, 70, 78, 88
Google Scholar, 31
GraphXML, 293
Grey sheep, 95

H

H-index, 31
Hypertexte, 276

I

Idiosyncrasie, 288
IEEE, 60
IFLA, 151
Indice d'immédiateté, 29
Infobésité, 36
Information overload, 36
Inverse definition frequency, 96
Isidore, 49, 238
ISO, 157
ISO 2709, 147
ISP Infomation Search Process, 124

J

JSON, 300

K

Kuhlthau, Carol, 124

L

LaTeX, 154
Lemmatisation, 65, 80
Library of Congress, 151
LOC, 151

M

Métamoteur, 80

MARC, 151
MARC21, 151
Marchonini, Gary, 131
Mind map, 285
MODS, 151
Mot clé, 67
Mot vide, 67
Moteur à curseur, 76
Moteur de recherche, 61
MS Word, 157
MSN, 78

O

OCLC, 2, 84
OCLC, Online Computer Library Center, 2, 258
Opérateur, 67
OPAC, 85

P

Pagerank, 111
Panda, 62, 113
Paratexte, 24, 130
Pearson, 93
Pertinence, 78, 82
Plugin, 235
Point de vue, 275
Polysémie, 69

Q

QBE, 89

R

Racinisation, 65, 80
Ranganathan, 72
Recommendation, 87
RIA, Rich Internet Application, 169, 170, 183, 184
Robot, 61
RSS, 83
Rue des Facs, 85

S

Sérendipité, 130
Salton Gerard, 118
Scientométrie, 24
Scientomics, 137
scrolling, 308
Segmentation, 64
SIGB, 275
Spider, 61
Spip, 235
STAR, Signalement des thèses électroniques, archivage et recherche, 2
Stemmer, 65
Stop list, 80
Stop word, 67
Style bibliographique, 162
Sudoc, 39

T

TAL, 63
Taxonomie, 255
Term Frenquency, 96
Term frequency - inverse definition frequency, 96
Thésaurus, 255
Tims, 127

Tokenisation, 64
Traitement de texte, 157
Troncature, 66
Turing, Alan, 63
Typicalité, 22

U

UNIMARC, 151

V

Visualisation, 283
Voilà, 78

W

Wikipédia, 60
Word, 157
WordPress, 235
Wrapper, 80
WYSIWYG, 157

X

XML, 151

Y

Yahoo, 76, 78

Quatrième partie

Annexes

Annexe A

Publications relatives à la thèse

Kembellec, G. (2008). Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques. In I. Porphyre, ed., *Actes de la deuxième conférence Toth*, 213–231, Annecy, France.

Kembellec, G., Saleh, I., Sauvaget, C. (2009). A model of cross language retrieval for IT domain papers through a map of ACM Computing Classification System In IEEE, ed., *International Conference on Multimedia Computing and Systems*, 162–168, Ouarzazate, Morocco.

Kembellec, G. (2011). Représentation de données et métadonnées dans une bibliothèque virtuelle pour une adéquation avec l'usager et les outils de glanage ou moissonnage scientifique. In ISKO, ed., *Colloque international ISKO-Maghreb*, ISKO, Hammamet.

Annexe B

Chiffres ISI

Disciplines	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	Total par discipline
Ingénérie	12 636	11 702	11 760	11 815	12 312	11 637	12 560	12 475	13 467	12 671	14 455	14 347	14 048	15 140	181 026
Astronomie	2 483	2 651	2 557	1 932	2 829	2 566	2 697	2 996	2 719	2 874	2 653	2 885	2 805	2 981	37 829
Chimie	14 915	15 219	14 375	14 414	14 491	14 560	14 342	14 043	15 763	15 110	17 138	16 266	15 268	16 684	212 589
Physique	19 709	18 906	18 048	17 966	18 074	16 897	17 385	17 301	18 657	17 653	20 551	19 797	18 212	20 701	259 857
Géologie	9 821	9 904	9 214	9 820	9 873	9 779	10 065	9 204	11 486	10 436	11 112	11 353	11 207	12 463	145 736
Mathématiques	3 190	3 272	3 051	3 483	3 561	3 758	3 657	3 556	3 760	3 562	3 922	3 930	3 810	4 477	50 988
Informatique	1 914	1 923	1 929	1 952	2 008	2 075	2 083	1 999	2 254	2 173	2 390	2 521	2 329	2 337	29 887
Agriculture	3 648	3 346	3 605	3 464	2 804	3 640	3 203	3 924	3 903	3 526	3 561	3 964	3 624	3 859	50 069
Biologie	54 076	53 301	51 718	52 076	51 443	50 997	52 696	50 108	52 981	51 502	55 502	51 783	51 865	53 403	733 451
Médecine	51 943	51 404	52 146	52 402	52 550	51 886	53 437	51 384	56 128	53 541	60 130	57 052	57 535	62 017	763 555
Sciences de la vie	2 338	2 811	3 009	3 009	2 876	2 942	3 146	2 958	2 964	2 699	3 670	4 456	3 947	4 706	45 530
Psychologie	7 830	8 258	8 130	7 736	7 806	7 499	7 809	7 691	7 892	7 410	8 482	8 821	8 665	10 210	114 239
Sciences sociales	9 751	10 094	9 935	9 745	9 763	9 888	9 985	9 763	10 501	9 754	10 790	10 700	10 993	12 935	144 599
Total par année	194 255	192 790	189 477	189 813	190 390	188 125	193 066	187 400	202 475	192 913	214 557	207 874	204 307	221 913	2 769 356

Figure B.1: Publication d'articles scientifiques dans des revues indexées par *Science Citation Index*(SCI) et *Social Sciences Citation Index* (SSCI) sur la période 1995 à 2008

Annexe C

Pratique bibliographique dans l'enseignement supérieur

Ce questionnaire s'inscrit dans le cadre d'une étude menée en milieu universitaire auprès de d'enseignants-chercheurs, de doctorants, d'ingénieurs et de personnels des SCD. Remplir ce questionnaire devrait vous prendre entre 5 et 10 minutes. Les champs intégrant la mention * sont obligatoires. D'avance merci.

1. Indiquez si vous exercez votre recherche principalement* :
   (a) en sciences humaines
   (b) en sciences dures
   (c) autre
   (d) sans objet
2. Indiquez si vous êtes plutôt :
   (a) post-doctorant
   (b) enseignant chercheur
   (c) personnel de SCD ou bibliothèque
   (d) autre
3. Quel est votre système d'exploitation favori (Celui que vous utilisez le plus souvent) * :
   (a) Linux

(b) Windows

(c) Mac OS

(d) autre

4. Pour la rédaction de votre mémoire de thèse, ou votre activité quotidienne utilisez-vous : *

   (a) un compilateur de document (LaTex, TEX)

   (b) un éditeur graphique (Word, OpenOffice . . .)

   (c) autre

5. Si vous êtes un utilisateur de LATEX quel outil d'édition de texte utilisez vous pour le corps du document (plusieurs réponses possibles) ?

   (a) Éditeur à compilation intégrée (TexWorks, Emacs...)

   (b) Éditeur basique (vi, ed, Notepad...)

   (c) Autre : vous pouvez préciser

6. Pour la constitution de votre bibliographie, Intégrez vous vos références et citation le plus souvent

   (a) à la main dans votre texte ?

   (b) Utilisez vous un fichier indépendant (BibTEX, OpenDocument XML) que vous alimentez manuellement ?

   (c) Utilisez vous les outils intégrés dans votre éditeur graphique (Word, OpenOffice. . .) ?

   (d) Utilisez vous un fichier indépendant (BibTEX, OpenDocument XML) que vous générez grâce à un logiciel ?

7. Quels formats bibliographiques connaissez vous parmi ceux là (Même de nom) : *

   (a) BibTEX

   (b) RIS

   (c) Ovid

   (d) XML OpenDocument

   (e) Aucun

8. Quels formats bibliographiques utilisez vous parmi ceux là :
   (a) BibTEX
   (b) RIS
   (c) Ovid
   (d) XML OpenDocument
   (e) Aucun
   (f) Autre (vous pouvez préciser) :
9. Pour la gestion de votre recherche bibliographique utilisez-vous un outil dédié (que ce soit un logiciel à installer ou accessible sur le web de gestion bibliographique) * ?
   (a) oui
   (b) non
   (c) sans objet
10. Si oui quel est-il ? (plusieurs choix possibles)
   (a) JabRef
   (b) Refworks
   (c) BibDesk
   (d) Zotero
   (e) Mendeley
   (f) EndNote
   (g) Bibus
   (h) RefBase
   (i) Autre : vous pouvez préciser.
11. Etes vous plus sensibles aux * :
   (a) Logiciels libres ou Open Source (gratuits ou payants)
   (b) Aux logiciels propriétaires (gratuits ou payants)
   (c) Indifférent à ce critère

C. PRATIQUE BIBLIOGRAPHIQUE DANS L'ENSEIGNEMENT SUPÉRIEUR

12. Votre choix potentiel d'un logiciel de gestion bibliographique s'oriente plutôt sur * (plusieurs choix possibles) :

    (a) Les fonctionnalités du produit
    (b) La facilité d'installation
    (c) La gratuité La disponibilité au sein de la structure (labo, SCD)
    (d) Le choix de vos collègues en la matière
    (e) NSP
    (f) Autre : vous pouvez préciser

13. Si vous avez à votre disposition un logiciel de gestion bibliographique est il :

    (a) choisi et installé par vos soins ?
    (b) le choix de votre organisme de rattachement ?
    (c) déployé par défaut ?
    (d) le choix de votre organisme de rattachement, installé par vos soins ?

14. De manière générale de quelle manière trouvez-vous votre information scientifique * (3 réponses maximum souhaitées) ?

