Université du Littoral Côte d'Opale **Université Gaston Berger de Saint-Louis**

Contribution à la modélisation de données multi-sources de type DATAWEB basé sur XML

THÈSE

Présentée et soutenue publiquement le xx xx xxxx

Pour l'obtention du

Doctorat de l'Université du Littoral Côte d'Opale

Et de

L'Université Gaston Berger de Saint-Louis

(Spécialité informatique)

Par

Ousmane SALL

Composition du jury :

Président :	Le président
Rapporteurs :	Pr. Claude Yves Chrisment, Université de Toulouse
	Pr. Mourad Chabane Oussalah, Université de Nantes
Examinateurs :	L'examinateur 1
	L'examinateur 2
Invité :	L'invité

Laboratoire d'Informatique d Littoral | EA 4029

Remerciements

Je remercie vivement le Pr. Henri BASSON, Directeur du Laboratoire d'informatique du Littoral, de m'avoir accueilli dans son Laboratoire durant ces longues années de thèse. Je le remercie aussi d'avoir bien voulu m'encadrer en cotutelle. Mes remerciements vont également au Pr. Moussa LO pour sa disponibilité, son encadrement, sa patience, et sa collaboration sans faille. Je remercie aussi le Pr. Mary Teuw NIANE sans qui cette thèse en cotutelle ne serait jamais possible.

Je tiens aussi à remercier très chaleureusement Nathalie PERNELLE pour les discussions fructueuses et ses nombreux conseils. Il en est de même pour M. Fabien GANDON, qui en plus de constituer une référence pour les étudiants en Informatique de l'Université Gaston berger de Saint-Louis, m'a fait l'honneur de bien vouloir participer à la réalisation de ce travail. Je n'oublierais pas de remercier aussi Cheikh NIANG pour son sérieux et son efficacité.

Je voudrais également exprimer mes sincères remerciements à mes rapporteurs Pr. Claude Yves CHRISMENT et Pr. Mourad Chabane OUSSALLAH de m'avoir fait l'immense honneur d'être les rapporteurs de ma thèse et pour leurs nombreuses remarques et suggestions, qui m'ont permis d'améliorer considérablement mon travail. Mes remerciements vont également à M. Mourad BOUNEFFA et au Pr. Aziz DOUKKALI pour avoir accepté d'évaluer mon travail.

Mes remerciements vont aussi à toutes les personnes qui m'ont aidé durant mes séjours difficiles à Calais. Je commencerais par M. Ali LAMINOU qui m'a accueilli chez lui durant tous mes séjours d'alternance à Calais. Il m'a hébergé et nourrit de manière désintéressée dans des moments difficiles et a souvent contribué au payement de mes billets d'avion. Sans son aide, je n'aurais jamais pu achever cette thèse. Je nommerais également mon ami et frère Antoine ABDOULLAHI qui m'a encouragé et soutenu dans des moments de survie à Calais. Sans oublier dans le même ordre le Dr. Pierre Tifi MAMBI ainsi que sa femme Guilaine qui m'ont accueilli et beaucoup aidé durant mes deux premières années à Calais. Il en est de même pour ma grande sœur Togolaise, comme je l'appelle Emilie AGBODJAN, qui m'a beaucoup assistée.

Merci également à mes camarades doctorants du LIL et l'équipe ModEL pour leur accueil, leur ouverture et disponibilité. Je remercie Christophe Soulié avec qui j'ai fait mes premiers pas dans la recherche au sein du laboratoire et pour ses encadrements en DEA.

Je remercie très sincèrement mes collègues et amis de l'Université de Thiès pour leur compréhension durant mes périodes d'alternance en France, ainsi que leurs encouragements. Je remercie en particulier Lamine BOUSSO, Cheikh SARR, Mouhamadou THIAM, Cheikh THIAM, Alioune CAMARA, Ibrahima MBAYE, Saliou NIDIAYE, Ibrahima NDIAYE, Oumar NDIAYE, Mamadou SARR.

Je ne peux pas oublier mes amis de toujours Abdoulaye FALL, Bamba SECK, Mamadou NDIAYE et leur famille, je les remercie pour leurs encouragements. En particulier, je réserverais une mention spéciale à mon ami d'enfance Pape El Hadji. A. DIAGNE pour ses encouragements, ses corrections et relectures.

Enfin, je ne trouve pas de mots assez forts pour exprimer mon sentiment de reconnaissance et de profonde gratitude à ma famille. Je voudrais citer surtout mon père, ma mère, mes frères et sœurs maternels, mes oncles afin de leur dire que cette reconnaissance leur sera toujours éternelle, afin de leur remercier aussi de l'effort et des sacrifices financiers qu'ils ont dus faire pour me permettre de poursuivre mes études en France. Leurs encouragements m'ont toujours permis de m'accrocher dans les moments difficiles et de ne pas me décourager. Merci aussi à ma tendre épouse pour son soutien et sa compréhension avec mes longues absences répétées. Je témoigne ma reconnaissance et ma gratitude à l'endroit de mon oncle Saliou SEYE, à Salimata SEYE et son époux Maguette SOW qui m'ont toujours encouragé dans mes ambitions et rendu service et assisté dans beaucoup de domaine.

A Tous je vous dis encore merci.

Je dédie cette thèse à mon défunt père
ainsi qu'à ma fille Aïcha.

Table des matières

Table des matières 7

Table des figures 11

Liste des tableaux 13

Chapitre 1 Introduction Générale 15

1.1 L'intégration de données 15
1.1.1. Problématique 16
1.1.2. Cadre applicatif 17
1.1.3. Approches d'intégration 18
1.2. Le travail de thèse 19
1.2.1. Approche dataweb sémantique 19
1.2.2. Contributions 21
1.3. Organisation de la thèse 22
Première partie Etat de l'art 25

Chapitre 2 Intégration de données issues de sources hétérogènes 27

2.1 Introduction 28
2.2 Problématique de l'hétérogénéité des données 29
2.2.1 Hétérogénéité syntaxique 29
2.2.2 Hétérogénéité structurelle 30
2.2.3 Hétérogénéité sémantique 30
2.3 Approches d'intégration de données 31
2.3.1 Architecture des systèmes d'intégration 31
2.3.1.1 Approche entrepôt 32
2.3.1.2 Approche médiateur 33
2.3.2 Formalisation d'un système d'intégration 35
2.3.3 Mise en correspondance entre schéma global et schémas locaux 37
2.3.3.1 Approche Local-As-View 37
2.3.3.1.1 Description de l'approche 37
2.3.3.1.2 Prototypes utilisant l'approche 40
2.3.3.1.2.1 Le prototype Styx 40
2.3.3.1.2.2 Le projet PICSEL 41
2.3.3.2 Approche Global-As-View 42
2.3.3.2.1 Description de l'approche 42
2.3.3.2.2 Prototypes utilisant l'approche 44
2.3.3.2.2.1 Projet TSIMMIS 44
2.3.3.2.2.2 Projet MIEL++ 45
2.3.3.3 La médiation distribuée ou hybride 47
2.3.3.3.1 Description de l'approche 47
2.3.3.3.2 Prototypes utilisant l'approche 48
2.3.3.3.2.1 XYLEME 48
2.3.3.3.2.2 ActiveXML 50
2.3.3.3.2.3 XLive 51
2.3.4 Approche dataweb basée sur XML 52
2.3.4.1 Dataweb basé sur XML pour l'intégration de données 53

2.3.4.1.1 Composants du modèle 55
2.3.4.1.2 Modèle pour l'intégration structurelle 56
2.3.4.1.3 Modèle pour l'intégration sémantique 56
2.3.4.2 Variante de l'approche 57
2.3.4.3 Approches *dataweb*, « data web » et celle web sémantique 59
2.3.4.2.1 Web Sémantique 59
2.3.4.2.2 Critiques du web sémantique 62
2.3.4.2.3 Web sémantique deuxième version 63
2.4 Conclusion 64
Chapitre 3 Utilisation et conception des ontologies pour l'intégration de données 67

3.1 Introduction 68
3.2 Ontologie 68
3.2.1 Définitions 68
3.2.2 Composants d'une ontologie 70
3.3 Modèles formels d'ontologie 71
3.3.1 Structure d'ontologies à base lexicale 72
3.3.2 Modèle d'ontologie SOWA 74
3.4 Langages de représentation et d'interrogation 74
3.4.1 Extensible Markup Language 75
3.4.2 Ressource Description Framework/RDF-Schema 76
3.4.2.2 Niveau de définition des types de base RDF 79
3.4.2.3 Niveau de définition des types complexes 79
3.4.2.4 Niveau de définition des schémas 80
3.4.3 Ontology Web Language 82
3.4.4 Simple Protocol And RDF Query Language (SPARQL) 87
3.5 Extraction automatique d'ontologie à partir de données 88
3.5.1 Extraction d'ontologie par apprentissage automatique 89
3.5.2 Extraction d'ontologie à partir de sources XML 89
3.6 Approches d'intégration de données basées sur les ontologies 92
3.6.1 Approche mono-ontologie 93
3.6.2 L'approche multi-ontologie 94
3.6.3 L'approche hybride 95
3.7 Conclusion 96
Deuxième partie Contributions 97

Chapitre 4 Du modèle de dataweb au dataweb sémantique basé sur XML 99

4.1 Introduction 100
4.2 De l'intégration des données vers l'intégration des connaissances 100
4.2.1 Architecture globale du système d'intégration 100
4.2.2 Un processus d'intégration en trois phases 103
4.2.2.1 Intégration structurelle des données au sein des *partenaires* 103
4.2.2.2 Intégration sémantique des données au sein des *partenaires* 104
4.2.2.3 Médiation entre les différents *partenaires* 104
4.2.3 Les différentes couches de l'architecture 104
4.2.3.1 Sources natives du *partenaire* 104
4.2.3.2 Représentation structurée des données *partenaires* 105
4.2.3.3 Représenter la sémantique des données *partenaires* 105
4.2.3.4 Le niveau médiateur 105
4.3 Composants et modèle du système d'intégration 106
4.3.1 Un modèle de système d'intégration par partenaire 106

4.3.2 Un modèle d'ontologie spécifique par partenaire 108
4.3.3 Un modèle formel de base d'annotations 109
4.3.4 Un modèle de base de connaissances 110
4.3.5 Un modèle formel de *dataweb sémantique* 110
4.3.6 Un modèle formel de système d'intégration par partenaire 110
4.3.7 Un modèle formel de système d'intégration globale 111
4.4 Conclusion 111
Chapitre 5 Intégration sémantique et structurelle 113
5.1 Introduction 114
5.2 Nature et structure des données sources initiales 115
5.3 Construction d'un entrepôt pour chaque partenaire 116
5.3.1 Extraction et transformation des données sources 117
5.3.2 Restructuration et nettoyage des données sources 119
5.3.2.1 Restructuration ciblant les caractéristiques spatiales 123
La figure 5.5 donne la forme générale de la restructuration que nous proposons. C'est le cas du nœud « Label_Saptial1 » dans la figure 5.6. Il est restructuré et devient la valeur de l'attribut nom d'un nouveau nœud « localite ». L'item (a) montre le nœud original et le (b) le nouveau nœud obtenu après restructuration spatiale. 124
5.3.2.2 Restructuration ciblant les caractéristiques temporelles 124
5.3.2.3 Restructuration ciblant les unités de mesure 125
5.4 Processus d'extraction des ontologies partenaires 126
5.4.1 Extraction et subsomption des concepts candidats 128
5.4.1.1 Extraction des concepts candidats 129
5.4.1.2 Subsomption des concepts candidats 132
5.4.2 Extraction et inférence des relations sémantiques 133
5.4.2.1 Extraction de relations à partir du dataweb partenaire 133
5.4.2.1.1 Les relations de subsomption 133
5.4.2.1.2 Les relations associatives 134
5.4.2.1.3 Les relations d'attribut 134
5.4.2.2 Extraction de relations sémantiques par inférence 135
5.5 Construction des bases d'annotations 136
5.6 Construction des ontologies génériques aux sources 137
5.7 Un système à base de hubs pour une médiation entre partenaires 138
5.7.1 Architecture du système 139
5.7.2 Adaptation du système par rapport à notre contexte 140
5.8 Construction de l'ontologie globale 141
5.9 Conclusion 142
Troisième partie 145
Validation 145
Chapitre 6 Les données environnementales comme domaine d'application cible 147
6.1 Introduction 148
6.2 Le projet SIC-Sénégal 149
6.2.1 Description du projet 149
6.2.2 Participation et tâches dans l'équipe BDISIC 150
6.2.2.1 Tâches de cette thèse dans ce contexte 151
6.2.2.2 Autres approches et contributions dans l'équipe 152
6.2.3 Ressources existantes 153
6.2.3.1 Relevés de données 153