    (a) Sur le catalogue de votre centre de documentation (OPAC)
    (b) Sur internet avec les moteurs de recherche traditionnels (Google, Bing, Yahoo ...)
    (c) Sur des moteurs scientifiques (Google Scholar, Scirus...)
    (d) Sur le SUDOC
    (e) Sur les sites d'éditeurs scientifiques (IEEE, ACM, Elsevier, Springer ...)
    (f) Sur archives ouvertes (HAL, ArXiV, Archivesic ....)
    (g) Autres
    (h) Sans objet

Annexe D

Questionnaire usabilité de l'ACM CCS et d'OntologyNavigator

Ce sondage à destination d'élèves ingénieurs en informatique est individuel, anonyme sur une base volontaire. Il est réalisé après 6 heures de théories de la recherche d'information scientifique et technique et 3 heures de TDs. Les champs signalés par le signe « * » sont obligatoires.

Avant ce cours aviez-vous connaissance de l'ACM CCS ?*

- Oui.
- Non.

Quel est votre état d'esprit face à la taille et la complexité de cette classification ? Pensez vous qu'une telle classification aide réellement l'utilisateur

final à classer et à retrouver de l'information ? *

- Oui.
- Non.
- Pas sûr.
- NSP.

D. QUESTIONNAIRE USABILITÉ DE L'ACM CCS ET D'ONTOLOGYNAVIGATOR

***A priori*, quel est votre modèle de représentation de l'ACM CCS préféré ? ***

- Le mode texte simple (site ACM).
- Le mode HTML (site ACM).
- Le mode HTML avec coloration et hyperliens vers les résultats (site de l'enseignant).
- L'hypergraphe, focus + context (site de l'enseignant).

Après avoir effectué une recherche sur le portail ACM via l'interface avancée puis depuis ontologynavigator, quelle interface trouvez-vous la plus pratique pour naviguer l'ACM CCS et accéder à l'IST ? *

- Le mode HTML avec coloration et hyperliens vers les résultats (site de l'enseignant).
- L'hypergraphe, focus + context (site de l'enseignant).
- Le formulaire classique de l'ACM CCS.
- Le formulaire avancé de l'ACM CCS.
- NSP (Pour ne se prononce pas choisissez autre et expliquez votre réponse).

En effectuant une recherche sur ontologynavigator (base locale ou externe) pour votre mémoire, évaluez l'adéquation des résultats avec la recherche de 1 à 10 (1 : Pas du tout en adéquation, 10 : Parfaitement en adéquation)

1	2	3	4	5	6	7	8	9	10

Annexe E

La classification informatique d'ACM

E.1 Le contenu de la classificatoire

THE ACM COMPUTING CLASSIFICATION SYSTEM (1998)
- JUILLET 2012 -

A. General Literature

- A.0 GENERAL
  - Biographies/autobiographies
  - Conference proceedings
  - General literary works (e.g., fiction, plays)
- A.1 INTRODUCTORY AND SURVEY
- A.2 REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
- A.m MISCELLANEOUS

B. Hardware

- B.0 GENERAL
- B.1 CONTROL STRUCTURES AND MICROPROGRAMMING (D.3.2)
  - B.1.0 General
  - B.1.1 Control Design Styles
    - Hardwired control [**] [1]

1. Le signe [**] indique que l'élément n'est plus utilisé pour la classification depuis janvier 1998, mais qu'il est toujours possible de l'utiliser pour rechercher des documents indexés avant cette période.

- Microprogrammed logic arrays [**]
- Writable control store [**]

B.1.2 Control Structure Performance
- Analysis and Design Aids
- Automatic synthesis [**]
- Formal models [**]
- Simulation [**]

B.1.3 Control Structure Reliability, Testing, and Fault-Tolerance [**] (B.8)
- Diagnostics [**]
- Error-checking [**]
- Redundant design [**]
- Test generation [**]

B.1.4 Microprogram Design Aids (D.2.2, D.2.4, D.3.2, D.3.4)
- Firmware engineering [**]
- Languages and compilers
- Machine-independent microcode generation [**]
- Optimization
- Verification [**]

B.1.5 Microcode Applications
- Direct data manipulation [**]
- Firmware support of operating systems/instruction sets [**]
- Instruction set interpretation
- Peripheral control [**]
- Special-purpose [**]

B.1.m Miscellaneous

B.2 ARITHMETIC AND LOGIC STRUCTURES

B.2.0 General

B.2.1 Design Styles (C.1.1, C.1.2)
- Calculator [**]
- Parallel Pipeline

B.2.2 Performance Analysis and Design Aids [**] (B.8)
- Simulation [**]
- Verification [**]
- Worst-case analysis [**]

B.2.3 Reliability, Testing, and Fault-Tolerance [**] (B.8)
- Diagnostics [**]
- Error-checking [**]
- Redundant design [**]

- Test generation [**]

B.2.4 High-Speed Arithmetic (NEW!)

- Algorithms NEW!
- Cost/performance NEW!

B.2.m Miscellaneous

B.3 MEMORY STRUCTURES

B.3.0 General

B.3.1 Semiconductor Memories (NEW!) (B.7.1)

- Dynamic memory (DRAM) NEW!
- Read-only memory (ROM) NEW!
- Static memory (SRAM) NEW!

B.3.2 Design Styles (D.4.2)

- Associative memories
- Cache memories
- Interleaved memories [**]
- Mass storage (e.g., magnetic, optical, RAID)
- Primary memory
- Sequential-access memory [**]
- Shared memory
- Virtual memory

B.3.3 Performance Analysis and Design Aids [**] (B.8, C.4)

- Formal models [**]
- Simulation [**]
- Worst-case analysis [**]

B.3.4 Reliability, Testing, and Fault-Tolerance [**] (B.8)

- Diagnostics [**]
- Error-checking [**]
- Redundant design [**]
- Test generation [**]

B.3.m Miscellaneous

B.4 INPUT/OUTPUT AND DATA COMMUNICATIONS

B.4.0 General

B.4.1 Data Communications

- Devices Processors [**]
- Receivers (e.g., voice, data, image) [**]
- Transmitters [**]

B.4.2 Input/Output Devices

- Channels and controllers
- Data terminals and printers
- Image display
- Voice

B.4.3 Interconnections (Subsystems)

- Asynchronous/synchronous operation
- Fiber optics Interfaces
- Parallel I/O NEW !
- Physical structures (e.g., backplanes, cables, chip carriers) [**]
- Topology (e.g., bus, point-to-point)

B.4.4 Performance

- Analysis and Design Aids [**] (B.8)
- Formal models [**]
- Simulation [**]
- Verification [**]
- Worst-case analysis [**]

B.4.5 Reliability, Testing, and Fault-Tolerance [**] (B.8)

- Built-in tests [**]
- Diagnostics [**]
- Error-checking [**]
- Hardware reliability [**]
- Redundant design [**]
- Test generation [**]

B.4.m Miscellaneous

B.5 REGISTER-TRANSFER-LEVEL IMPLEMENTATION

B.5.0 General

B.5.1 Design

- Arithmetic and logic units
- Control design
- Data-path design
- Memory design Styles (e.g., parallel, pipeline, special-purpose)

B.5.2 Design Aids

- Automatic synthesis
- Hardware description languages
- Optimization
- Simulation
- Verification

B.5.3 Reliability and Testing [**] (B.8)

- Built-in tests [**]
- Error-checking [**]
- Redundant design [**]
- Test generation [**]
- Testability [**]

B.5.m Miscellaneous

B.6 LOGIC DESIGN

B.6.0 General

B.6.1 Design Styles

- Cellular arrays and automata
- Combinational logic
- Logic arrays
- Memory control and access [**]
- Memory used as logic [**]
- Parallel circuits
- Sequential circuits

B.6.2 Reliability and Testing [**] (B.8)

- Built-in tests [**]
- Error-checking [**]
- Redundant design [**]
- Test generation [**]
- Testability [**]

B.6.3 Design Aids

- Automatic synthesis
- Hardware description languages
- Optimization
- Simulation
- Switching theory
- Verification

B.6.m Miscellaneous

B.7 INTEGRATED CIRCUITS

B.7.0 General

B.7.1 Types and Design Styles

- Advanced technologies
- Algorithms implemented in hardware
- Gate arrays
- Input/output circuits
- Memory technologies

- Microprocessors and microcomputers
- Standard cells [**]
- VLSI (very large scale integration)

B.7.2 Design Aids

- Graphics
- Layout
- Placement and routing
- Simulation
- Verification

B.7.3 Reliability and Testing [**] (B.8)

- Built-in tests [**]
- Error-checking [**]
- Redundant design [**]
- Test generation [**]
- Testability [**]

B.7.m Miscellaneous

B.8 PERFORMANCE AND RELIABILITY (NEW !) (C.4)

B.8.0 General (NEW !)

B.8.1 Reliability, Testing, and Fault-Tolerance (NEW !)

B.8.2 Performance Analysis and Design Aids (NEW !)

B.8.m Miscellaneous (NEW !)

B.m MISCELLANEOUS

- Design management

C. Computer Systems Organization

C.0 GENERAL

- Hardware/software interfaces
- Instruction set design (e.g., RISC, CISC, VLIW)
- Modeling of computer architecture
- System architectures
- Systems specification methodology

C.1 PROCESSOR ARCHITECTURES

C.1.0 General

C.1.1 Single Data Stream Architectures

- Multiple-instruction-stream, single-data-stream processors (MISD) [**]
- Pipeline processors [**]

- RISC/CISC, VLIW architectures NEW !
- Single-instruction-stream, single-data-stream processors (SISD) [**]
- Von Neumann architectures [**]

C.1.2 Multiple Data Stream Architectures (Multiprocessors)

- Array and vector processors
- Associative processors
- Connection machines
- Interconnection architectures (e.g., common bus, multi port memory, crossbar switch)
- Multiple-instruction-stream, multiple-data-stream processors (MIMD)
- Parallel processors [**]
- Pipeline processors [**]
- Single-instruction-stream, multiple-data-stream processors (SIMD)

C.1.3 Other Architecture Styles

- Adaptable architectures
- Analog computers NEW !
- Capability architectures [**]
- Cellular architecture (e.g., mobile)
- Data-flow architectures
- Heterogeneous (hybrid) systems NEW !
- High-level language architectures [**]
- Neural nets
- Pipeline processors NEW !
- Stack-oriented processors [**]

C.1.4 Parallel Architectures (NEW !)

- Distributed architectures NEW !
- Mobile processors NEW !