6.2.3.2 Le Service d'Ontologie Agricole (SOA) 154
6.2.3.3 Jena 156
6.2.3.4 Corese (Conceptual Resource Search Engine) 157
6.2.3.5 SeWeSe (Semantic Web Server) 158
6.2.4 Evaluations 159
6.3 Construction des dataweb partenaires 160
6.3.1 Extraction et transformation des données sources 161
6.3.2 Restructuration des données sources 162
6.3.2.1 Restructuration ciblant les caractéristiques spatiales 163
6.3.2.2 Restructuration des caractéristiques temporelles 164
6.3.2.3 Restructuration ciblant les unités de mesure 165
6.4 Construction des bases de connaissances 167
6.4.1 Processus d'extraction des ontologies partenaires 167
6.4.2 Construction des bases d'annotations 172
6.4.3 Construction des ontologies génériques 173
6.4.4 Construction de l'*ontologie globale* 174
6.5 Une médiation utilisant un système à base de hubs 174
6.6 Validation des ontologies 175
6.7 Conclusion 176
Chapitre 7 Le prototype AIDE-ISH 179
7.1 Introduction 179
7.2 Manager un partenaire 180
7.3 Générer un dataweb sémantique partenaire 180
7.4 Visualiser la base d'annotations d'un partenaire 181
7.5 Envoyer une requête partenaire ou la distribuer 181
7.6 Conclusion 182
Chapitre 8 Conclusion et perspectives 185
8.1 Conclusion 185
8.2 Contributions 186
8.2.1 Intégration de données par *partenaire* par une approche dataweb sémantique 187
8.2.1.1 Approche d'homogénéisation structurelle basée sur la notion de *dataweb* 187
8.2.1.2 Approche dataweb sémantique pour l'intégration sémantique 188
8.2.2 Une approche d'intégration des dataweb sémantiques 189
8.3 Travaux et perspectives de recherches 189
8.3.1 Maintenance évolutive 190
8.3.2 Langage et interface de requêtes 191
8.3.3 Extension aux autres formats et aux applications distribuées 191
8.3.4 Construction de dataweb thématiques 191
8.3.5 Imputation des données manquantes 192
Bibliographie 195
Annexe A 207
A.1 Sources Tabulaires initiales 207
A.2 Dataweb de la SAED 207
A.3 Ontologie Partenaire de la SAED 210
A.4 Ontologie générique de la SAED 222
A.5 Base d'annotations de la SAED 226

Table des figures

Figure 2-1 Architecture générale d'un système d'intégration 31
Figure 2-2 Illustration de l'approche entrepôt 32
Figure 2-3 Illustration de l'approche médiateur 34
Figure 2-4 Architecture conceptuel d'un système d'intégration de données 35
Figure 2-5 Modèle structurel de l'approche LAV 38
Figure 2-6 Exemple construction d'un schéma médiateur en approche LAV 39
Figure 2-7 Architecture du médiateur STYX 40
Figure 2-8 Architecture du médiateur PICSEL 41
Figure 2-9 Modèle structurel de l'approche GAV 42
Figure 2-10 Exemple de construction d'un schéma médiateur en approche GAV 44
Figure 2-11 Architecture de TSIMMIS 45
Figure 2-12 Architecture de MIEL++ 46
Figure 2-13 Modèle structurel de l'approche GLAV 47
Figure 2-14 Architecture fonctionnelle de XYLEME 49
Figure 2-15 Architecture d'ActiveXML 50
Figure 2-16 Architecture d'XLive 51
Figure 2-17 Processus de construction du dataweb (extrait de [Lo, 2002]) 54
Figure 2-18 Schématisation des unités d'informations d'un dataweb (extrait de [Lo, 2002]). 54
Figure 2-19 Architecture d'un dataweb basé sur XML 55
Figure 2-20 Architecture du modèle d'intégration de données 58
Figure 2-21 Architecture du web sémantique 60
Figure 2-22 Limites de la recherche par mot-clé 60
Figure 3-1 Illustration graphique des composantes d'une ontologie 73
Figure 3-2 Exemple de schéma d'un triplet RDF 77
Figure 3-3 Exemple RDF 78
Figure 3-4 Exemple de Schéma-RDF 82
Figure 3-5 Exemple de structure ontologique 83
Figure 3-6 Cas d'utilisation de multiples ontologies 84
Figure 3-7 Cas d'utilisation d'une *ontologie globale* 93
Figure 3-8 Cas d'utilisation de multiples ontologies 94
Figure 3-9 Architecture de l'approche hybride 95
Figure 4-1 Architecture globale du système d'intégration 101
Figure 4-2 Découpage du processus d'intégration des données et des partenaires 102
Figure 4-3 Description des 4 niveaux du processus d'intégration 103
Figure 5-1 Trois composants de base du système d'intégration 114
Figure 5-2 Structure XML résultant de la transformation du tableau 5.1 117
Figure 5-3 Document XML résultant de la transformation du tableau 5.1 118
Figure 5-4 Une vue sur la structure XML résultant de l'XMLisation du tableau 5.2 122
Figure 5-5 Forme générale du nœud inséré pour la restructuration spatiale 123
Figure 5-6 Exemple de restructuration ciblant les caractéristiques spatiales 123
Figure 5-7 Forme générale de restructuration ciblant les caractéristiques temporelles 124
Figure 5-8 Exemple de restructuration ciblant les caractéristiques temporelles 124
Figure 5-9 Forme générale de restructuration ciblant les unités de mesure 125
Figure 5-10 Exemple de restructuration ciblant les unités de mesure 125
Figure 5-11 Une vue de la structure résultant de la restructuration de la figure 5.4 126

Figure 5-12 Exemple d'extrait d'un format ontologie OWL d'un partenaire 127
Figure 5-13 Illustration de l'extraction d'un concept à la recherche de relations 128
Figure 5-14 Processus de construction des ontologies 129
Figure 5-15 Intégration semi-structurée 135
Figure 5-16 Extrait d'un format de base d'annotations partenaire d'un partenaire 136
Figure 5-17 Extrait d'un format d'une ontologie générique en construction 137
Figure 5-18 Architecture globale d'un système à base de hubs 139
Figure 5-19 Architecture globale d'un hub 140
Figure 5-20 Schéma sur la construction de l'*ontologie globale* 141
Figure 6-1 La vallée du fleuve Sénégal 150
Figure 6-2 Problématique du besoin d'intégration dans la vallée du fleuve Sénégal 151
Figure 6-3 Contributions au projet SIC-Sénégal 152
Figure 6-4 Architecture de Jena 157
Figure 6-5 Architecture de Corese 158
Figure 6-6 Architecture de SeWeSe 159
Figure 6-7 Structure XML résultant de la transformation du tableau 6.1 161
Figure 6-8 Une vue sur la structure XML résultant de l'XMLisation du tableau 6.1 163
Figure 6-9 Exemple de restructuration ciblant les caractéristiques spatiales 164
Figure 6-10 Exemple de restructuration ciblant les caractéristiques temporelles 165
Figure 6-11 Exemple de restructuration ciblant les unités de mesure 165
Figure 6-12 Une vue de la structure résultant de la restructuration de la figure 6.8 167
Figure 6-13 Structure graphique de l'ontologie de la DRDR-SL 169
Figure 6-14 Graphe de concepts de l'ontologie de la DRDR-SL 170
Figure 6-15 Illustration d'une subsomption à un concept générique 171
Figure 6-16 Illustration d'une extraction de relations type r_261 d'un document XML 171
Figure 6-17 Définition OWL de la relation « r_261 » dans $\mathcal{AOS}$ 172
Figure 6-18 Extrait de base d'annotations d'un partenaire 173
Figure 6-19 Extrait d'une ontologie générique d'un partenaire 174
Figure 6-20 Architecture conceptuelle d'un Hub 175
Figure 6-21 Trace du chargement de l'ontologie de l'ADRAO sous Corese 176
Figure 7-1 Interface pour manager un partenaire 180
Figure 7-2 Interface pour générer le dataweb et le dataweb sémantique d'un partenaire 181
Figure 7-3 Interface pour visualiser la base d'annotations d'un partenaire 181
Figure 7-4 Interface pour soumettre une requête à un ou plusieurs partenaires 182

Liste des tableaux

Tableau 3-1Tableau des mapping d'XML-Schéma à OWL ([Bohring et Auer, 2005]) 92
Tableau 5-1 Structure d'un tableau « individus * variables » général 115
Tableau 6-1 Exemple de tableau de données partenaire (Ici de la DRDR-SL)..... 154
Tableau 6-2 Partenaires et part de chacun dans les échantillons de données..... 160
Tableau A-1 Evolution des fourchettes de prix des riziers 207
Tableau A-2 Prix au producteur de la tomate 207
Tableau A-3 Prix au producteur en l'an 2000 207

Chapitre 1

Introduction Générale

Sommaire

1.1 L'intégration de données 15
1.1.1. Problématique 16
1.1.2. Cadre applicatif 17
1.1.3. Approches d'intégration 18
1.2. Le travail de thèse 19
1.2.1. Approche dataweb sémantique 19
1.2.2. Contributions 21
1.3. Organisation de la thèse 22

Le présent travail concerne l'intégration de données issues de sources hétérogènes, de données de nature environnementale provenant d'observations ou de statistiques sur l'évolution de caractéristiques communes à plusieurs individus. L'intérêt sans cesse renouvelé de l'étude des impacts environnementaux de l'homme sur la nature et l'évolution de celle-ci, d'une part, et la mise à disposition sous forme de serveurs de données sur le web, d'autre part, donnent une importance particulière à ce domaine avec une masse de données considérable. Mais cette dernière est dépourvue d'homogénéité structurelle et sémantique pour leur mise en commun au sein d'un organisme fournisseur de données dit *partenaire* dans le cadre de ce travail et aussi la mise en commun des connaissances de différents systèmes d'informations.

L'objectif de ce travail est de développer, pour des données de cette nature, une approche permettant l'intégration structurelle et sémantique des données d'un *partenaire* d'une part, et la mise en commun des connaissances de plusieurs sources de données d'autre part.

1.1 L'intégration de données

Dans sa première vocation, le web est un système développé pour l'accès, la consultation et la mise en relation de documents contenant des données. Cette mise en relation des pages web est essentiellement basée sur l'hypertexte. Un rapport de *Netcraft*, spécialisé

dans les solutions d'analyses et de statistiques réseaux, a montré que du nombre d'un million de site web en 1997[1], on est actuellement passé à celui de 240 millions en juillet 2009[2]. Face à la masse de données, les moteurs de recherche proposent de localiser les informations en procédant à une indexation du contenu du web.

Cette forme de collecte et de localisation de l'information commence à montrer ses limites avec le monopole sur des connaissances disponibles et leur accès limité à un groupe restreint. Cette approche souffre également du manque de sémantique liée aux descripteurs des données, ce qui compromet les recherches en renvoyant à une masse colossale de données non souhaitées. Comme l'exemple simple de recherche du mot « *rice* » en anglais effectuée avec google[3] qui renvoie à cent treize millions de résultats incluant des noms d'université, d'organisations et de personnes. Une sémantique associée à ces données aurait permis de choisir le contexte dans lequel le mot clé est recherché. La mise en relation des données englobées dans le niveau de granularité des pages web ne suffira pas, de plus en plus d'initiatives s'orientent avec le web sémantique vers une mise en relation directe des données inspirées de l'hypertexte. Certains théoriciens évoquent de la notion d'« hyperdata » [Spivack, 2007] en référence à l'interconnexion des données.

Quelle que soit l'approche adoptée, une mise en relation ou une combinaison des données passe par la disponibilité d'un système intégré. Les données de ce système seront, soit représentées de la même manière du point de vue structurel pour faciliter leur accès et leur exploitation automatique, soit représentées dans un format standardisé. Dans le contexte de la normalisation du format d'échange, il est nécessaire de s'entendre sur l'utilisation et la disponibilité d'un système normalisant la description des données. Cela permettra d'utiliser un vocabulaire contrôlé pour décrire les données et leurs éventuelles relations.

1.1.1. Problématique

L'intégration de données pose essentiellement deux grandes problématiques : l'accès aux données et leur compréhension du point de vue de leur sens en vue de leur combinaison pour répondre à des requêtes.

La problématique d'accès aux données est liée à la multiplicité des formats de représentation (structuré, textuel, base de données). Du point de vue architectural, les données peuvent se trouver en un seul et même endroit ou totalement dupliquées chez plusieurs

[1] http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html
[2] http://news.netcraft.com/archives/2009/12/24/december_2009_web_server_survey.html
[3] Http://www.google.com

partenaires. Dans un contexte où plusieurs serveurs de données proposent un accès à leurs données, il est essentiel que cette forme d'hétérogénéité soit prise en compte, permettant ainsi de bien situer les données susceptibles de répondre à une requête donnée et de savoir comment y accéder.

La problématique liée à la compréhension des données concerne la manière dont les données sont décrites en fonction du vocabulaire utilisé. Pour combiner des masses de données, l'approche classique des bases de données consistant à lancer des requêtes pour identifier les champs par des chaînes de caractères ne suffit pas. Cela nous amène à examiner, dans la perspective de la sémantique, ce qui se trouve derrière ce descriptif, ce qui le motive et les caractéristiques de l'entité décrite. C'est la problématique de la sémantique des données. Elle permet de répondre à des questions auxquelles il est nécessaire de s'élever à un niveau d'abstraction supérieur à celui des données. C'est la problématique des connaissances décrivant les données. Dans l'optique environnementale, les données décrites sous un format tabulaire ou textuel véhicule implicitement un sens et les concepts utilisés dans le même contexte ne sont pas orphelins de toute relation. Il est aussi important de distinguer les éléments du vocabulaire utilisé et leurs interactions pour dégager le sens global du document.

1.1.2. Cadre applicatif

La mise en valeur de la vallée du fleuve Sénégal fait intervenir depuis un certain nombre d'années des experts appartenant à plusieurs organismes (Ministère de l'Agriculture, OMVS - Organisation pour la Mise en Valeur du Fleuve Sénégal, ISRA – Institut Sénégalais de Recherche Agronomique, SAED - Société d'Aménagement et d'Exploitation des terres du Delta, OMS - Organisation Mondiale de la Santé, etc.), de différents domaines de compétence (hydraulique, activités agricoles, recherche agronomique, santé, etc.) mais aussi localisés dans différents pays (Sénégal, Mali, Mauritanie, organismes internationaux). Tous ces experts mènent des travaux qui aboutissent généralement à la production et à l'exploitation de gros volumes de données.

La gestion et l'exploitation des données sont loin d'être satisfaisantes à cause de leur distribution, hétérogénéité, volume et appartenance (propriétaires différents). La problématique de l'hétérogénéité dans ce contexte est liée à la distribution des données sur plusieurs sources (SAED, OMVS,...). En plus, chaque source dispose de ses propres moyens et techniques de stockage (SGBD, matériels différents...), avec des précisions et périodes de collecte distinctes, d'un vocabulaire propre au *partenaire* ou partagé avec une sémantique diverse.

La nécessité s'est faite sentir de mettre en place des outils permettant une intégration de ces sources hétérogènes afin de disposer d'une exploitation des données intégrées pour fournir des services et faciliter la prise de décision ainsi que la recherche d'informations pertinentes sur un ensemble de données pour un besoin précis. Un projet nommé *SIC-Sénégal* (Système d'Information et de Connaissances) a été initié pour cela [Bdisic, 2004]. L'objectif du projet est de mettre une plate-forme logicielle à la disposition des producteurs de données (experts, organismes), et des consommateurs de données (décideurs, bailleurs) pour faciliter l'intégration, la gestion, l'organisation, la diffusion, et l'exploitation des données produites sur la région.

1.1.3. Approches d'intégration

Dans la section précédente, nous avons vu que l'intégration de données dans notre contexte vise essentiellement à résoudre un problème lié à leur hétérogénéité des points de vue de leur représentation (hétérogénéité structurelle), de l'organisation des structures (hétérogénéité organisationnelle) les regroupant ainsi que la compréhension et organisation des connaissances les décrivant (hétérogénéité sémantique).

D'un point de vue structurel, la solution proposée consiste à utiliser un format de représentation unique pour l'ensemble des données ou un système d'adaptateurs permettant de passer facilement entre les formats. Le seul but de la résolution de ce problème vise à l'homogénéisation du point de vue utilisateur de la représentation des données. Dans notre contexte l'aspect structurel de l'hétérogénéité se retrouve lorsque des partenaires utilisent des formats tabulaires pour représenter leurs données, d'autres des bases de données ou les rapports au format textuel. En plus de la diversité des formats au sein même d'un partenaire les données peuvent être réparties en plusieurs endroits et la nature distribuée des données se remarque le plus lorsqu'il faut localiser l'information recherchée. Du point de vue sémantique, l'objectif est de pouvoir combiner des données de thématiques diverses. Un même mot peut avoir des sens différents selon les domaines d'utilisation, et l'occurrence de deux mots dont les relations sémantiques ne peuvent être appréhendées qu'en ayant une bonne connaissance du contexte des données, par exemple la relation entre une occurrence de « riz » et « panicule ».

S'agissant de l'architecture des sources de données, il existe deux grandes approches, celle dite entrepôt et celle virtuelle ou médiateur [Hull et. Zhou, 1996], [Florescu et. al 1998]. La première approche permet de représenter les données dans un seul et même endroit ; elle est difficilement réalisable dans un cadre de répartition des données et de respect des aspects

propriétaires et privés des données. La difficulté de cette approche tient au respect de la propriété et au caractère confidentiel de certaines données, les fournisseurs de données étant généralement très « jaloux » de leurs données. La deuxième approche consiste d'abord à laisser les données dans leur source et, ensuite, offrir un système permettant à l'utilisateur d'avoir virtuellement l'impression que les données se trouvent en un seul et même endroit. Cette approche commande de combiner des données ayant des structures de représentation lié à la divergence sémantique entre données ayant des sources multiples pouvant varier selon leurs sources respectueuse et la résolution de l'aspect sémantique. C'est ce qui introduit la troisième forme d'hétérogénéité.

D'un point de vue sémantique, pour résoudre l'hétérogénéité des données, la solution la plus opératoire consiste à s'entendre sur l'ensemble des termes à utiliser pour représenter les objets cibles des observations et la description de leurs caractéristiques. C'est ce que l'on appelle un vocabulaire contrôlé. L'étape suivante voudrait que l'on s'entende sur des caractéristiques, relations et hiérarchie reconnues comme existants entre les objets considérés. Le vocabulaire est donc contrôlé et hiérarchisé. Nous optons pour cette démarche logique, avec les différents langages que nous utilisons pour communiquer (communiquer, dans le sens d'établir une relation avec autrui, de diffuser un message auprès d'une audience plus ou moins vaste et hétérogène). Le langage comprend un ensemble de concepts décrivant les objets reconnus comme existants ainsi que leur manière d'être et leurs interactions.

1.2. Le travail de thèse

Dans cette thèse, notre objectif est d'une part, de fournir une solution d'intégration structurelle et sémantique des données d'un partenaire désirant partager ses données et, d'autre part, de proposer une solution à l'intégration des connaissances entre les partenaires. Ces solutions doivent intégrer la contrainte du respect de l'aspect propriétaire et confidentiel dans le cadre de l'approche de représentation architecturale des données.

1.2.1. Approche dataweb sémantique

L'approche d'intégration que nous proposons peut être qualifiée de structurelle et sémantique. D'une part, elle vise à résoudre la problématique de l'hétérogénéité structurelle et sémantique des données internes à un système d'information. D'autre part, elle propose une solution permettant une combinaison de leurs données à des organismes désireux de partager leurs données dans un cadre partenarial.

Sur le plan architectural, cette approche est chevillée autour de deux axes. Le premier concerne un système d'intégration de données internes à un *partenaire* et le second est celui de l'intégration des systèmes d'intégrations. Elle est basée sur la définition d'un modèle d'homogénéisation structurelle et sémantique des données reposant sur les possibilités offertes par le langage XML et les ontologies.

L'homogénéisation structurelle est fondée sur une phase de pré-intégration transformant l'ensemble des données sources au format XML sous forme d'*entrepôt de documents XML* dit *dataweb* [Hocine et. Lo, 2000], [Lo et. Hocine, 2005]. Une phase de réorganisation des données permet, grâce à l'intervention de l'expert du domaine, une restructuration des documents combinés par une extraction et mise en exergue des caractéristiques spatio-temporelles des données. Ainsi, nous disposerons du même format de représentation des données sur chaque serveur de données, facilitant ainsi l'échange des données.

Nous avons ensuite développé une méthodologie permettant d'extraire le vocabulaire utilisé pour décrire les données, et la hiérarchie entre les différents éléments de ce vocabulaire. Elle permet aussi d'enrichir ce système sémantique grâce l'utilisation d'une ontologie existante du domaine. Cette ontologie extraite est associée à une *base d'annotations* associant chaque concept du vocabulaire extrait aux données qu'il décrit, constituant ainsi une ontologie du *partenaire* et une partie de sa *base de connaissances*. L'apport de sémantique au dataweb par l'intermédiaire de la *base de connaissances* justifie la notion de *dataweb sémantique* qui qualifie notre approche. Nous assimilons dans tout le document la notion d'*ontologie partenaire* à celle décrivant l'ensemble des données d'un partenaire ; plus que l'ontologie du domaine, elle conceptualise explicitement tous ses domaines d'intervention.

Le partage des données entre plusieurs *partenaires* passe par une homogénéisation des connaissances utilisées pour représenter localement leurs données. Cette homogénéisation doit tenir compte de l'aspect confidentiel de certaines connaissances que le *partenaire* veut rendre disponible localement mais pas forcément consultable en dehors de son système d'intégration. Il faut donc que le *partenaire* décide des connaissances qu'il désire partager avec les autres. Nous avons alors introduit dans la démarche d'intégration une étape dans laquelle après la construction de l'*ontologie partenaire*, le *partenaire* désigne les connaissances qu'il désire partager avec les autres à travers une ontologie dite *générique* du *partenaire*.

Dans notre contexte, ce partage est basé sur le principe que l'on ne partage que ce que l'on a en commun. Ainsi un mécanisme d'échange des *ontologies génériques* permet de

constituer une *ontologie globale* avec uniquement les connaissances que tout le monde a en commun. L'ontologie *partenaire*, l'ontologie générique et la *base d'annotations* constituent la *base de connaissances* du *partenaire*. Cette *base de connaissances* à base ontologique et le *dataweb* intégrant structuralement les données *partenaires* constituent le *dataweb sémantique* du *partenaire*.

Dans ce système, chaque organisme devra pouvoir envoyer des requêtes aux autres serveurs de données et aussi en recevoir. Nous utilisons ainsi un système à base de hubs dans lequel chaque hub héberge le *dataweb sémantique* d'un *partenaire*.

1.2.2. Contributions

L'approche que nous proposons est essentiellement fondée sur un système d'intégration contenant : les données au format XML, la structure décrivant sémantiquement ces données et celle pour le partage et la combinaison de ces connaissances avec les autres. Nous avons ainsi deux composantes principales : le système d'intégration des *partenaires* et le système d'intégration sémantique des connaissances sur les données des différents *partenaires*. Une première des trois contributions de cette approche est son application à l'intégration de données environnementales.

Nous avons ainsi introduit un modèle de *dataweb sémantique* dont la mise en œuvre permet l'intégration structurelle et sémantique des données d'un *partenaire* au moyen d'un *entrepôt de documents XML* (*dataweb*) et d'une *base de connaissances*. Cette première caractéristique nous a permis de développer une approche permettant l'extraction d'une ontologie décrivant les connaissances sur les données de chaque *partenaire* à partir de ses sources enrichie grâce à l'utilisation d'une ontologie existante du domaine. Chaque concept extrait est alors associé à un concept existant dans l'ontologie de référence permettant ainsi l'extraction de relations sémantiques supplémentaires. L'intervention de l'expert du domaine permet de corriger la structuration des connaissances. L'expert est dans notre contexte le scientifique ou spécialiste intervenant indépendamment d'un organisme partenaire ou pas et par ses activités est reconnu comme maîtrisant les connaissances dans son (ses) domaines(s) de compétence.

Pour l'intégration des différents *dataweb sémantique*, nous avons proposé un système d'intégration exploitant une *architecture à base de hubs*. Ce système permet à chaque *partenaire* de déclarer aux autres les données qu'il désire partager évitant ainsi le processus onéreux de l'indexation des données de chaque *partenaire*. Alors que dans certains systèmes

la médiation des ontologies est effectuée deux par deux, notre approche permet une médiation globale grâce à l'*ontologie générique* de chaque partenaire.