C.1.m Miscellaneous

- Analog computers [**]
- Hybrid systems [**]

C.2 COMPUTER-COMMUNICATION NETWORKS

C.2.0 General

- Data communications
- Open Systems
- Interconnection reference model (OSI)
- Security and protection (e.g., firewalls)

C.2.1 Network Architecture and Design

- Asynchronous

- Transfer Mode (ATM) NEW !
- Centralized networks [**]
- Circuit-switching networks
- Distributed networks
- Frame relay networks NEW !
- ISDN (Integrated Services Digital Network)
- Network communications
- Network topology
- Packet-switching networks
- Store and forward networks
- Wireless communication NEW !

C.2.2 Network Protocols

- Applications (SMTP, FTP, etc.) NEW !
- Protocol architecture (OSI model)
- Protocol verification
- Routing protocols NEW !

C.2.3 Network Operations

- Network management
- Network monitoring
- Public networks

C.2.4 Distributed Systems

- Client/server NEW !
- Distributed applications
- Distributed databases
- Network operating systems

C.2.5 Local and Wide-Area Networks

- Access schemes Buses
- Ethernet (e.g., CSMA/CD) NEW !
- High-speed (e.g., FDDI, fiber channel, ATM) NEW !
- Internet (e.g., TCP/IP) NEW !
- Token rings

C.2.6 Internetworking (NEW !) (C.2.2)

- Routers NEW !
- Standards (e.g., TCP/IP) NEW !

C.2.m Miscellaneous

C.3 SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS (J.7)

- Microprocessor/microcomputer applications
- Process control systems
- Real-time and embedded systems

- Signal processing systems
- Smartcards NEW !

C.4 PERFORMANCE OF SYSTEMS

- Design studies
- Fault tolerance NEW !
- Measurement techniques
- Modeling techniques
- Performance attributes
- Reliability, availability, and serviceability

C.5 COMPUTER SYSTEM IMPLEMENTATION

C.5.0 General

C.5.1 Large and Medium ("Mainframe")

- Computer
- Super (very large) computers

C.5.2 Minicomputers [**]

C.5.3 Microcomputers

- Microprocessors
- Personal computers
- Portable devices (e.g., laptops, personal digital assistants) NEW !
- Workstations

C.5.4 VLSI Systems

C.5.5 Servers (NEW !)

C.5.m Miscellaneous

C.m MISCELLANEOUS

D. Software

D.0 GENERAL

D.1 PROGRAMMING TECHNIQUES (E)

D.1.0 General

D.1.1 Applicative (Functional) Programming

D.1.2 Automatic Programming (I.2.2)

D.1.3 Concurrent Programming

- Distributed programming
- Parallel programming

D.1.4 Sequential Programming

D.1.5 Object-oriented Programming

D.1.6 Logic Programming

D.1.7 Visual Programming

D.1.m Miscellaneous

D.2 SOFTWARE ENGINEERING (K.6.3)

D.2.0 General (K.5.1)
- Protection mechanisms
- Standards

D.2.1 Requirements/Specifications (D.3.1)
- Elicitation methods (e.g., rapid prototyping, interviews, JAD) NEW !
- Languages
- Methodologies (e.g., object-oriented, structured)
- Tools

D.2.2 Design Tools and Techniques
- Computer-aided software engineering (CASE)
- Decision tables
- Evolutionary prototyping NEW !
- Flow charts
- Modules and interfaces
- Object-oriented design methods NEW !
- Petri nets
- Programmer workbench [**]
- Software libraries
- State diagrams NEW !
- Structured programming [**]
- Top-down programming [**]
- User interfaces

D.2.3 Coding Tools and Techniques
- Object-oriented programming NEW !
- Pretty printers
- Program editors
- Reentrant code [**]
- Standards Structured programming NEW !
- Top-down programming NEW !

D.2.4 Software/Program Verification (F.3.1)
- Assertion checkers
- Class invariants NEW !
- Correctness proofs

- Formal methods NEW !
- Model checking NEW !
- Programming by contract NEW !
- Reliability Statistical methods NEW !
- Validation

D.2.5 Testing and Debugging

- Code inspections and walk-throughs
- Debugging aids
- Diagnostics
- Distributed debugging NEW !
- Dumps [**]
- Error handling and recovery Monitors
- Symbolic execution
- Testing tools (e.g., data generators, coverage testing)
- Tracing

D.2.6 Programming Environments

- Graphical environments NEW !
- Integrated environments NEW !
- Interactive environments
- Programmer workbench NEW !

D.2.7 Distribution, Maintenance, and Enhancement Corrections [**]

- Documentation Enhancement [**]
- Extensibility [**]
- Portability
- Restructuring, reverse engineering, and reengineering
- Version control

D.2.8 Metrics (D.4.8)

- Complexity measures
- Performance measures
- Process metrics NEW !
- Product metrics NEW !
- Software science [**]

D.2.9 Management (K.6.3, K.6.4)

- Copyrights [**]
- Cost estimation
- Life cycle
- Productivity
- Programming teams
- Software configuration management
- Software process models (e.g., CMM, ISO, PSP)

- NEW ! Software quality assurance (SQA)
- Time estimation

D.2.10 Design [**] (D.2.2)
- Methodologies [**]
- Representation [**]

D.2.11 Software Architectures (NEW !)
- Data abstraction NEW !
- Domain-specific architectures NEW !
- Information hiding NEW !
- Languages (e.g., description, interconnection, definition) NEW !
- Patterns (e.g., client/server, pipeline, blackboard) NEW !

D.2.12 Interoperability (NEW !)
- Data mapping NEW !
- Distributed objects NEW !
- Interface definition languages NEW !

D.2.13 Reusable Software (NEW !)
- Domain engineering NEW !
- Reusable libraries NEW !
- Reuse models NEW !

D.2.m Miscellaneous
- Rapid prototyping [**]
- Reusable software [**]

D.3 PROGRAMMING LANGUAGES

D.3.0 General Standards

D.3.1 Formal Definitions and Theory (D.2.1, F.3.1, F.3.2, F.4.2, F.4.3)
- Semantics
- Syntax

D.3.2 Language Classifications
- Applicative (functional) languages
- Concurrent, distributed, and parallel languages
- Constraint and logic languages NEW !
- Data-flow languages
- Design languages
- Extensible languages
- Macro and assembly languages
- Microprogramming languages [**]
- Multiparadigm languages NEW !
- Nondeterministic languages [**]

- Nonprocedural languages [**]
- Object-oriented languages
- Specialized application languages
- Very high-level languages

D.3.3 Language Constructs and Features (E.2)

- Abstract data types
- Classes and objects NEW !
- Concurrent programming structures
- Constraints NEW !
- Control structures
- Coroutines
- Data types and structures
- Dynamic storage management
- Frameworks NEW !
- Inheritance NEW !
- Input/output
- Modules, packages
- Patterns NEW !
- Polymorphism NEW !
- Procedures, functions, and subroutines
- Recursion

D.3.4 Processors

- Code generation
- Compilers
- Debuggers NEW !
- Incremental compilers NEW !
- Interpreters
- Memory management (garbage collection) NEW !
- Optimization
- Parsing
- Preprocessors
- Retargetable compilers NEW !
- Run-time environments
- Translator writing systems and compiler generators

D.3.m Miscellaneous

D.4 OPERATING SYSTEMS (C)

D.4.0 General

D.4.1 Process Management

- Concurrency

- Deadlocks
- Multiprocessing/multiprogramming/multitasking
- Mutual exclusion
- Scheduling
- Synchronization
- Threads NEW !

D.4.2 Storage Management

- Allocation/deallocation strategies
- Distributed memories
- Garbage collection NEW !
- Main memory
- Secondary storage
- Segmentation [**]
- Storage hierarchies
- Swapping [**]
- Virtual memory

D.4.3 File Systems Management (E.5)

- Access methods
- Directory structures
- Distributed file systems
- File organization
- Maintenance [**]

D.4.4 Communications Management (C.2)

- Buffering
- Input/output
- Message sending
- Network communication
- Terminal management [**]

D.4.5 Reliability

- Backup procedures
- Checkpoint/restart
- Fault-tolerance
- Verification

D.4.6 Security and Protection (K.6.5)

- Access controls
- Authentication
- Cryptographic controls
- Information flow controls
- Invasive software (e.g., viruses, worms, Trojan horses)
- Security kernels [**]

- Verification [**]

D.4.7 Organization and Design
- Batch processing systems [**]
- Distributed systems
- Hierarchical design [**]
- Interactive systems
- Real-time systems and embedded systems

D.4.8 Performance (C.4, D.2.8, I.6)
- Measurements
- Modeling and prediction
- Monitors
- Operational analysis
- Queueing theory
- Simulation
- Stochastic analysis

D.4.9 Systems Programs and Utilities
- Command and control languages
- Linkers [**]
- Loaders [**]
- Window managers

D.4.m Miscellaneous

D.m MISCELLANEOUS
- Software psychology [**]

E. Data

E.0 GENERAL

E.1 DATA STRUCTURES
- Arrays
- Distributed data structures NEW !
- Graphs and networks
- Lists, stacks, and queues
- Records NEW !
- Tables [**]
- Trees

E.2 DATA STORAGE REPRESENTATIONS
- Composite structures [**]
- Contiguous representations [**]
- Hash-table representations
- Linked representations

- Object representation NEW !
- Primitive data items [**]

E.3 DATA ENCRYPTION

- Code breaking NEW !
- Data encryption standard (DES) [**]
- Public key cryptosystems
- Standards (e.g., DES, PGP, RSA) NEW !