1.3. Organisation de la thèse

Ce mémoire s'articule autour de neuf chapitres présentant notre approche d'intégration, la bibliographie ayant servie de base à développer cette approche et l'application de la démarche d'intégration au contexte applicatif des données environnementales.

Le **chapitre 2** présente un état de l'art des systèmes d'intégration de données. Nous présentons dans ce chapitre les différents types d'hétérogénéité et approches d'intégration de données. Des systèmes basés sur ces différentes approches d'intégration sont présentés. Ce chapitre présente également l'*approche dataweb basée sur XML* pour l'intégration structurelle des données.

Le **chapitre 3** présente un état de l'art sur l'utilisation des ontologies pour l'intégration sémantique de données. Nous nous attelons dans ce chapitre à présenter la notion d'ontologie, les langages et modèles formels proposés. Nous présentons aussi les différentes approches de construction des ontologies et leur utilisation dans les approches d'intégration de données.

Le **chapitre 4** présente le modèle formel et les composantes de notre démarche d'intégration. Nous présenterons également dans ce chapitre l'architecture du système d'intégration.

Le **chapitre 5** présente les détails du processus de construction de chaque *dataweb partenaire*. Cette construction cible des données structurées et représentées d'une manière particulière. Nous présentons également dans cette partie la méthodologie globale d'identification, d'extraction et la mise en exergue des caractéristiques spatio-temporelles constituant les cibles de la phase de restructuration des données. Nous y abordons également dans une deuxième partie le processus de construction de la *base de connaissances* à base ontologique de chaque *partenaire*. Il détaille la méthodologie d'extraction des ontologies et de leurs bases d'annotations associées pour chaque *partenaire* à partir des données de son *dataweb*. Nous présentons aussi dans cette partie l'approche d'enrichissement de l'ontologie locale grâce à la réutilisation d'une ontologie plus générique existante. La troisième partie du chapitre présente l'approche d'intégration des systèmes d'intégration des *partenaires*. L'objectif de chaque organisme participant au processus d'intégration est de pouvoir disposer d'un cadre où ses données sont intégrées structurellement et sémantiquement mais aussi à un

niveau de granularité entre structures organisationnelles de mettre en commun les connaissances que chacun désire partager. Cette mise en commun est effectuée grâce à un *système à base de hubs* et une *ontologie générique* dupliquée chez tous les organismes participant au projet d'intégration.

Le **chapitre 6** présente le projet *SIC-Sénégal* qui constitue notre cadre applicatif. Nous présentons donc dans ce chapitre l'applicatif de construction des différents composants identifiés dans le chapitre 5.

Le **chapitre 7** présente le prototype du système d'intégration nommé AIDE-ISH implémentant la démarche d'intégration dans le contexte du projet *SIC*-Sénégal.

Première partie

Etat de l’art

Chapitre 2

Intégration de données issues de sources hétérogènes

Sommaire

2.1 Introduction 28

2.2 Problématique de l'hétérogénéité des données 29

2.2.1 Hétérogénéité syntaxique 29

2.2.2 Hétérogénéité structurelle 30

2.2.3 Hétérogénéité sémantique 30

2.3 Approches d'intégration de données 31

2.3.1 Architecture des systèmes d'intégration 31

2.3.1.1 Approche entrepôt 32

2.3.1.2 Approche médiateur 33

2.3.2 Formalisation d'un système d'intégration 35

2.3.3 Mise en correspondance entre schéma global et schémas locaux 37

2.3.3.1 Approche Local-As-View 37

2.3.3.1.1 Description de l'approche 37

2.3.3.1.2 Prototypes utilisant l'approche 40

2.3.3.1.2.1 Le prototype Styx 40

2.3.3.1.2.2 Le projet PICSEL 41

2.3.3.2 Approche Global-As-View 42

2.3.3.2.1 Description de l'approche 42

2.3.3.2.2 Prototypes utilisant l'approche 44

2.3.3.2.2.1 Projet TSIMMIS 44

2.3.3.2.2.2 Projet MIEL++ 45

2.3.3.3 La médiation distribuée ou hybride 47

2.3.3.3.1 Description de l'approche 47

2.3.3.3.2 Prototypes utilisant l'approche 48

2.3.3.3.2.1 XYLEME 48

2.3.3.3.2.2 ActiveXML 50
2.3.3.3.2.3 XLive 51
2.3.4 Approche dataweb basée sur XML 52
2.3.4.1 Dataweb basé sur XML pour l'intégration de données 53
2.3.4.1.1 Composants du modèle 55
2.3.4.1.2 Modèle pour l'intégration structurelle 56
2.3.4.1.3 Modèle pour l'intégration sémantique 56
2.3.4.2 Une approche similaire 57
2.3.4.3 Approches *dataweb*, « data web » et celle web sémantique 59
2.3.4.2.1 Web Sémantique 59
2.3.4.2.2 Critiques du web Sémantique 62
2.3.4.2.3 Web sémantique deuxième version 63
2.4 Conclusion 64

2.1 Introduction

La disponibilité croissante de sources de données variées et dispersées contenant des informations cruciales pour la prise de décision au sein des organisations pose souvent le problème de leur accès. Dans la plupart des cas, les sources ont été développées indépendamment et sont par conséquent hétérogènes. Cependant les organisations ont de nos jours le souci de faire inter opérer ces différentes sources pour offrir à leurs utilisateurs une vision globale de leur système d'information. La réalisation de cette interopérabilité passe par ce que l'on nomme un processus d'intégration de données hétérogènes [Halevy, 2001], [Hull, 1997], [Ullman, 1997].

Ce chapitre présente la problématique et les approches d'intégration de données. Il met en évidence les différentes facettes du besoin d'intégration et l'état de l'art des solutions proposées. Ce chapitre est organisé comme suit.

La partie 2.2 présente la problématique et les différentes facettes de ce besoin d'intégration.

Différentes approches sont proposées pour pallier à chacune des besoins d'intégration. Ces solutions peuvent être classifiées selon l'approche utilisée du point de vue matérialisation des données et celles permettant la mise en relation entre le schéma global et les schémas locaux des sources à intégrer. La présentation de ces approches est l'objet de la partie 2.3 de ce chapitre.

Nous présenterons également dans la partie 2.3 pour chacune de ces approches, des exemples de projets utilisant leur démarche d'intégration. Ces projets sont intéressants, car ils présentent des idées originales par rapport aux orientations de recherche retenues dans notre approche d'intégration. Parmi ces dernières l'*approche dataweb* basé sur XML que nous réutilisons dans ce travail.

2.2 Problématique de l'hétérogénéité des données

L'hétérogénéité permet de qualifier le caractère de ce qui est constitué d'éléments de nature différente (Le petit Larousse). En traitement de données, cette hétérogénéité est fortement liée aux choix de représentation, structuration et terminologie des données. Nous pouvons distinguer trois familles d'hétérogénéité selon ces choix. Différents modèles et standards de représentation peuvent être utilisés pour représenter la même information induisant ainsi un conflit de représentation lié à l'hétérogénéité dite syntaxique des données. Le choix de la représentation des données peut pousser deux sources de données à structurer et organiser la même information de manière différente. Cette hétérogénéité est dite structurelle. Des terminologies différentes peuvent être utilisées pour représenter la même information, c'est une hétérogénéité dite sémantique.

Nous allons discuter dans cette partie de ces différentes familles d'hétérogénéité ainsi que des solutions proposées pour les résoudre.

2.2.1 Hétérogénéité syntaxique

L'hétérogénéité syntaxique découle des choix effectués sur le format de représentation ou d'encodage des données constituant ainsi un blocage à la coopération de sources de données et leur interrogation. Il existe actuellement une grande variété de formats et standards de représentation de données. Les données peuvent être représentées de manière structurée comme dans les bases de données ou de manière semi-structurée.

Pour résoudre cette forme d'hétérogénéité, deux solutions peuvent être utilisées. La première consiste à utiliser un modèle en commun dit aussi modèle pivot que toutes les sources de données vont utiliser pour représenter leurs données. C'est le cas de XML dans le web actuel, son adoption dans de nombreux contextes en a fait un pivot pour l'échange de données [Laurent, 2004]. La deuxième solution consiste à mettre en place des fonctions de traduction ou de mise en correspondance permettant de passer d'un format à l'autre.

2.2.2 Hétérogénéité structurelle

La structure rend compte de la manière dont les éléments constitutifs d'un système (attributs, classes,... constitutifs de la donnée représentée) sont organisés. L'hétérogénéité structurelle est liée à l'utilisation de différents attributs, schémas, hiérarchie, propriétés, unités de mesure pour représenter le même concept. Cette hétérogénéité peut donc être liée à un conflit de schémas lorsque deux sources utilisent des représentations différentes pour représenter un même concept. Par exemple, le nombre d'attributs ou de classes décrivant un concept peut varier d'une source à l'autre.

La résolution de l'hétérogénéité passe par une intégration des schémas, il existe de nombreux travaux relatifs à l'intégration des schémas dans les bases de données. Nous y reviendrons dans l'état de l'art sur les solutions d'intégration de données.

2.2.3 Hétérogénéité sémantique

L'hétérogénéité sémantique est liée aux différentes représentations d'une information pour lui affecter un sens dans son « contexte ». Nous pouvons distinguer deux types d'hétérogénéité sémantique selon le niveau de granularité du sens de la représentation de l'information ou de sa valeur.

Selon des sources hétérogènes du point de vue sémantique, elle peut être liée à l'utilisation de (i) termes différents pour désigner la même information (problème de synonymie), (ii) deux termes identiques désignant en réalité des significations distinctes (problème d'homonymie), (iii) d'un même terme qui change de signification selon le contexte d'occurrence (problème de polysémie). Elle peut aussi être liée au langage utilisé ou à des variations syntaxiques telles que l'utilisation d'abréviations, préfixes par exemple.

Concernant la représentation de la valeur, des codages différents peuvent être utilisés pour représenter la valeur d'une information d'une source à l'autre. C'est le cas lorsque qu'un système utilise comme valeurs booléennes les valeurs « *0* » et « *1* » au moment où un autre utilise « *vrai* » et « *faux* ». La représentation des valeurs diffère mais est la même du point de vue sémantique. Il peut également se poser un problème d'hétérogénéité sémantique lié à l'utilisation par deux sources d'unité ou de précision de mesure différentes.

Pour résoudre l'hétérogénéité sémantique, il faut passer par l'établissement des correspondances entre les termes utilisés ou harmoniser les différentes représentations dans l'optique de la sémantique des connaissances sur les données si cette sémantique est disponible. Dans le cas contraire, il faudra procéder à une sémantisation des données pour

apporter un sens aux données. Il existe différentes approches de sémantisation, la sémantisation consistant à l'ajout d'une couche sémantique aux données.

2.3 Approches d'intégration de données

Nous pouvons classifier les systèmes selon deux types d'architectures d'intégration de données hétérogènes : l'approche entrepôt et celle dite « de médiation ».

La figure 2.1 illustre l'architecture physique de ces deux approches. Dans l'approche entrepôt, une migration et matérialisation dans un seul et même endroit est effectuée et les requêtes sur les données sont envoyées sur l'entrepôt. Tandis que pour l'approche dite de médiation ou virtuelle, les données sont laissées telles quelles et l'interrogation des sources est effectuée via une requête sur le schéma global qui sera réécrite en fonction des schémas locaux impliqués, les réponses aux requêtes locales étant ensuite migrées sur le site où est formulée la requête, puis agrégées et restituées à l'utilisateur.

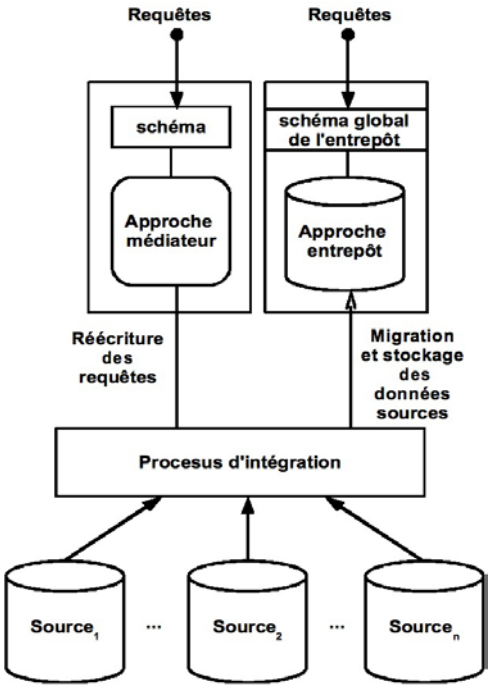


Figure 2-1 Architecture générale d'un système d'intégration

2.3.1 Architecture des systèmes d'intégration

Il existe globalement deux architectures d'intégration. L'une est basée sur la migration des données (l'approche entrepôt) et l'autre est basée sur la réécriture des requêtes. Cependant, quelque soit l'approche utilisée, il est important de disposer d'une (des) interface (s) entre un utilisateur et (entre aussi) les ressources : c'est la médiation. Nous pouvons

distinguer deux familles d'approches de médiation, celles centralisées et celles distribuées que nous allons aborder dans la suite.

2.3.1.1 Approche entrepôt

L'approche entrepôt applique le principe des vues matérialisées et intègre les données en accord avec les schémas globaux. Le résultat est un entrepôt de données qui peut directement être interrogé à travers un langage adapté [Widom, 1995].

Cette approche a été utilisée par plusieurs produits (avec des documents XML) dont les approches sont intéressantes dans le cadre de notre contexte applicatif. Il s'agit, entre autres, de Xyleme [Delobel et al., 2003], Xedix [Gerat, 2007], Enosys [Papakonstantinou et al., 2003], Tamino [Schöning et Wäch, 2000]. On procède pour cela à une copie des données sources en procédant à une migration des données vers l'entrepôt, ce qui est problématique lorsque l'on dispose d'une masse de données colossale.

Cette migration massive est réalisée par l'intégrateur comme le montre l'architecture de la figure 2.2. Ce composant est une sorte de middleware permettant la gestion de toutes les phases nécessaires à la construction de l'entrepôt identifié dans le domaine des entrepôts de données comme un processus ETL (Extract, Transform and Load - extraction, transformation et chargement).

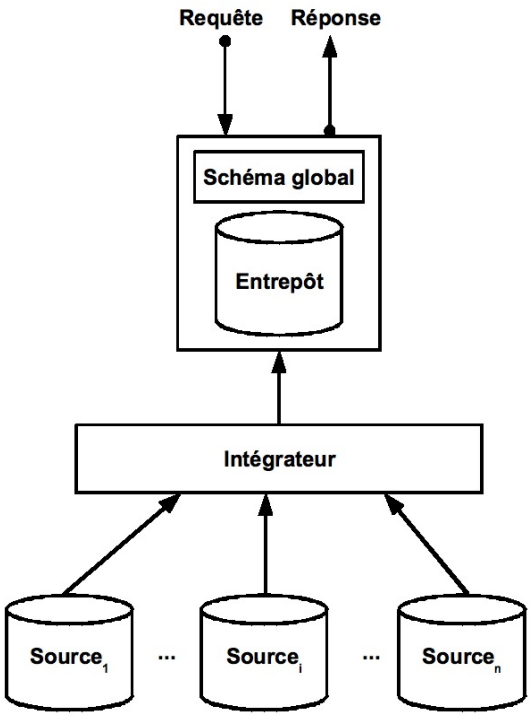


Figure 2-2 Illustration de l'approche entrepôt

L'extraction est la phase d'identification et de collecte des informations pertinentes pour la construction de l'entrepôt, la transformation vise généralement entre autres à disposer d'une structuration et la mise en évidence de certaines caractéristiques des données comme c'est le cas dans notre contexte.

Cette approche a plusieurs avantages, en termes de performance, avec la possibilité de procéder à des optimisations et la génération d'index. Les données étant stockées localement, il est aussi possible de les organiser, annoter, personnaliser en somme, adapter à différents besoins et aussi les archiver et faire des versions. Dans notre contexte, il n'est pas possible de constituer un entrepôt avec l'ensemble des données de tous les *partenaires*, une contrainte posée par ceux-ci est l'aspect privé de certaines de ces données.

Elle offre aussi la possibilité d'interroger des ressources passives, ce qui n'est pas le cas avec l'approche médiateur. Des inconvénients de cette approche sont aussi liés au rafraîchissement, à la mise à jour des données avec le volume des données très important.

2.3.1.2 Approche médiateur

L'approche médiateur [Wiederhold, 1995] ou paresseuse est fondée sur la définition de correspondances permettant la traduction de requêtes : une requête formulée par l'utilisateur dans les termes du schéma global est traduite en une ou plusieurs sous-requêtes qui sont évaluées sur les données sources. Les réponses sont combinées et transformées afin d'être compatibles avec le schéma global et conformes à la requête posée par l'utilisateur.

Gio Wiederhold a défini un médiateur comme « un module logiciel qui exploite la connaissance de certains ensembles ou sous-ensembles de données pour créer de l'information pour des applications à un niveau supérieur »[4].

C'est une approche qui présente plusieurs inconvénients, notamment en termes de performance avec le transfert des requêtes sur les sources. La traduction des requêtes n'est pas évidente non plus. Il y'a aussi la dépendance vis à vis des sources, des fonctionnalités proposées, du langage utilisé par la source. L'avantage est que l'on interroge toujours les données sources, les données d'origine. Nous avons une migration des requêtes vers les sources mais pas de matérialisation des données. Les premières solutions de médiation consistent à mettre en place un schéma global pour l'ensemble des sources et à décrire les correspondances entre les schémas locaux et le schéma global.

[4]http://www.m-grimaud.com/memoire-business-intelligence/etat-de-lart/faire-le-lien-entre-lamont-et-laval/approche-virtuelle/

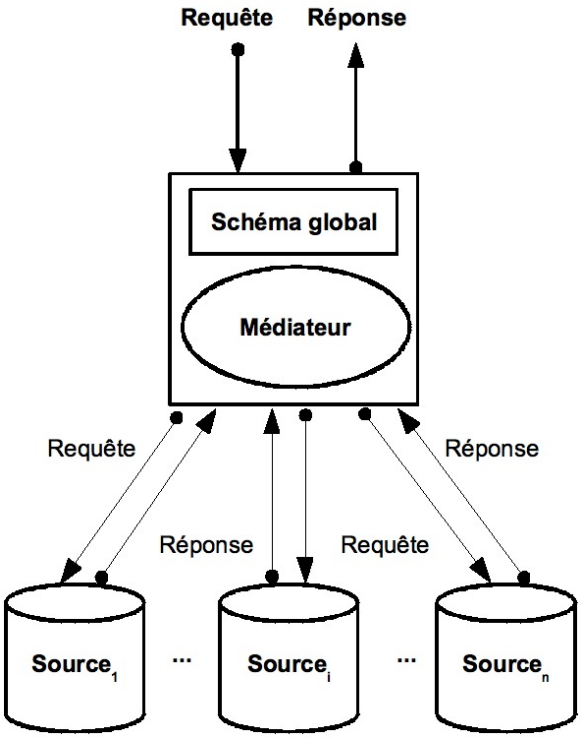


Figure 2-3 Illustration de l'approche médiateur

Cette solution présente une certaine lourdeur lorsque les schémas locaux évoluent ou lorsqu'un nouveau schéma local est introduit. Une autre solution consiste à introduire une ontologie pour rechercher automatiquement les correspondances existantes entre les schémas locaux et la requête de l'utilisateur. C'est cette approche que nous avons adoptée. Dans notre contexte d'application, chaque source joue à la fois un rôle de serveur et de client .

Les requêtes reçues par le médiateur sont exprimées dans le vocabulaire du schéma global ; ensuite chaque requête est analysée puis dépliée et envoyée aux sources de données concernées dans le langage de leur schéma, ensuite les réponses des différentes sources sont combinées pour fournir à l'utilisateur la réponse à sa requête. Un point important, c'est l'expression de cette requête et dans ce cas une question à résoudre est celle de la liaison entre le schéma global et les schémas locaux. Dans la plupart des travaux, cette liaison est basée sur un système de vues, selon que l'on définit le schéma médiateur comme une vue sur les sources ou chaque source comme étant une vue sur le schéma global, une approche hybride consiste quand à elle à allier les deux.

Nous allons étudier dans la suite le modèle global de formalisation des systèmes d'intégration ainsi que les approches de mise en correspondance entre le schéma global et les schémas locaux selon le système de vue que nous appelons modèle de description. Nous présenterons des exemples de systèmes utilisant les ontologies comme schéma local ou

global. Nous n'assimilons pas exclusivement les ontologies à des schémas des données, mais au schéma des connaissances décrivant les données. Il faut remarquer que dans de nombreux travaux, comme [Bellatreche et al., 2004], [Amann et al., 2002(a)], les ontologies sont utilisées comme schéma global.

2.3.2 Formalisation d'un système d'intégration

Les systèmes d'intégration sont généralement basés sur la mise en correspondance d'un schéma global et un ensemble de sources données. Ce schéma global offre une vue unifiée et intégrée sur un ensemble de sources de données. Une interrogation des sources passant par le schéma global nécessite la mise en place d'un système de mise en correspondance entre le schéma global et les schémas de sources. Du point de vue conceptuel, trois composantes sont donc prises en compte pour la formalisation d'un système d'intégration, le schéma global, les schémas locaux et le système de mise en correspondance.

La figure 2.4 donne une illustration des composantes structurelles d'un système d'intégration de données selon ces trois composantes. Une interrogation des sources passant par le schéma global nécessite la mise en place d'un système de correspondances entre le schéma global et les schémas de sources.

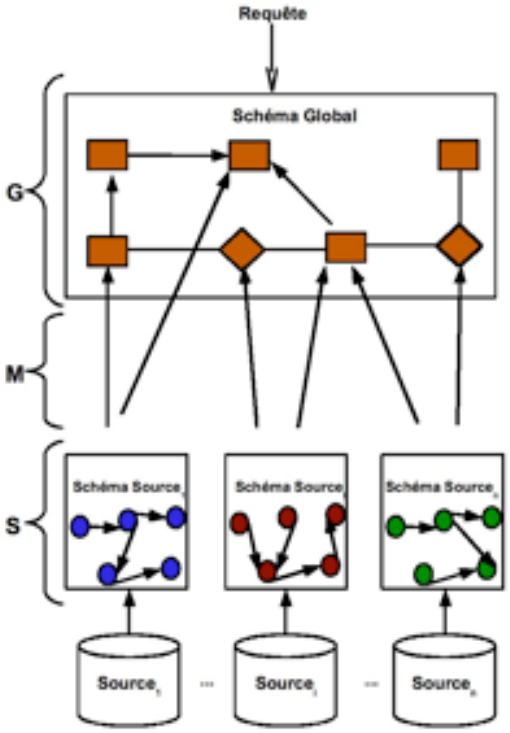


Figure 2-4 Architecture conceptuel d'un système d'intégration de données

Un système d'intégration de données I est formalisé selon le triplet [Lenzerini, 2002] :

$I := \{\mathcal{G}, \mathcal{S}, \mathcal{M}\}$ Où :

— $\mathcal{G}$ est le schéma global ou médiateur représentant le domaine d'intérêt du système d'intégration. Il exprime avec des contraintes d'intégrité dans un langage $\mathcal{L}_{\mathcal{G}}$ sur un alphabet AG comprenant un symbole pour chaque élément de $\mathcal{G}$. Dans [Calì et al., 2005] il est exprimé dans un fragment de la logique du premier ordre avec une équivalence dans un alphabet formé par, éventuellement un ensemble infini d'ensemble Γ de symboles.

— $\mathcal{S} := \{Si,\ 1 \leq i \leq n \in \mathcal{N}^*\}$ noté comme un ensemble fini de n schémas sources décrivant la structure des sources composant le système d'intégration, il est exprimé dans un langage $\mathcal{L}_{\mathcal{S}}$ sur an alphabet $\mathcal{A}_{\mathcal{S}}$ comprenant un symbole pour les éléments de chacune des sources;

— $\mathcal{M} := \{\mathcal{M}i\ ,\ 1 \leq i \leq n \in \mathcal{N}^*\}$ est le système de correspondances établissant une relation entre $\mathcal{G}$ et $\mathcal{S}$. C'est un ensemble de n correspondances entre les sources et le schéma global, tel que pour tout schéma de source Si, il existe $\mathcal{M}_i$ de $\mathcal{S}_i$ à $\mathcal{G}$; il existe différentes approches de spécification de ces correspondances qui sont cruciales pour le système d'intégration, constitué par un ensemble d'axiomes de la forme :

$q_{\mathcal{S}}\ \zeta\ q_{\mathcal{G}}$ où :

— $q_{\mathcal{S}}$ et $q_{\mathcal{G}}$ sont deux requêtes de même arité, respectivement sur le schéma source $\mathcal{S}$, et le schéma global $\mathcal{G}$;

— La requête $q_{\mathcal{S}}$ exprimée dans un langage de requête $\mathcal{L}_{\mathcal{M},\mathcal{S}}$ sur l'alphabet $\mathcal{A}_{\mathcal{S}}$;

— La requête $q_{\mathcal{G}}$ dans un langage de requête $\mathcal{L}_{\mathcal{M},\mathcal{G}}$ sur l'alphabet $\mathcal{A}_{\mathcal{G}}$;

— ζ est un symbole de l'alphabet {<, ≤, ≅, ≥, >}. Si ζ était ≅, alors une assertion $q_{\mathcal{S}} \cong q_{\mathcal{G}}$ spécifie que le concept représenté par la requête $q_{\mathcal{S}}$ sur les sources correspond au concept dans le schéma global représenté par la requête $q_{\mathcal{G}}$ (idem pour une assertion du genre $q_{\mathcal{G}} \cong q_{\mathcal{S}}$).