E.4 CODING AND INFORMATION THEORY (H.1.1)

- Data compaction and compression
- Error control codes
- Formal models of communication
- Non secret encoding schemes [**]

E.5 FILES (D.4.3, F.2.2, H.2)

- Backup/recovery
- Optimization [**]
- Organization/structure
- Sorting/searching

E.m MISCELLANEOUS

F. Theory of Computation

F.0 GENERAL

F.1 COMPUTATION BY ABSTRACT DEVICES

F.1.0 General

F.1.1 Models of Computation (F.4.1)

- Automata (e.g., finite, push-down, resource-bounded)
- Bounded-action devices (e.g., Turing machines, random access machines)
- Computability theory
- Relations between models
- Self-modifying machines (e.g., neural networks)
- Unbounded-action devices (e.g., cellular automata, circuits, networks of machines)

F.1.2 Modes of Computation

- Alternation and nondeterminism
- Interactive and reactive computation
- Online computation NEW !
- Parallelism and concurrency
- Probabilistic computation
- Relations among modes [**]

– Relativized computation

F.1.3 Complexity Measures and Classes (F.2)

– Complexity hierarchies
– Machine-independent complexity [**]
– Reducibility and completeness
– Relations among complexity classes
– Relations among complexity measures

F.1.m Miscellaneous

F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY (B.6, B.7, F.1.3)

F.2.0 General

F.2.1 Numerical Algorithms and Problems (G.1, G.4, I.1)

– Computation of transforms (e.g., fast Fourier transform)
– Computations in finite fields
– Computations on matrices
– Computations on polynomials
– Number-theoretic computations (e.g., factoring, primality testing)

F.2.2 Nonnumerical Algorithms and Problems (E.2, E.3, E.4, E.5, G.2, H.2, H.3)

– Complexity of proof procedures
– Computations on discrete structures
– Geometrical problems and computations
– Pattern matching
– Routing and layout
– Sequencing and scheduling
– Sorting and searching

F.2.3 Tradeoffs between Complexity Measures (F.1.3)

F.2.m Miscellaneous

F.3 LOGICS AND MEANINGS OF PROGRAMS

F.3.0 General

F.3.1 Specifying and Verifying and Reasoning about Programs (D.2.1, D.2.4, D.3.1, E.1)

– Assertions
– Invariants
– Logics of programs
– Mechanical verification

- Pre-and post-conditions
- Specification techniques

F.3.2 Semantics of Programming Languages (D.3.1)

- Algebraic approaches to semantics
- Denotational semantics
- Operational semantics
- Partial evaluation NEW!
- Process models NEW!
- Program analysis NEW!

F.3.3 Studies of Program Constructs (D.3.2, D.3.3)

- Control primitives
- Functional constructs
- Object-oriented constructs NEW!
- Program and recursion schemes
- Type structure

F.3.m Miscellaneous

F.4 MATHEMATICAL LOGIC AND FORMAL LANGUAGES

F.4.0 General

F.4.1 Mathematical Logic (F.1.1, I.2.2, I.2.3, I.2.4)

- Computability theory
- Computational logic
- Lambda calculus and related systems
- Logic and constraint programming
- Mechanical theorem proving
- Modal logic NEW!
- Model theory
- Proof theory
- Recursive function theory
- Set theory NEW!
- Temporal logic NEW!

F.4.2 Grammars and Other Rewriting Systems (D.3.1)

- Decision problems
- Grammar types (e.g., context-free, context-sensitive)
- Parallel rewriting systems (e.g., developmental systems, L-systems)
- Parsing Thue systems

F.4.3 Formal Languages (D.3.1)

- Algebraic language theory
- Classes defined by grammars or automata (e.g., context-free languages, regular sets, recursive sets)

- Classes defined by resource-bounded automata [**]
- Decision problems
- Operations on languages

F.4.m Miscellaneous

F.m MISCELLANEOUS

G. Mathematics of Computing

G.0 GENERAL

G.1 NUMERICAL ANALYSIS

G.1.0 General

- Computer arithmetic Conditioning (and ill-conditioning)
- Error analysis
- Interval arithmetic NEW!
- Multiple precision arithmetic NEW!
- Numerical algorithms
- Parallel algorithms Stability (and instability)

G.1.1 Interpolation (I.3.5, I.3.7)

- Difference formulas [**]
- Extrapolation
- Interpolation formulas
- Smoothing
- Spline and piecewise polynomial interpolation

G.1.2 Approximation

- Approximation of surfaces and contours NEW!
- Chebyshev approximation and theory
- Elementary function approximation
- Fast Fourier transforms (FFT) NEW!
- Least squares approximation
- Linear approximation
- Minimax approximation and algorithms
- Nonlinear approximation
- Rational approximation
- Special function approximations NEW!
- Spline and piecewise polynomial approximation
- Wavelets and fractals NEW!

G.1.3 Numerical Linear

- Algebra
- Conditioning Determinants [**]
- Eigenvalues and eigenvectors (direct and iterative methods)

- Error analysis
- Linear systems (direct and iterative methods)
- Matrix inversion
- Pseudoinverses [**]
- Singular value decomposition NEW !
- Sparse, structured, and very large systems (direct and iterative methods)

G.1.4 Quadrature and Numerical Differentiation (F.2.1)

- Adaptive and iterative quadrature
- Automatic differentiation NEW !
- Equal interval integration [**]
- Error analysis
- Finite difference methods
- Gaussian quadrature Iterative methods
- Multidimensional (multiple) quadrature

G.1.5 Roots of Nonlinear Equations

- Continuation (homotopy) methods NEW !
- Convergence
- Error analysis
- Iterative methods
- Polynomials, methods for Systems of equations

G.1.6 Optimization

- Constrained optimization
- Convex programming NEW !
- Global optimization NEW !
- Gradient methods
- nteger programming
- Least squares methods
- Linear programming
- Nonlinear programming
- Quadratic programming methods NEW !
- Simulated annealing NEW !
- Stochastic programming NEW !
- Unconstrained optimization NEW !

G.1.7 Ordinary Differential Equations

- Boundary value problems
- Chaotic systems NEW !
- Convergence and stability
- Differential-algebraic equations NEW !
- Error analysis

- Finite difference methods NEW!
- Initial value problems
- Multistep and multivalue methods
- One-step (single step) methods
- Stiff equations

G.1.8 Partial Differential Equations

- Domain decomposition methods NEW!
- Elliptic equations
- Finite difference methods
- Finite element methods
- Finite volume methods NEW!
- Hyperbolic equations
- Inverse problems NEW!
- Iterative solution techniques NEW!
- Method of lines
- Multigrid and multilevel methods NEW!
- Parabolic equations
- Spectral methods NEW!

G.1.9 Integral Equations

- Delay equations NEW!
- Fredholm equations
- Integro-differential equations
- Volterra equations

G.1.10 Applications NEW!

G.1.m Miscellaneous

G.2 DISCRETE MATHEMATICS

G.2.0 General

G.2.1 Combinatorics (F.2.2)

- Combinatorial algorithms
- Counting problems
- Generating functions
- Permutations and combinations
- Recurrences and difference equation

G.2.2 Graph Theory (F.2.2)

- Graph algorithms
- Graph labeling NEW!
- Hypergraphs NEW!
- Network problems

- Path and circuit problems
- Trees

G.2.3 Applications (NEW!)

G.2.m Miscellaneous

G.3 PROBABILITY AND STATISTICS

- Contingency table analysis NEW!
- Correlation and regression analysis NEW!
- Distribution functions NEW!
- Experimental design NEW!
- Markov processes NEW!
- Multivariate statistics NEW!
- Nonparametric statistics NEW!
- Probabilistic algorithms (including Monte Carlo)
- Queueing theory NEW!
- Random number generation
- Reliability and life testing NEW!
- Renewal theory NEW!
- Robust regression NEW!
- Statistical computing
- Statistical software
- Stochastic processes NEW!
- Survival analysis NEW!
- Time series analysis NEW!

G.4 MATHEMATICAL SOFTWARE

- Algorithm design and analysis
- Certification and testing
- Documentation NEW!
- Efficiency
- Parallel and vector implementations NEW!
- Portability [**]
- Reliability and robustness
- User interfaces NEW!
- Verification [**]

G.m MISCELLANEOUS

- Queueing theory [**]

H. Information Systems

H.0 GENERAL

H.1 MODELS AND PRINCIPLES

H.1.0 General

H.1.1 Systems and Information Theory (E.4)
- General systems theory
- Information theory
- Value of information

H.1.2 User/Machine
- Systems
- Human factors
- Human information processing
- Software psychology NEW

H.1.m Miscellaneous

H.2 DATABASE MANAGEMENT (E.5)

H.2.0 General Security, integrity, and protection [**]

H.2.1 Logical Design
- Data models
- Normal forms
- Schema and subschema

H.2.2 Physical Design
- Access methods
- Deadlock avoidance
- Recovery and restart

H.2.3 Languages (D.3.2)
- Data description languages (DDL)
- Data manipulation languages (DML)
- Database (persistent) programming languages
- Query languages
- Report writers

H.2.4 Systems Concurrency
- Distributed databases
- Multimedia databases NEW!
- Object-oriented databases NEW!
- Parallel databases NEW!
- Query processing
- Relational databases NEW!
- Rule-based databases NEW!
- Textual databases NEW!
- Transaction processing

H.2.5 Heterogeneous Databases

- Data translation [**]
- Program translation [**]

H.2.6 Database Machines

H.2.7 Database Administration
- Data dictionary/directory
- Data warehouse and repository NEW !
- Logging and recovery Security, integrity, and protection NEW !

H.2.8 Database Applications
- Data mining NEW !
- Image databases NEW !
- Scientific databases NEW !
- Spatial databases and GIS NEW !
- Statistical databases NEW !

H.2.m Miscellaneous

H.3 INFORMATION STORAGE AND RETRIEVAL

H.3.0 General

H.3.1 Content Analysis and Indexing
- Abstracting methods
- Dictionaries
- Indexing methods
- Linguistic processing
- Thesauruses

H.3.2 Information Storage
- File organization
- Record classification [**]

H.3.3 Information Search and Retrieval
- Clustering Information filtering NEW !
- Query formulatio
- Relevance feedback NEW !
- Retrieval models
- Search process
- Selection process

H.3.4 Systems and Software
- Current awareness systems (selective dissemination of information–SDI) [**]
- Distributed systems NEW !
- Information networks
- Performance evaluation (efficiency and effectiveness) NEW !

- Question-answering (fact retrieval) systems [**]
- User profiles and alert services NEW !

H.3.5 Online Information Services

- Commercial services NEW !
- Data sharing
- Web-based services NEW !