Exemple : Soit l'exemple suivant d'un système d'intégration de données dont le schéma $\mathcal{G}_0$ fait référence aux critiques de relecteurs d'articles soumis à des conférences en informatique,

$I_0 := \{\mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0\}$ où :

– $\mathcal{G} := \{article\ (Titre,\ Année,\ Conférence),\ informatique(Conférence),\ reviewer(Titre,\ Critique)\}$ est le schéma global, donc l'alphabet $\mathcal{A}_{\mathcal{G}0} := \{article,\ informatique,\ reviewer\}$;

– Le schéma source $\mathcal{S}_0 := \{S_0^i,\ 1 \leq i \leq 2\}$ a son alphabet $\mathcal{A}_{\mathcal{S}0}$ constitué de 2 sources S_0^1 et S_0^2 :

- la source sources S_0^1 d'arité 3 contient les informations sur les publications selon le schéma *r1(Titre, Année, Conférence)*, les titres de publications et conférences depuis 1994;
- La source S_0^2 d'arité 2 contient les informations selon le schéma *r2(Titre, Critique)* depuis 2001.

Nous allons nous baser sur cet exemple pour voir dans la suite comment on pourrait modéliser le schéma global, dans chacune des approches suivantes, dont le but cherché est de répondre à la question de savoir comment faire la liaison ou mapping entre les sources ou schémas décrivant les sources données et le schéma global exprimant en quelque sorte la sémantique de ses sources. Le symbole ζ équivaut à la notion d'équivalence symboliquement représentée par ≅ dans les tous les exemples qui suivront.

2.3.3 Mise en correspondance entre schéma global et schémas locaux

Une des étapes importantes de la mise en place d'un système d'intégration est la mise en correspondance ou mapping entre le schéma global et les schémas locaux. Selon la stratégie adoptée, on peut définir le schéma global comme une vue sur les schémas des sources de données ou les schémas des sources comme des vues sur les schémas globaux ou aussi adopter une approche hybride. Ces ensembles peuvent être repartis en deux groupes selon un contexte de médiation centralisée ou distribuée. Nous allons étudier ces différentes approches en étudiant la stratégie de formalisation du système de mapping. Dans la littérature, nous avons des travaux dans ce sens tels que dans [Calì et al., 2005], [Lenzerini, 2002] et [Abiteboul et Duschka, 1998].

2.3.3.1 Approche Local-As-View

Il est possible de décrire chaque ressource indépendamment des autres par rapport à un schéma global ou une ontologie. . C'est ce que l'on appelle modèle de description *LAV* ou *Local-as-View*, les ressources étant des vues sur le système global. Le schéma médiateur lui, est défini indépendamment des sources qui, elles, sont définies comme des vues sur le schéma médiateur.

2.3.3.1.1 Description de l'approche

Dans cette approche, lorsque les schémas locaux et le schéma global sont des ontologies comme dans notre cas, l'approche fait correspondre à chaque concept C_L de l'ontologie locale une vue V_G sur l'*ontologie globale*.

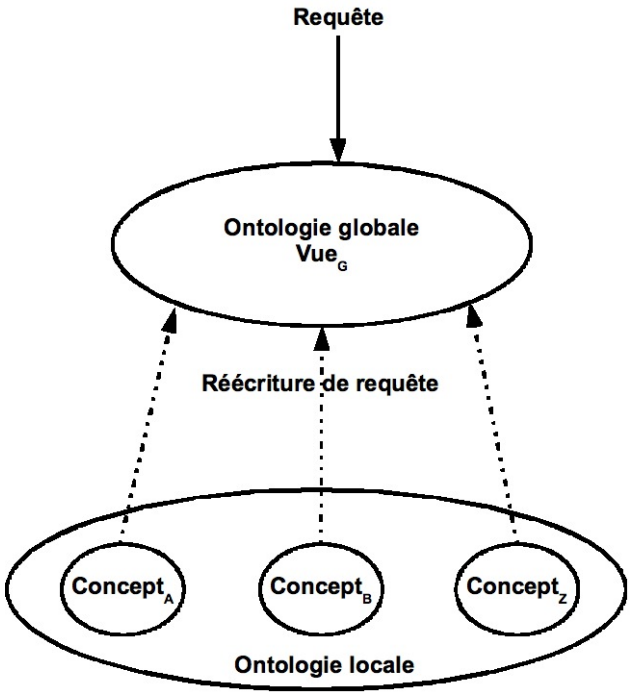


Figure 2-5 Modèle structurel de l'approche LAV

Une requête qui est exprimée en fonction des concepts de l'*ontologie globale*, est réécrite en termes de vues pour être envoyée aux différentes sources concernée. Dans le cas du *SIC-Sénégal* une discrimination est faite entre les vues en fonction des dimensions spatio-temporelles de la requête exprimée.

En approche *LAV*, le mapping $\mathcal{M}$ du système d'intégration est l'ensemble des assertions associant à tout élément s du schéma global S, une requête ou vue $q_{\mathcal{G}}$ sur le schéma global $\mathcal{G}$ de la forme :

$$s \zeta q_{\mathcal{G}}$$

En approche *LAV*, le mapping $\mathcal{M}$ du système d'intégration est un ensemble de triplets (AS, Q(AG), ζ) avec ζ est un symbole de l'alphabet $\{<, \leq, \cong, \geq, >\}$.

Reprenons l'exemple fourni précédemment d'un modèle d'intégration dont le schéma $\mathcal{G}0$ fait référence aux critiques de relecteurs d'articles en soumission à des conférences en informatique,

$$I_0 := \{\mathcal{G}_0, S_0, \mathcal{M}_0\} \text{ où :}$$

1. $\mathcal{G}:=\{article\ (Titre,\ Année,\ Conférence),\ informatique(Conférence),\ reviewer(Titre,\ Critique)\}$ est le schéma global, donc l'alphabet $\mathcal{A}_{\mathcal{G}0} := \{article, informatique, reviewer\}$;
2. Le schéma source $S_0:=\{S_0^i,\ 1 \leq i \leq 2\}$ a son alphabet A_{S0} constitué de 2 sources S_0^1 et S_0^2 :

- la source sources S_0^1 d'arité 3 contient les informations sur les publications selon le schéma *r1(Titre, Année, Conférence)*, les titres de publications et conférences depuis 1994 ;
- La source S_0^2 d'arité 2 contient les informations selon le schéma *r2(Titre, Critique)* depuis 2001.

3. Selon le principe de l'approche *LAV*, à chaque relation ou vue des sources de S_0, M_0 associe par la relation ζ une vue sur le schéma global, donc $M_0:=\{M_0^i, 1 \leq i \leq 2\}$ défini par :
   - $M_0^1 : r1(t, a, c) \zeta \{(t, a, c) \mid article(t, a, c), informatique(d), y \geq 1994\};$
   - $M_0^2 : r2(t, r) \zeta \{(t, r) \mid article(t, a, c), reviewer(t, r), a \geq 1994\}.$

Dans la figure 2.5, les flèches indiquent que les vues des schémas de sources sont définies sur des vues du schéma médiateur. Cette architecture illustre les avantages de cette approche qui permet de décrire les sources indépendamment les unes des autres et il est aussi très simple de rajouter une nouvelle source. Cependant, elle présente un inconvénient majeur : la traduction de requêtes est un processus complexe.

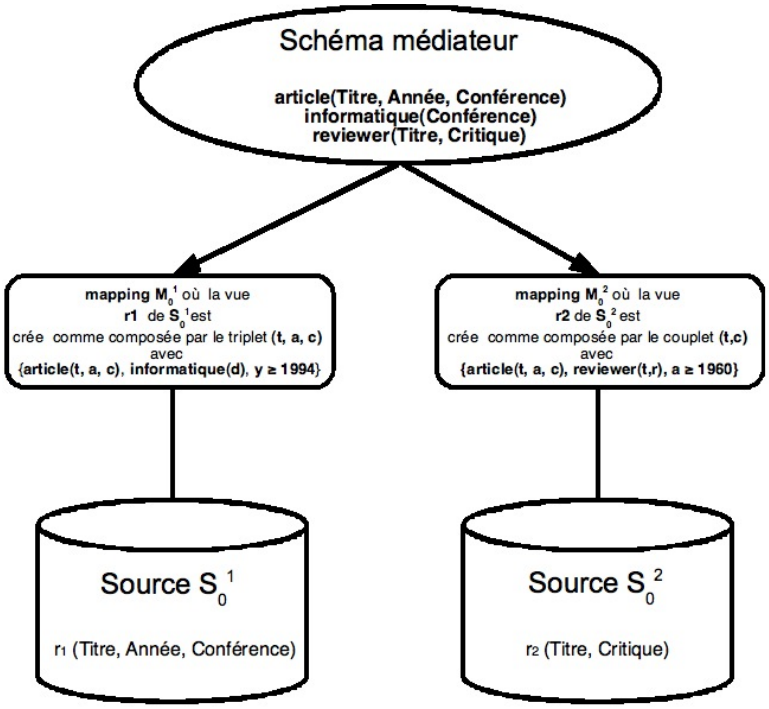


Figure 2-6 Exemple construction d'un schéma médiateur en approche LAV

C'est une approche qui est surtout utilisée lorsque l'on a beaucoup de sources à intégrer. Cette approche a été utilisée dans plusieurs projets comme PICSEL[5] qui est à sa

[5] http://www.lri.fr/ sais/picsel3/index.php

troisième version, Ontobroker[6] dont une version commerciale est actuellement disponible. Nous allons en étudier l'architecture et les composantes conceptuelles de son système d'intégration.

2.3.3.1.2 Prototypes utilisant l'approche

2.3.3.1.2.1 Le prototype Styx

STYX [Fundulaki et al., 2002], [Amann et al., 2002(a)] est un médiateur pour communautés web en vue de concevoir un modèle de médiation plus riche permettant de décrire des ressources XML dans les termes d'une ontologie composée de concepts ainsi que de rôles [Amann et al., 2002(a)]. STYX permet de décrire des ressources XML en termes d'une *ontologie globale*. Il a été développé en tant que partie du projet C-web (la Communauté Web) dont l'objectif principal est de soutenir le partage, l'intégration et la récupération d'informations dans les communautés Web au sujet d'un domaine spécifique d'intérêt [Fundulaki et al., 2002].