H.3.6 Library Automation

- Large text archives

H.3.7 Digital Libraries (NEW !)

- Collection NEW !
- Dissemination NEW !
- Standards NEW !
- Systems issues NEW !
- User issues NEW !

H.3.m Miscellaneous

H.4 INFORMATION SYSTEMS APPLICATIONS

H.4.0 General

H.4.1 Office Automation (I.7)

- Desktop publishing NEW !
- Equipment [**]
- Groupware NEW !
- Spreadsheets
- Time management (e.g., calendars, schedules)
- Word processing
- Workflow management NEW !

H.4.2 Types of Systems

- Decision support (e.g., MIS)
- Logistics

H.4.3 Communications Applications

- Bulletin boards
- Computer conferencing, teleconferencing, and videoconferencing
- Electronic mail Information browsers NEW !
- Videotex

H.4.m Miscellaneous

H.5 INFORMATION INTERFACES AND PRESENTATION (e.g., HCI) (I.7)

H.5.0 General

H.5.1 Multimedia

- Information Systems
- Animations
- Artificial, augmented, and virtual realities
- Audio input/output
- Evaluation/methodology
- Hypertext navigation and maps [**]
- Video (e.g., tape, disk, DVI)

H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6)

- Auditory (non-speech) feedback NEW!
- Benchmarking NEW!
- Ergonomics
- Evaluation/methodology
- Graphical user interfaces (GUI) NEW!
- Haptic I/O NEW!
- Input devices and strategies (e.g., mouse, touchscreen)
- Interaction styles (e.g., commands, menus, forms, direct manipulation)
- Natural language NEW!
- Prototyping NEW!
- Screen design (e.g., text, graphics, color)
- Standardization NEW!
- Style guides NEW!
- Theory and methods
- Training, help, and documentation
- User-centered design NEW!
- User interface management systems (UIMS)
- Voice I/O NEW!
- Windowing systems

H.5.3 Group and Organization Interfaces

- Asynchronous interaction
- Collaborative computing NEW!
- Computer-supported cooperative work NEW!
- Evaluation/methodology
- Organizational design
- Synchronous interaction
- Theory and models
- Web-based interaction

H.5.4 Hypertext/Hypermedia (NEW!) (I.7, J.7)

- Architectures NEW!
- Navigation NEW!

- Theory NEW!
- User issues NEW!

H.5.5 Sound and Music Computing (NEW!) (J.5)

- Methodologies and techniques NEW!
- Modeling NEW!
- Signal analysis, synthesis, and processing NEW!
- Systems NEW!

H.5.m Miscellaneous (NEW!)

H.m MISCELLANEOUS

I. Computing Methodologies

I.0 GENERAL

I.1 SYMBOLIC AND ALGEBRAIC MANIPULATION

I.1.0 General

I.1.1 Expressions and Their Representation (E.1, E.2)

- Representations (general and polynomial)
- Simplification of expressions

I.1.2 Algorithms (F.2.1, F.2.2)

- Algebraic algorithms
- Analysis of algorithms
- Nonalgebraic algorithms

I.1.3 Languages and Systems (D.3.2, D.3.3, F.2.2)

- Evaluation strategies
- Nonprocedural languages [**]
- Special-purpose algebraic systems
- Special-purpose hardware [**]
- Substitution mechanisms [**]

I.1.4 Applications

I.1.m Miscellaneous

I.2 ARTIFICIAL INTELLIGENCE

I.2.0 General

- Cognitive simulation
- Philosophical foundations

I.2.1 Applications and Expert Systems (H.4, J)

- Cartography
- Games
- Industrial automation

- Law
- Medicine and science
- Natural language interfaces
- Office automation

I.2.2 Automatic Programming (D.1.2, F.3.1, F.4.1)

- Automatic analysis of algorithms
- Program modification
- Program synthesis
- Program transformation
- Program verification

I.2.3 Deduction and Theorem Proving (F.4.1)

- Answer/reason extraction
- Deduction (e.g., natural, rule-based)
- Inference engines NEW !
- Logic programming
- Mathematical induction
- Metatheory [**]
- Nonmonotonic reasoning and belief revision
- Resolution
- Uncertainty, "fuzzy" and probabilistic reasoning

I.2.4 Knowledge Representation

- Formalisms and Methods (F.4.1)
- Frames and scripts
- Modal logic NEW !
- Predicate logic
- Relation systems
- Representation languages
- Representations (procedural and rule-based)
- Semantic networks
- Temporal logic NEW !

I.2.5 Programming Languages and Software (D.3.2)

- Expert system tools and techniques

I.2.6 Learning (K.3.2)

- Analogies
- Concept learning
- Connectionism and neural nets
- Induction
- Knowledge acquisition
- Language acquisition
- Parameter learning

I.2.7 Natural Language Processing
- Discourse
- Language generation
- Language models
- Language parsing and understanding
- Machine translation
- Speech recognition and synthesis
- Text analysis

I.2.8 Problem Solving, Control Methods, and Search (F.2.2)
- Backtracking
- Control theory NEW !
- Dynamic programming
- Graph and tree search strategies
- Heuristic methods
- Plan execution, formation, and generation
- Scheduling NEW !

I.2.9 Robotics
- Autonomous vehicles NEW !
- Commercial robots and applications NEW !
- Kinematics and dynamics NEW !
- Manipulators
- Operator interfaces NEW !
- Propelling mechanisms
- Sensors
- Workcell organization and planning NEW !

I.2.10 Vision and Scene Understanding (I.4.8, I.5)
- 3D/stereo scene analysis NEW !
- Architecture and control structures [**]
- Intensity, color, photometry, and thresholding
- Modeling and recovery of physical attributes
- Motion
- Perceptual reasoning
- Representations, data structures, and transforms
- Shape
- Texture
- Video analysis NEW !

I.2.11 Distributed Artificial Intelligence
- Coherence and coordination
- Intelligent agents NEW !
- Languages and structures

- Multiagent systems NEW !

I.2.m Miscellaneous

I.3 COMPUTER GRAPHICS

I.3.0 General

I.3.1 Hardware Architecture (B.4.2)

- Graphics processors
- Hardcopy devices [**]
- Input devices
- Parallel processing
- Raster display device
- Storage devices [**]
- Three-dimensional displays [**]
- Vector display devices [**]

I.3.2 Graphics Systems (C.2.1, C.2.4, C.3)

- Distributed/network graphics
- Remote systems [**]
- Stand-alone systems [**]

I.3.3 Picture/Image

- Generation
- Antialiasing [**]
- Bitmap and framebuffer operations
- Digitizing and scanning
- Display algorithms
- Line and curve generation
- Viewing algorithms

I.3.4 Graphics Utilities

- Application packages
- Device drivers [**]
- Graphics editors
- Graphics packages
- Meta files [**]
- Paint systems
- Picture description languages [**]
- Software support
- Virtual device interfaces

I.3.5 Computational Geometry and Object Modeling

- Boundary representations
- Constructive solid geometry (CSG) [**]

- Curve, surface, solid, and object representations
- Geometric algorithms, languages, and systems
- Hierarchy and geometric transformations
- Modeling packages
- Object hierarchies
- Physically based modeling
- Splines

I.3.6 Methodology and Techniques

- Device independence [**]
- Ergonomics
- Graphics data structures and data types
- Interaction techniques
- Languages Standards

I.3.7 Three-Dimensional Graphics and Realism

- Animation
- Color, shading, shadowing, and texture
- Fractals
- Hidden line/surface removal
- Radiosity
- Raytracing
- Virtual reality
- Visible line/surface algorithms

I.3.8 Applications

I.3.m Miscellaneous

I.4 IMAGE PROCESSING AND COMPUTER VISION

I.4.0 General

- Image displays
- Image processing software

I.4.1 Digitization and Image

- Capture
- Camera calibration NEW!
- Imaging geometry NEW!
- Quantization
- Radiometry NEW!
- Reflectance NEW!
- Sampling
- Scanning

I.4.2 Compression (Coding) (E.4)

- Approximate methods
- Exact coding [**]

I.4.3 Enhancement

- Filtering
- Geometric correction
- Grayscale manipulation
- Registration
- Sharpening and deblurring [**]
- Smoothing

I.4.4 Restoration

- Inverse filtering [**]
- Kalman filtering
- Pseudoinverse restoration [**]
- Wiener filtering [**]

I.4.5 Reconstruction

- Series expansion methods
- Summation methods [**]
- Transform methods

I.4.6 Segmentation

- Edge and feature detection
- Pixel classification
- Region growing, partitioning
- Relaxation NEW !

I.4.7 Feature Measurement

- Feature representation NEW !
- Invariants
- Moments
- Projections
- Size and shape
- Texture

I.4.8 Scene Analysis Color NEW !

- Depth cues
- Motion NEW !
- Object recognition NEW !
- Photometry
- Range data
- Sensor fusion
- Shading NEW !
- Shape NEW !

- Stereo Surface fitting NEW !
- Time-varying imagery
- Tracking NEW !

I.4.9 Applications

I.4.10 Image Representation
- Hierarchical
- Morphological
- Multidimensional
- Statistical
- Volumetric

I.4.m Miscellaneous

I.5 PATTERN RECOGNITION

I.5.0 General

I.5.1 Models Deterministic [**]
- Fuzzy set
- Geometric
- Neural nets
- Statistical
- Structural

I.5.2 Design
- Methodology
- Classifier design and evaluation
- Feature evaluation and selection
- Pattern analysis

I.5.3 Clustering

Algorithms

Similarity measures

I.5.4 Application
- Computer vision
- Signal processing
- Text processing
- Waveform analysis

I.5.5 Implementation (C.3)
- Interactive systems
- Special architectures

I.5.m Miscellaneous

I.6 SIMULATION AND MODELING (G.3)

I.6.0 General

I.6.1 Simulation
- Theory
- Model classification
- Systems theory
- Types of simulation (continuous and discrete) [*] [1]

I.6.2 Simulation Languages

I.6.3 Applications

I.6.4 Model Validation and Analysis

I.6.5 Model Development
- Modeling methodologies

I.6.6 Simulation Output Analysis

I.6.7 Simulation Support
- Systems
- Environments

I.6.8 Types of Simulation
- Animation
- Combined
- Continuous
- Discrete event
- Distributed
- Gaming
- Monte Carlo
- Parallel Visual

I.6.m Miscellaneous

I.7 DOCUMENT AND TEXT PROCESSING (H.4, H.5)

I.7.0 General

I.7.1 Document and Text Editing
- Document management NEW !
- Languages [**]
- Spelling [**]
- Version control NEW !

I.7.2 Document Preparation
- Desktop publishing
- Format and notation
- Hypertext/hypermedia

1. Le signe [*] indique que l'élément n'est plus utilisé pour la classification depuis janvier 1991, mais qu'il est toujours possible de l'utiliser pour rechercher des documents indexés avant cette période.

- Index generation NEW!
- Languages and systems
- Markup languages NEW!
- Multi/mixed media
- Photocomposition/typesetting
- Scripting languages NEW!
- Standards

I.7.3 Index Generation [**]

I.7.4 Electronic Publishing NEW! (H.5.4, J.7)

I.7.5 Document Capture NEW!(I.4.1)
- Document analysis NEW!
- Graphics recognition and interpretation NEW!
- Optical character recognition (OCR) NEW!
- Scanning NEW!

I.7.m Miscellaneous

I.m MISCELLANEOUS

J. Computer Applications

J.0 GENERAL

J.1 ADMINISTRATIVE DATA PROCESSING
- Business
- Education
- Financial (e.g., EFTS)
- Government
- Law
- Manufacturing
- Marketing
- Military

J.2 PHYSICAL SCIENCES AND ENGINEERING
- Aerospace
- Archaeology NEW!
- Astronomy
- Chemistry
- Earth and atmospheric sciences
- Electronics
- Engineering
- Mathematics and statistics
- Physics

J.3 LIFE AND MEDICAL SCIENCES
- Biology and genetics

- Health
- Medical information systems

J.4 SOCIAL AND BEHAVIORAL SCIENCES
- Economics
- Psychology
- Sociology

J.5 ARTS AND HUMANITIES
- Architecture NEW!
- Arts, fine and performing [**]
- Fine arts NEW!
- Language translation
- Linguistics
- Literature
- Music [**]
- Performing arts (e.g., dance, music) NEW!

J.6 COMPUTER-AIDED ENGINEERING
- Computer-aided design (CAD)
- Computer-aided manufacturing (CAM)

J.7 COMPUTERS IN OTHER SYSTEMS (C.3)
- Command and control
- Consumer products
- Industrial control
- Military
- Process control
- Publishing
- Real time

J.m MISCELLANEOUS

K. Computing Milieux

K.0 GENERAL

K.1 THE COMPUTER INDUSTRY
- Markets
- Standards
- Statistics
- Suppliers

K.2 HISTORY OF COMPUTING
- Hardware
- People
- Software
- Systems

- Theory

K.3 COMPUTERS AND EDUCATION

K.3.0 General

K.3.1 Computer Uses in Education
- Collaborative learning NEW !
- Computer-assisted instruction (CAI)
- Computer-managed instruction (CMI)
- Distance learning NEW !

K.3.2 Computer and Information
- Science Education
- Accreditation NEW !
- Computer science education
- Curriculum
- Information systems education
- Literacy NEW !
- Self-assessment

K.3.m Miscellaneous
- Accreditation [**]
- Computer literacy [**]

K.4 COMPUTERS AND SOCIETY

K.4.0 General

K.4.1 Public
- Policy Issues
- Abuse and crime involving computers NEW !
- Computer-related health issues NEW !
- Ethics NEW !
- Human safety
- Intellectual property rights NEW !
- Privacy
- Regulation
- Transborder data flow
- Use/abuse of power NEW !

K.4.2 Social Issues
- Abuse and crime involving computers [**]
- Assistive technologies for persons with disabilities NEW !
- Employment
- Handicapped persons/special needs [**]

K.4.3 Organizational
- Impacts

- Automation NEW!
- Computer-supported collaborative work NEW!
- Employment NEW!
- Reengineering NEW!

K.4.4 Electronic Commerce (NEW!) (J.1)

- Cybercash, digital cash NEW!
- Distributed commercial transactions NEW!
- Electronic data interchange (EDI) NEW!
- Intellectual property NEW!
- Payment schemes NEW!
- Security NEW!

K.4.m Miscellaneous

K.5 LEGAL ASPECTS OF COMPUTING

K.5.0 General

K.5.1 Hardware/Software

- Protection
- Copyrights
- Licensing NEW!
- Patents
- Proprietary rights
- Trade secrets [**]

K.5.2 Governmental Issues

- Censorship NEW!
- Regulation
- Taxation

K.5.m Miscellaneous

- Contracts [**]
- Hardware patents [**]

K.6 MANAGEMENT OF COMPUTING AND INFORMATION SYSTEMS

K.6.0 General Economics

K.6.1 Project and People

- Management
- Life cycle
- Management techniques (e.g., PERT/CPM)
- Staffing
- Strategic information systems planning NEW!
- Systems analysis and design
- Systems development

- Training

K.6.2 Installation

- Management
- Benchmarks
- Computer selection
- Computing equipment management
- Performance and usage measurement
- Pricing and resource allocation

K.6.3 Software Management (D.2.9)

- Software development
- Software maintenance
- Software process NEW !
- Software selection

K.6.4 System Management

- Centralization/decentralization
- Management audit
- Quality assurance

K.6.5 Security and Protection (D.4.6, K.4.2)

- Authentication Insurance [**]
- Invasive software (e.g., viruses, worms, Trojan horses)
- Physical security [**]
- Unauthorized access (e.g., hacking, phreaking) NEW !

K.6.m Miscellaneous

- Insurance [*]
- Security [*]

K.7 THE COMPUTING PROFESSION

K.7.0 General

K.7.1 Occupations

K.7.2 Organizations

K.7.3 Testing, Certification, and Licensing

K.7.4 Professional Ethics (NEW !) (K.4)

- Codes of ethics NEW !
- Codes of good practice NEW !
- Ethical dilemmas NEW !

K.7.m Miscellaneous

- Codes of good practice [**]
- Ethics [**]

K.8 PERSONAL COMPUTING

- Games [*]

K.8.0 General Games

K.8.1 Application Packages

- Data communications
- Database processing
- Freeware/shareware NEW !
- Graphics
- Spreadsheets
- Word processing

K.8.2 Hardware

K.8.3 Management/Maintenance

K.8.m Miscellaneous (NEW !)

K.m MISCELLANEOUS

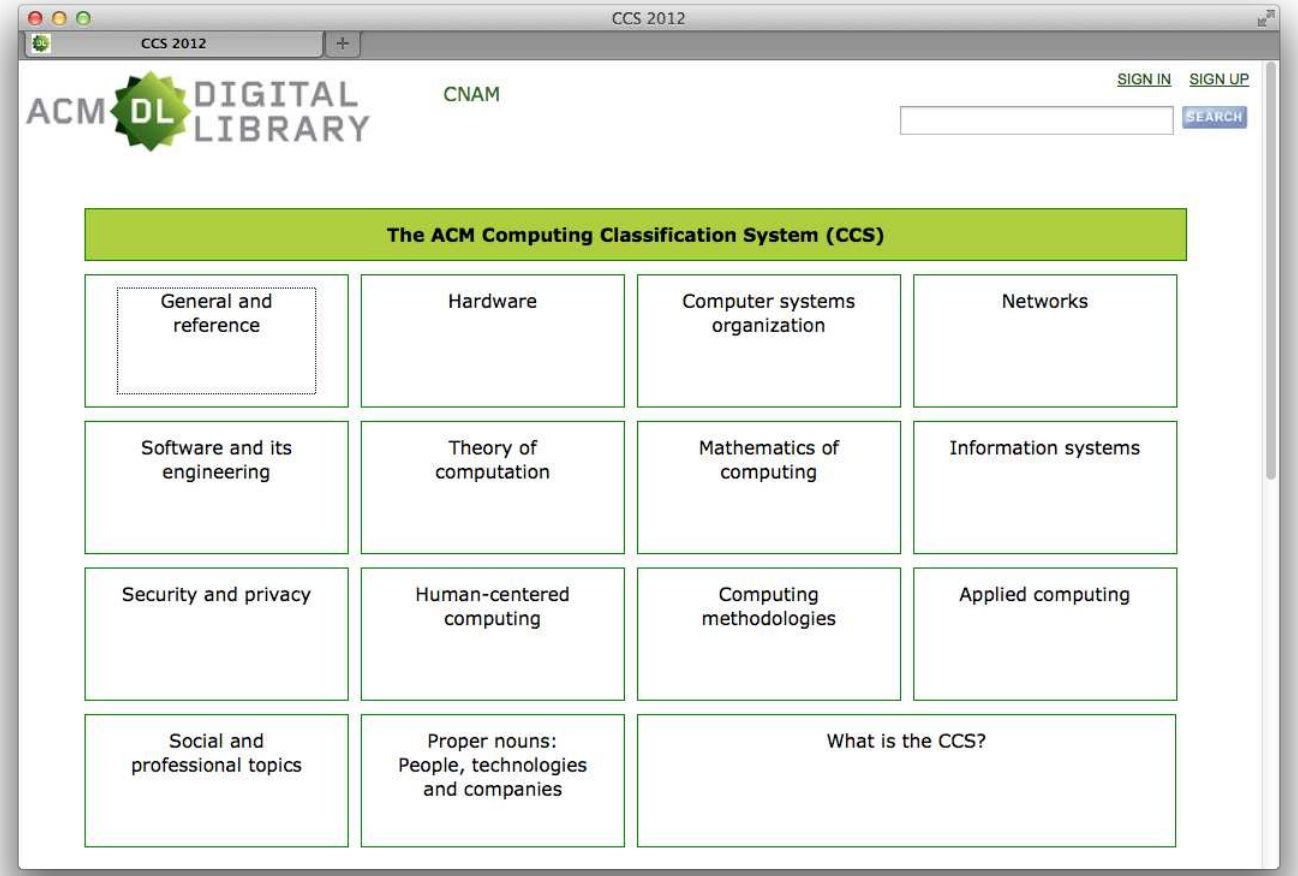


Figure E.1: Nouvelle version de navigation« tabulaire » de l'ACM CCS, *publiée en ligne le 22 septembre 2012*

E.2 Nouvelle interface graphique d'ACM

Annexe F

A propos de DBLP

Dans cette annexe nous présentons le détail de nos réflexions autour de la base de connaissance DBLP. Cette partie aurait été fastidieuse à lire dans le cadre d'une thèse en sciences de l'information et de la communication. Cependant, ces réflexions méthodologiques et techniques peuvent intéresser les personnes ayant à réaliser de volumineux exports de données.