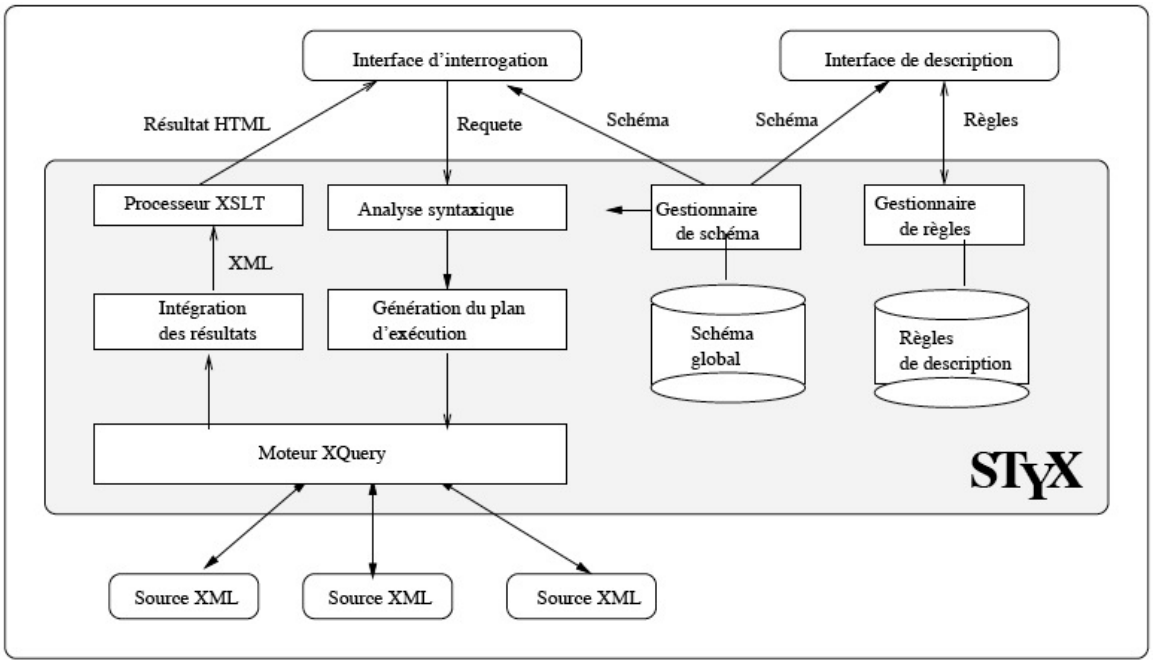


Figure 2-7 Architecture du médiateur STYX

Le schéma global est décrit comme une ontologie, et les ressources XML sont décrites comme des vues du schéma global [Amann et al., 2002(b)]. Dans Styx l'approche tire profit de la présence de DTDs XML (Document Type Definitions) qui capturent la structure du document XML. Une source XML est éditée dans un portail STYX par un ensemble de règles

[6] http://www.ontoprise.de/content/index.html

de mapping qui effectuent la correspondance entre les fragments de documents XMLs indiqués par les chemins de localisation de XPath aux chemins de localisation de l'ontologie.

Les requêtes sont réécrites à partir des termes du schéma global en une ou plusieurs requêtes XML sur les sources locales. Le plan de requête est généré automatiquement avec une possibilité de décomposer une requête en plusieurs requêtes sur des sources multiples. L'utilisateur interroge l'ensemble des sources en formulant des requêtes simples selon les chemins de l'ontologie. L'objectif est de pouvoir envoyer les requêtes de l'utilisateur aux diverses sources XML tout en cachant leur hétérogénéité structurelle à l'utilisateur [Amann et al., 2002(b)].

2.3.3.1.2.2 Le projet PICSEL

PICSEL (Production d'Interfaces à bases de Connaissances pour des Services En Ligne) est un modèle d'intégration de sources XML distribuées avec un système de médiation centralisé basé sur les DTD. Dans cette solution, la description du contenu des sources permet de représenter un grand nombre de structures XML. Les descriptions abstraites des contenus permettent de localiser des sources appropriées.

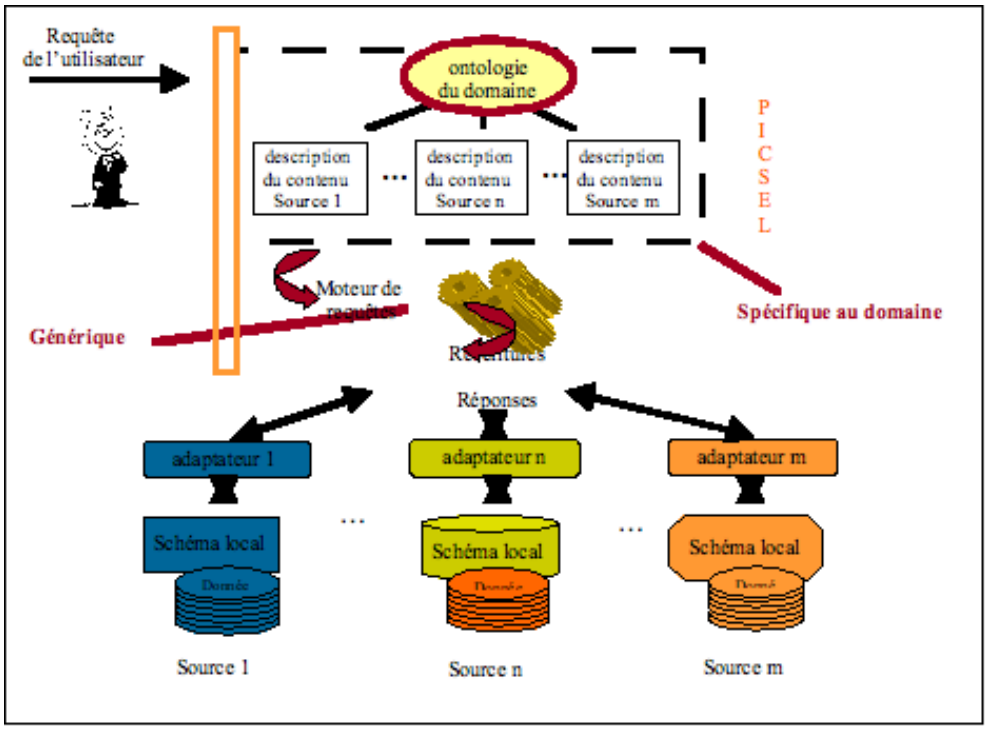


Figure 2-8 Architecture du médiateur PICSEL[7]

La figure 2.8 illustre l'architecture générale de PICSEL. Un moteur de requête prend en charge l'accès aux sources afin d'obtenir les réponses aux requêtes de l'utilisateur. Pour

[7] Http://www.lri.fr/ sais/picsel3/Architecture.php

une requête globale Q : décomposition en un ensemble de requêtes locales fournissant l'ensemble de toutes les réponses possibles à Q. Le système repose sur la construction d'une ontologie de domaine et ce de façon manuelle pour la description du contenu des sources.

Les DTDs associées à des documents XML sont vues comme des descriptions conceptuelles d'ontologies locales et spécialisées dans un domaine d'application. Elles construisent alors une *ontologie globale* et formelle unifiant un ensemble de DTDs relatives à un même domaine d'application. Cette approche centralisée est généralisée afin d'établir des correspondances sémantiques entre ontologies (sans passer nécessairement par une *ontologie globale* unificatrice).

2.3.3.2 Approche Global-As-View

2.3.3.2.1 Description de l'approche

En approche *Global-as-View* ou *(GAV)*, le schéma global devient une vue sur les sources. Elle consiste à définir le schéma médiateur comme un ensemble de vues des sources, ce sont donc ici les vues qui définissent le schéma médiateur. Pour chaque relation dans le schéma global, on définit une vue composée de termes des relations des sources.

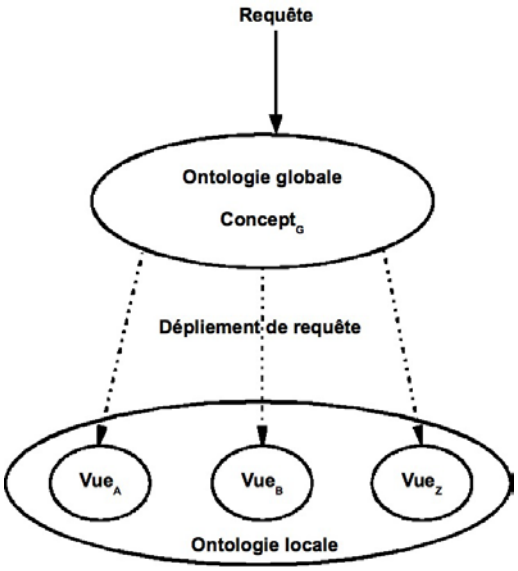


Figure 2-9 Modèle structurel de l'approche GAV

Dans cette approche, lorsque les schémas locaux et le schéma global sont des ontologies comme dans notre cas, l'approche *GAV* fait correspondre à chaque concept C_G de l'*ontologie globale* une vue $\mathcal{V}_L$ sur l'ontologie locale, comme le montre la figure 2.9.

En approche *GAV*, le mapping $\mathcal{M}$ du système d'intégration est obtenu en associant à tout élément g du schéma global g une requête q_S sur le schéma des sources $\mathcal{S}$, le langage de

requête $\mathcal{L}_{\mathcal{M},\mathcal{G}}$ basé sur l'alphabet $\mathcal{A}_{\mathcal{G}}$. En approche *GAV*, un mapping est un ensemble d'assertions, une pour tout élément g de $\mathcal{G}$ de la forme $q_S \zeta$ g.

$$g \zeta q_S$$

Donc en résumé, le mapping dans une approche *GAV* est un ensemble de triplets ($\mathcal{A}_{\mathcal{G}}$, $Q(\mathcal{A}_S)$, ζ) ζ est un symbole de l'alphabet $\{<, \leq, \cong, \geq, >\}$.

Reprenons l'exemple fourni précédemment d'un modèle d'intégration dont le schéma $\mathcal{G}_0$ fait référence aux critiques de relecteurs d'articles en soumission à des conférences en informatique,

$I_0 := \{\mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0\}$ où :

— $\mathcal{G}:=\{article\ (Titre,\ Année,\ Conférence),\ informatique(Conférence),\ reviewer(Titre,Critique)\}$ est le schéma global, donc l'alphabet $\mathcal{A}_{\mathcal{G}0} := \{\ article,\ informatique,\ reviewer\}$;

— Le schéma source $\mathcal{S}_0 := \{S_0^i,\ 1 \leq i \leq 2\}$ a son alphabet $\mathcal{A}_S{}^0$ constitué de 2 sources $\mathcal{S}_0{}^1$ et $\mathcal{S}_0{}^2$:

- la source sources $\mathcal{S}_0{}^1$ d'arité 3 contient les informations sur les publications selon le schéma $r1(Titre,\ Année,\ Conférence)$, les titres de publications et conférences depuis 1994 ;
- La source $\mathcal{S}_0{}^2$ d'arité 2 contient les informations selon le schéma $r2\ (Titre,\ Critique)$ depuis 2001.

— Selon le principe de l'approche *GAV*, à chaque relation du schéma global $\mathcal{G}_0$, $\mathcal{M}_0 := \{\mathcal{M}_0{}^i\ ,\ 1 \leq i \leq 3\}$ associe par la relation ζ une vue sur les sources, donc $\mathcal{M}_0$ sera défini par ses composantes :

1. $\mathcal{M}_0{}^1 : \{(t, a, c) \mid r1(t, a, c)\}\ \zeta\ article(t, a, c)\}$
2. $\mathcal{M}_0{}^2 : \{(c) \mid r1(t, a, c)\}\ \zeta\ informatique(c)\}$
3. $\mathcal{M}_0{}^3 : \{(t, r) \mid r2(t, r)\}\ \zeta\ reviewer(t, r)\}$

Dans la figure 2.9 illustrant l'architecture conceptuelle de l'exemple ci-dessus, les flèches indiquent que le schéma médiateur est défini (sur), à partir des vues sur les sources. La figure 2.10 illustre la définition du système de mapping entre les deux principales composantes du système d'intégration.

L'approche *GAV* présente l'avantage d'être une approche naturelle et la traduction de requêtes se fait facilement par une « expansion » de la requête dans la vue. Elle présente l'inconvénient de nécessiter une modification du modèle global lorsqu'une nouvelle source est ajoutée. Il faut considérer l'interaction de la nouvelle source avec les autres. Cette

approche a été l'objet dans plusieurs expérimentations parmi lesquelles *TSIMMIS (The Stanford-IBM Manager of Multiple databases)*[8]*, SIMS, MOMIS et MIEL++*.

SIMS [Arens et al., 1996] vise l'intégration de bases de connaissances et sources de données hétérogènes qu'il représente en utilisant un langage basé sur la logique de description. Le médiateur dans *SIMS* est spécialisé dans un seul domaine d'application.
MOMIS (Mediator environment for Multiple Information Sources) [Beneventano et al., 2000] est également un environnement d'intégration utilisant une *ontologie globale* appelée *GVV (Global Virtual View)* générée semi-automatiquement. Le mapping entre l'*ontologie globale* et les sources locales est réalisé grâce à l'approche *GAV*.