Choix motivé de DBLP

Peu de bases de connaissances scientifiques exposent gratuitement l'ensemble de leurs données au moissonnage massif. Cependant, un projet de qualité recense la plupart des publications scientifiques en informatique scientifique. Il s'agit du projet DBLP initié par Michael Ley de l'Université de Trèves (Trier) en Allemagne. Le suivi, le sérieux de ce travail mirent en avant le projet sur la scène internationale. En 2003, M. Ley a reçu le prix de contribution ACM SIGMOD (groupe d'intérêt en gestion de données [1]) pour ses travaux.

Depuis, la base DBLP est une référence incontournable dans le monde de l'informatique scientifique. L'ACM a même répliqué le site DBLP sur le site de SIGMOD. Cette base propose un SRI avec moteur classique et une navigation avec recherche à facettes au travers du projet *Faceted DBLP* [2] de Diederich *et al.* (2007).

1. `http://www.sigmod.org/sigmod-awards/award-people/michael-ley`
2. `http://dblp.l3s.de/`

DBLP offre également un accès exploratoire par conférence . Le grand intérêt de cette base est que M. Ley propose l'intégralité des 1 631 850 notices dans un seul fichier XML. Il n'est pas possible de se lancer dans l'utilisation d'une telle masse de données en faisant l'économie d'une étude de son historique et de ses évolutions techniques. En effet, l'historique de ce projet est étalé sur deux décennies. Son succès lui a valu d'être plusieurs fois répliqué, et ce dans plusieurs contextes d'usages avec différentes IHM.

Historique et spécificités de DBLP

La bibliothèque DBLP a évolué à partir d'un embryon de petit serveur Web expérimental en 1993 à un service populaire pour la communauté informatique (sur une vingtaine d'années). DBLP a été mis en œuvre en 1993 comme une petite collection de fichiers HTML qui ont été entrés manuellement depuis un éditeur de texte standard. Très vite, les notices bibliographiques ont été extraites des fichiers HTML. Puis ce projet fut publié en 1997.

En 2006, le travail de M. Ley fut présenté à la conférence internationale francophone Extraction et Gestion de Connaissances (EGC [1]). Cependant, beaucoup de décisions de conception et les détails des fichiers XML composants DBLP n'ont jamais été documentés.

En 2009, M. Ley a écrit un article officiel pour pallier au déficit de documentation relatif à DBLP. La même année, plusieurs appendices techniques décrivant des aspects particuliers ont été également publiés Ley (2009a,b). Le contenu de cette base est accessible par des fichiers au format XML avec un fichier de Définition de Type de Document (DTD) Ley (2001). Ce fichier permet de décrire un modèle de document XML. Il s'agit à la fois d'une grammaire qui décrit l'imbrication des entités (balises) les unes par rapport aux autres, mais aussi les attributs possibles de chaque balise. Un fichier de normalisation typologique XML (DTD [2]) permet de spécifier l'ordre, le nombre d'occurrences et le type des entités et attributs.

1. `http://www.egc.asso.fr/`
2. Un fichier *Document Type Definition* (DTD) permet donc de valider un document XML en imposant une syntaxe (combinaison d'éléments) et une grammaire descriptive strictes.

Il est indispensable dans le cadre de la base de données XML de posséder une DTD qui va être le gage de formalisme du document. En effet, si le document est mal formé, un outil de parcours de document XML (un simple navigateur) ne sera pas « autorisé » à l'ouvrir et affichera un message d'erreur.

Interfaces de consultation de DBLP

Différentes aides à la recherche sont proposées sur l'interface. Voici un aperçu des avantages offerts par une bonne exposition des métadonnées associée aux technologies récentes de recherche d'informations :

- La recherche rapide. Une recherche est déclenchée, puis affinée, après chaque frappe, avec un temps de réponse immédiat.
- Les options de recherche par groupe d'intérêt. Un préfixe de recherche peut être indiqué en complément de la recherche. Il est ainsi possible de préfixer sa requête avec « SIG : ». Cela signifie que l'on ne cherche que des articles parus dans les prestigieuses conférences « *Special Interest Group* » d'ACM, comme SIGIR.
- Le portail offre également une tolérance à l'orthographe et une syntaxe de recherche. Par exemple, pour un mot exact : il faut saisir le caractère dollar en fin de mot. Les opérateurs booléens sont de mise avec le caractère pipe (tuyau) qui signifie OR et le signe moins devant un mot signifie NOT. Un point entre les mots, par exemple *information.retrieval*, équivaut à un encadrement par guillemets pour une recherche d'expression.
- La recherche à facette. La barre de droite offre une ventilation du résultat courant par catégories.
- recherche sur métadonnée. Il est possible de préciser le nombre d'auteurs « na :1 » ou spécifier la place d'un auteur dans une publication (s'il est premier auteur par exemple) avec « prenomnom :1 ».

Notre travail d'intégration et de mise à jour de DBLP

Comme notre projet de recherche a été amorcé en 2006, nous avons pu suivre les évolutions et problématiques de M. Ley de manière synchrone. Notre première intégration du fichier principal DBLP dans une base de données MySQL date de 2007.

La partie de l'annexe expose dans un premier temps la modélisation de notre base de connaissances. Ensuite, nous proposons la chronologie des différences méthodes utilisées pour parcourir un aussi grand fichier XML et l'intégrer dans une base MySQL. Préambule : méthodologie de parcours de fichiers XML. L'extraction des données d'un document XML nécessite un analyseur syntaxique ou parseur (*parser*). Cet outil permet de parcourir le document et d'en extraire les informations qu'il contient. Les données évaluées de manière événementielle ou séquentielle sont retournées brutes ou sous forme d'objets. Nous distinguons deux approches technologiques pour parcourir les gros fichiers XML. La première approche concerne la validité du document XML :

1. Les parseurs validant qui permettent de vérifier qu'un document XML est conforme à sa DTD ;
2. Les parseurs non validant (non-validating) se contentent de vérifier que le document XML est bien formé (c'est-à-dire respectant la syntaxe XML de base).

Nous savons que le document que nous devons parcourir est valide. A priori une méthode validante ne peut pas apporter de plus value à notre processus. Dès 2006, Grün (2006) se proposait de parcourir de vastes fichiers XML.
Les parseurs XML sont également divisés selon l'approche qu'ils utilisent pour traiter le document :

1. Les API utilisant une approche hiérarchique : les analyseurs utilisant cette technique construisent une structure hiérarchique contenant des objets représentant les éléments du document, et dont les méthodes permettent d'accéder aux propriétés. La principale API utilisant cette approche est DOM (Document Object Model) dont les spécifications sont référencées par le WC3, (Le Hégaret *et al.*, 2005).
2. Les API basées sur un mode événementiel (ou séquentiel) permettent de réagir à des événements (comme le début d'un élément, la fin d'un élément) et de renvoyer le résultat à l'application utilisant cette API. SAX (Simple API for XML) est la principale interface utilisant l'aspect événementiel. Contrairement à DOM, SAX n'est pas issu d'une réflexion commune à un groupe, mais l'œuvre d'un individu isolé, Megginson (1998). SAX a ensuite été repris par saxproject et intégré au sein de xml.org par son fondateur pour devenir SAX 1.0 puis SAX 2.0, Saxproject (1998, 2001).

Ainsi, on tend à associer l'approche hiérarchique avec DOM et l'approche événementielle avec SAX. L'utilisation d'objets stockés en mémoire est séduisante. Cette méthode pourrait permettre de manipuler facilement et rapidement les notices bibliographiques pour les intégrer à une base de données.
En 2007, Pan *et al.* (2007) avaient proposé une méthode pour parcourir un gros fichier XML en DOM par parallélisation. Cette méthode appelée *Parallel Algorithm for XML DOM Parsing* (PXP). En 2009, Shah *et al.* (2009) ont proposé une approche basée sur un traitement DOM parallèle baptisée ParDOM. Ces deux équipes ont utilisé leur technologie sur le fichier DBLP.xml comme corpus de test. Les deux équipes décrivaient de la même façon le problème posé par ce type travaux. Cette opération est connue pour causer des goulets d'étranglement des performances dans les applications et les systèmes qui traitent de gros volumes de données XML. Le parallélisme est une façon naturelle d'améliorer les performances en segmentant une tâche répétitives en plusieurs tâches exécutées de matière simultanées sur plusieurs processeurs, machines ou un processeur multicœurs. S'appuyer sur des processeurs multicœurs peut offrir une solution rentable, car à l'avenir des processeurs multicœurs prendront en charge des centaines de cœurs dans les *clusters* de calcul, ce qui offrira un degré élevé de parallélisme matériel. ParDOM offre un meilleur traitement du parallélisme que PXP Ce système adopte un système de *chunking* flexible : Chaque bloc peut contenir un nombre arbitraire de début et de fin des balises XML qui sert d'identifiant. ParDOM peut donc être facilement mise en œuvre par programmation parallèle, notamment sur des ordinateurs à processeurs multicœurs. ParDOM comporte deux phases :

1. dans la première phase, un document XML est partitionné en morceaux qui sont analysés en parallèle ;
2. dans la seconde phase, les différentes structures partielles d'arbre DOM créé lors de la première phase, sont reliées (en parallèle) pour construire une arborescence complète DOM.

Nous allons, dans un premier temps, nous inspirer de la méthode ParDOM pour le *chunking* flexible. Mais au lieu de reconstruire un objet, nous allons faire une intégration dans une base de données MySQL.

Première approche : exploitation du fichier XML par script

Développement d'une base locale en DOM

Lors de notre découverte du travail de M. Ley, nous avons pensé construire la base de données bibliographique en parsant le fichier XML d'un bloc par un script PHP en ligne de commande. La modélisation initiale de ce processus s'organisait autour de la bibliothèque de programmation PHP (ou librairie) DOM simpleXML. Cette librairie avait la particularité d'une manipulation aisée basée sur un modèle objet. La simplicité de manipulation des objets au sein du fichier XML était malheureusement contrebalancée par deux fortes contraintes. SimpleXML est une API validante, ce qui signifie que seuls les fichiers « bien formés » peuvent être parcourus. M. Ley a expliqué que pour gérer l'accentuation des noms propres, il s'est contraint à encoder les caractères accentués en HMTL. Cette décision est intégrée dans la DTD accompagnant le fichier XML. Le fichier est donc en théorie valide pour l'affichage.

Cependant, l'affichage et le parcours par une API sont deux choses différentes. L'entête du fichier XML annonçant un encodage ISO 8859-1, la librairie SimpleXML refusait de parcourir un fichier qui propose des caractères HTML comme « **é** » pour le caractère « **é** ». Cela signifie implicitement que le fichier DBLP n'est pas conforme à l'encodage mentionné dans sa DTD.

Pour résoudre le problème d'encodage, une première solution native XML est basée la balise **CDATA**. Le XML propose une manière souple de gérer ce type de problème tant pour l'affichage que pour la compatibilité avec les API validantes. Il suffit d'encadrer les noms propres, au sein de balise **<author>** source de conflit, dans une balise **<![CDATA[...]]>** pour régler ce problème de caractères accentués. Cependant, si cette solution normalise le fichier XML, elle gène la manipulation des balises impactées.

Cette solution n'ayant pas retenu notre attention car nous souhaitons une manipulation aisée des éléments bibliographiques. Nous avons adapté notre démarche et avons écrit un script l'outil Linux de manipulation de texte ***sed*** pour rendre le fichier DBLP réellement « valide » en remplaçant les caractères au format HTML par leur équivalant en texte unicode (UTF-8). Cette méthode avait pour avantage de résoudre le problème simplement. Le temps de calcul nécessaire à cet encodage se compte cependant en minutes. L'exécution du script occupe la machine hôte de manière quasi systématique, tant du point de vue de la mémoire que du processeur. Cette solution permet donc de

fournir un fichier XML valide avec des caractères accentués correspondants à l'encodage ISO 8859-1 déclaré dans l'en-tête.

L'autre problème posé par le fichier XML est sa taille. Il est difficile de parser un fichier XML de plus de 800 Mo avec une API DOM. En effet, l'ensemble du fichier est chargé entièrement en mémoire avant de séparer chaque entrée en objet. Les tests effectués sous Mac OS X avec 4 Go de RAM ont montré que seuls des fichiers de taille inférieure à 100 mégas pouvaient être parcourus de manière efficace en DOM/simpleXML. Enfin, les mises à jour hebdomadaires de DBLP ne sont pas incrémentales, un différentiel entre deux versions du fichier ne peut être généré avec une commande Linux « ***diff*** ».

Nous avons écrit un script Shell autour des commandes Linux « ***tail*** et « ***split***. Ce traitement par lot se charge de transformer le fichier XML en n fichiers de taille moindre. Le fichier initial étant coupé avec la commande « ***split***, le découpage en objets ne pouvait pas être respecté. C'est ici que la commande ***tail*** associée à une sélection ***grep*** grâce à un tuyau Linux (***pipe***). Enfin, l'usage d'une commande ***cut*** a permis de déplacer les objets tronqués d'un fichier à un autre. En ajoutant une entête et une fin XML à chaque fichier, nous obtenons n fichiers DBLP bien formés de taille moindre. Ces fichiers alors peuvent alors être parcourus en traitement par lots.

Pour ce qui est de la partie mise à jour, nous n'avions pas trouvé d'autre solution que d'industrialiser le processus décrit jusqu'ici afin de générer un fichier ***SQL*** unique qui était injecté en ligne de commande dans la base de données ***MySQL***. Le processus était lancé chaque semaine grâce au planificateur Linux ***cron***.
Cette première approche, bien qu'efficace, était assez peu satisfaisante. En effet, l'exécution séquentielle de ce laborieux processus basé uniquement sur des scripts Shell et PHP prenait trop de temps. Nous n'avions pas accès à un serveur dédié, et étions obligés de brider l'utilisation de mémoire vive par PHP pour stabiliser le serveur.

Seconde approche de traitement local Dans un second temps, nous avons décidé d'accélérer le traitement en nous appuyant sur une technologie Java. Dans ce langage, les principales APIs DOM et SAX sont disponibles.

- Java DOM : L'approche DOM consiste à créer une arborescence d'objets qui représente l'organisation des données dans le document. Cette arborescence est enregistrée en mémoire. L'application peut alors accéder à l'information voulue pour l'utiliser ou même pour la modifier.

– Java SAX : L'approche SAX consiste à parcourir le fichier XML et passer chaque partie du fichier à l'application. L'application peut alors utiliser les informations reçues par le parseur, mais elle ne peut pas les modifier, car rien n'est enregistré dans la mémoire.

Nous avons pensé dédier le processus de parcours du fichier DBLP.xml à une machine externe au serveur web hébergeant l'application. En abordant notre problématique sous un angle de programme « compilé », sur une machine dédiée, plus puissante et dotée du double de mémoire vive, nous étions confiants sur notre capacité à parser le fichier en DOM d'une seule traite.

Nous avons lancé le processus en allouant trois gigaoctets à la JVM. Nous conservions un gigaoctet pour le système d'exploitation. Cependant, même si le résultat était spectaculairement plus rapide pour l'analyse de fichiers dont la taille ne dépassait pas les 200 mégaoctets, notre programme ne pouvait pas charger l'intégralité du fichier XML en mémoire sous forme d'objets.

Nous aurions pu réécrire le code en C, qui est plus léger et complètement compilé. Cependant, nous avons jugé qu'il était plus économe en temps de travailler de manière séquentielle avec Java SAX 2.

Notre programme, lancé dans la JVM (machine virtuelle Java) ne prenait plus qu'une petite dizaine de minutes (hors traitement) pour parser les 1 631 850 entrées du fichier XML de 822 mégaoctets. Il ne restait plus qu'à tenir compte de l'expérience acquise lors de la première approche pour intégrer les données à la base de connaissances.

Le choix d'intégration des données issues du fichier se posait alors. Soit, nous décidions d'ouvrir un connecteur avec la base de données, soit nous générions directement un fichier au format SQL. Cette deuxième option offrait la possibilité d'un accès limité à la base de données. En effet, l'import se fait en une fois grâce à la commande Shell MySQL (rapide et scriptable) ou par une sur-couche logicielle comme PHPMyAdmin (convivial). Notre choix s'est arrêté sur une intégration séquentielle par un connecteur Java/MySQL. Cette option offre la possibilité d'une mise à jour sans écraser la base à chaque mise à jour. De plus, la base reste accessible pour l'exploitation tout au long de la procédure, même si son accès est sensiblement ralenti.

Bibliographies scientifiques : de la recherche d'informations à la production de documents normés

Résumé : Dans un cycle ou chaque document scientifique, s'inspirant lui même d'autres productions, sera lu et commenté par des chercheurs qui le citeront, l'écrit sera une production tour à tour finie, réactualisée et toujours une source d'appropriation et de citation. Après une introduction à la recherche d'information, nous examinerons la typologie des documents scientifiques, les modalités de stockage et de diffusion, ainsi que les normes et protocoles associés. Nous décrivons plusieurs méthodes de recherche documentaire et différents outils d'interrogation des bases de connaissances. Nous postulons qu'aujourd'hui la recherche documentaire peut être techniquement automatisée, de la première étape d'établissement du périmètre de recherche jusqu'à l'écriture de la bibliographie. Les étapes de sélection et de gestion documentaire peuvent aussi être facilitées par des outils et normes dédiés. Nous proposons une étude centrée sur l'usager pour faire émerger des profils utilisateurs et les usages associés, puis nous soumettons une démarche conceptuelle et expérimentale d'accompagnement visuel à la recherche de documentation scientifique. Nous nous intéressons tant aux méthodologies d'évaluation et de recommandation de ce type de littérature qu'aux formats et normes documentaires. Nous synthétisons l'ensemble des procédures d'automatisation bibliographiques pour modéliser un outil de recherche alliant respect des normes, souplesse d'usage et considération des besoins cognitifs et documentaires de l'usager. Cette interface servira de support à une recherche naviguée dans les corpus documentaires avec l'intégration de services d'exposition de métadonnées.

Mots clés : Information scientifique et technique, recherche d'information, bibliothèque numérique, bibliographie, hypergraphes, ontologies, taxonomies.

Scientific bibliographies : from information retrieval to standardized bibliographies

Abstract : In a cycle of research, in which each scientific document, itself inspired by previous papers, will be read and commented by researchers who will quote it, a document will be an alternately finished, an updated production and a potential source of appropriation and quotation.

After an introduction to information retrieval, we will examine the typology of scientific documents, what the storage and broadcasting of information model can be, as well as their standards and related protocols.

We describe several methods of document retrieval and various tools of interrogation of knowledge bases. We postulate that nowadays, the document retrieval can be technically automated, since the first stage of establishment of the research's scope up to the writing of the bibliography. We introduce dedicated tools and standards can also facilitate the stages of selection and documentary management.

We propose a user focused study in order to establish several users' profiles and associated uses. Then we submit a conceptual process of visual information retrieval. To do so, we analyze methodologies of rating this specific type of paper as well as scientific papers' criterions and standards.

We synthesize all the bibliographical procedures of automation to model a search tool allying respect for the standards, the flexibility of use and consideration of the of the users' cognitive needs. This interface will be used as support to browse conceptual domain as an access to scientific corpuses. It will also expose of document's metadata with LIS good practices and semantic web's rules. After evaluation, this tool is generalized as a model for scientific information retrieval design. **Keywords :** Scientific and technic information, e-Library, bibliography, hypergraph, ontologies, taxonomies.