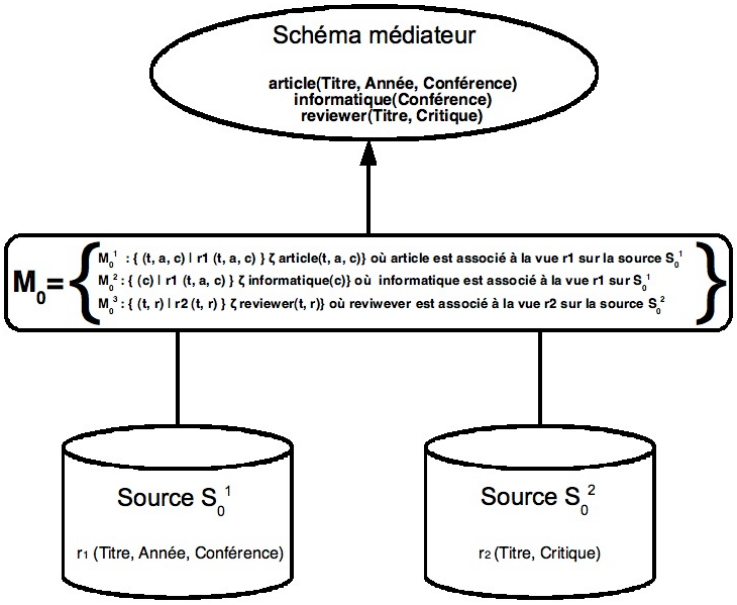


Figure 2-10 Exemple de construction d'un schéma médiateur en approche GAV

2.3.3.2.2 Prototypes utilisant l'approche

Nous présentons dans cette partie projets utilisant l'approche *Global-As-View*.

2.3.3.2.2.1 Projet TSIMMIS

Le projet TSIMMIS (The Stanford-IBM Manager of Multiple databases) avait pour but de développer des outils qui facilitent l'intégration d'informations hétérogènes provenant de données structurées et semi-structurées.

Dans TSIMMIS c'est l'approche *GAV* qui est utilisée avec le modèle semi-structuré OEM (Object Exchange Model) permettant de définir les objets sous forme de quadruplet : <id, nom, type, valeur> et le langage de spécification de médiateur MSL permettant de définir des règles.

[8] http://infolab.stanford.edu/tsimmis/

Dans TSIMMIS, chaque source d'information est associée à un adaptateur permettant de convertir les données de la source dans le modèle de données commun OEM. Au niveau médiateur, chaque médiateur obtient les informations d'un ou plusieurs adaptateurs ou d'autres médiateurs et affine cette information par intégration et résolution de conflits entre les données extraites des différentes sources, et fournit l'information résultante à l'utilisateur ou à d'autres médiateurs [Lo, 2002].

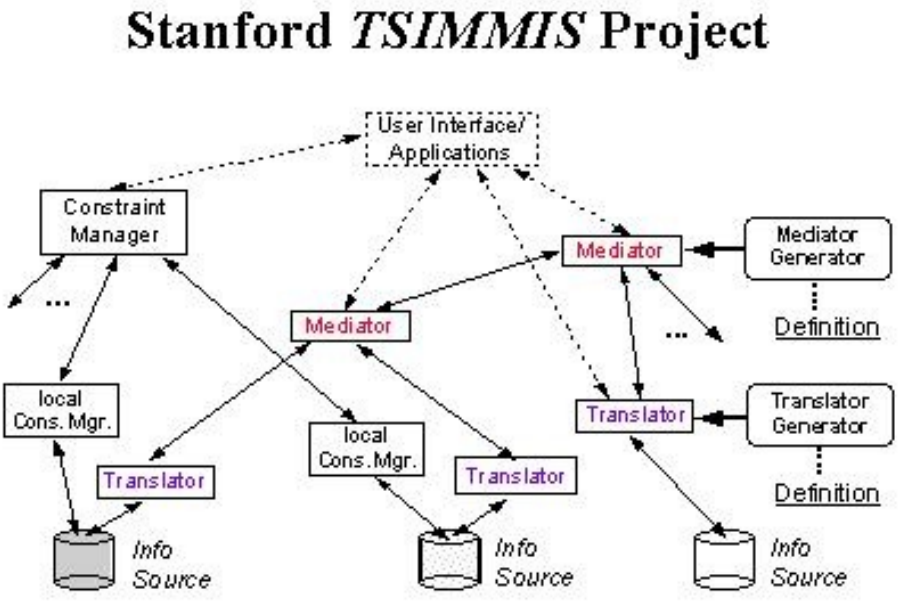


Figure 2-11 Architecture de TSIMMIS

Chaque source d'information est associée à un adaptateur permettant une conversion des données de la source dans le modèle de données commun OEM.

2.3.3.2.2.2 Projet MIEL++

MIEL++[Buche et al., 2005] est un sous projet de E.DOT-RNTL[9], un moteur d'interrogation, pour permettre à un utilisateur d'interroger différentes sources de données portant sur la même thématique.

Le modèle d'intégration choisi est un modèle de type *Global as View* (ou *GAV*) dans lequel le schéma médiateur est défini en fonction des schémas des sources à intégrer : les différentes sources de données sont connues au moment de l'élaboration du schéma médiateur qui est en fait une vue globale sur les schémas locaux.

L'exécution d'une requête consiste alors à remplacer les vues utilisées dans la requête d'origine par leur traduction du point de vue des bases locales. L'extraction et l'intégration des données sont dirigées par une ontologie.

[9] http://gemo.futurs.inria.fr/projects/edot/

L'entrepôt thématique est construit avec des données provenant des tables de documents HTML ou PDF. Les tâches qui sont spécifiques au traitement d'un format d'une source sont explicitement séparées des tâches indépendantes de n'importe quelle source. Ainsi, la transformation automatique des diverses tables en représentation générique de XML appelée XTAB est d'abord effectuée. En raison de l'hétérogénéité sémantique parmi des sources, extraire des données à partir des pages web est insuffisant pour l'intégration de données. Les données doivent être organisées d'une manière différente avec un vocabulaire différent, c'est-à-dire qu'il est nécessaire de trouver une représentation XML où la plupart des valeurs et des étiquettes appartiennent à l'ontologie.

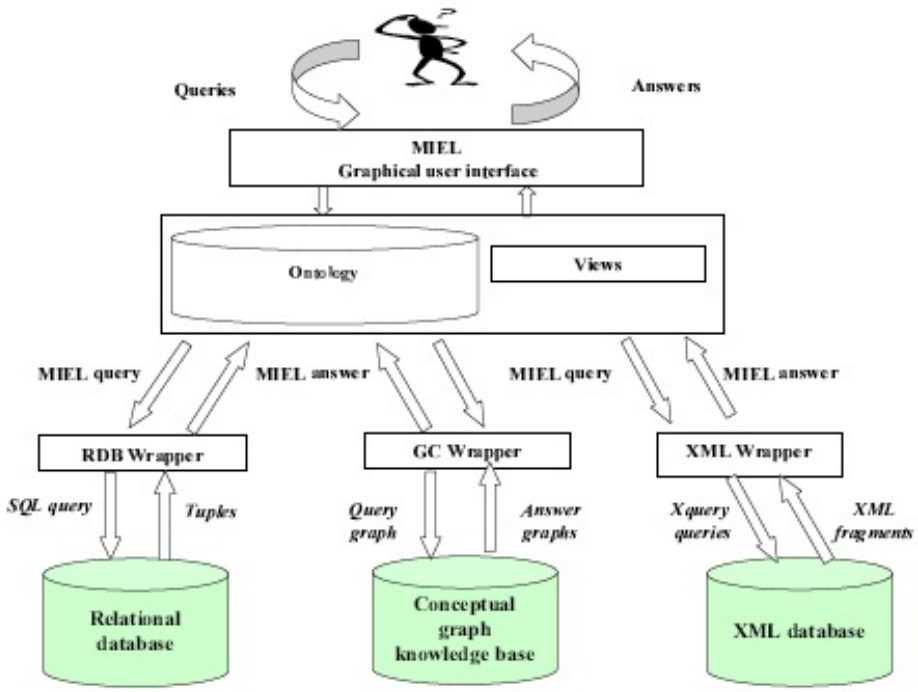


Figure 2-12 Architecture de MIEL++

Dans l'approche, la transformation a été faite de sorte qu'elle soit la plus automatique et flexible possible en étant seulement dirigée par l'ontologie et la manière dont les données ont été structurées dans la table originale. Ainsi, une DTD appelée SML (Semantic Markup Language) été définie. Elle peut automatiquement être générée en utilisant l'ontologie et peut traiter une information additionnelle ou incomplète dans une relation sémantique, des ambiguïtés ou des erreurs d'interprétation. Cette transformation a été implémentée et expérimentée sur des données du projet E.DOT.

Les données extraites sont exploitées en utilisant le moteur d'interrogation. Cette interrogation s'effectue via des requêtes composées exclusivement de termes issus de l'une

des ontologies (Sym'Previus) alors que les réponses attendues sont composées de documents annotés par les termes appartenant aux deux ontologies.

L'approche de mise en correspondance proposée s'exécute en deux temps. D'abord, des mappings dits « probables » sont automatiquement découverts. Ensuite, des « indicateurs» sont proposés ainsi que des mappings potentiels pour aider l'expert du domaine à mettre en correspondance les éléments pour lesquels un mapping probable n'a pu être trouvé automatiquement. La méthode d'alignement proposée repose sur des techniques variées appliquées séquentiellement : terminologiques, structurelles et sémantiques.

2.3.3.3 La médiation distribuée ou hybride

2.3.3.3.1 Description de l'approche

Les projets en contexte de médiation distribuée appliquent le principe de l'approche hybride ou *Global Local As View* (G*LAV*). Dans le contexte d'une approche hybride, il n'y a pas de schéma médiateur à ajouter, mais les correspondances entre les schémas des différentes sources doivent être définies.

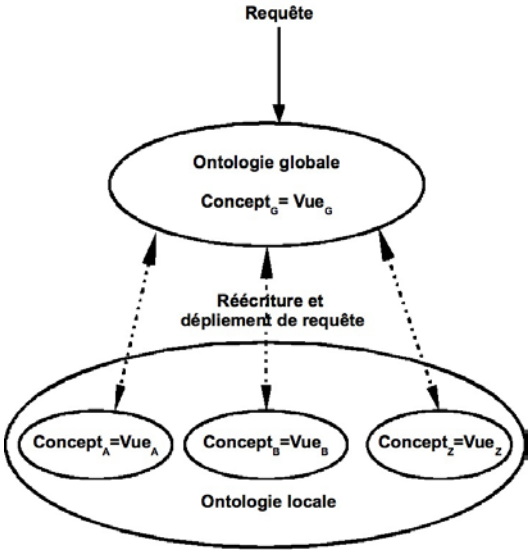


Figure 2-13 Modèle structurel de l'approche GLAV

Ces correspondances, selon les besoins, peuvent être plus moins complexes nécessitant de combiner les approches *Global-as-View* et *Local-as-View* (G*LAV*) pour permettre l'assertion d'inclusions de requêtes combinant des relations provenant de différentes sources[10]. Dans cette approche, lorsque les schémas locaux et le schéma global sont des

[10] Http://www.lirmm.fr/ libourel/MM/PourBDAAS2.doc

ontologies comme dans notre cas, l'approche G*LAV* fait correspondre à chaque concept $\mathcal{V}_G$ de l'*ontologie globale* une vue $\mathcal{V}_L$ sur l'ontologie locale.

En approche G*LAV*, le mapping $\mathcal{M}$ du système d'intégration est un ensemble d'assertions, associant à toute requête q_S sur le schéma des sources $\mathcal{S}$ une requête q_G du schéma global. En approche G*LAV*, un mapping est un ensemble d'assertions de la forme

$$q_S \; \zeta \; q_G$$

Avec q_S une requête sur $\mathcal{S}$ et q_G une requête sur $\mathcal{G}$ de même arité et ζ est un symbole de l'alphabet $\{<, \leq, \cong, \geq, >\}$.

Cette approche a été expérimentée dans plusieurs projets comme XYLEME, XLIVE, ActiveXML.

Reprenons l'exemple fourni précédemment d'un modèle d'intégration dont le schéma $\mathcal{G}_0$ fait référence aux critiques de reviewers d'articles soumis à des conférences en informatique,

$$I_0 := \{\mathcal{G}_0, \mathcal{S}_0, \mathcal{M}_0\} \text{ où :}$$

— $\mathcal{G}:=\{article\ (Titre, Année, Conférence),\ informatique(Conférence),\ reviewer(Titre, Critique)\}$ est le schéma global, donc l'alphabet $\mathcal{A}_{G0} :=\{ article, informatique, reviewer\}$;

— Le schéma source $\mathcal{S}_0:=\{S_0^i\ ,\ 1 \leq i \leq 2\}$ a son alphabet $\mathcal{A}_{S0}$ constitué de 2 sources $\mathcal{S}_0^1$ et $\mathcal{S}_0^2$:

- la source sources $\mathcal{S}_0^1$ d'arité 3 contient les informations sur les publications selon le schéma $r1(Titre, Année, Conférence)$, les titres de publications et conférences depuis 1994;
- La source $\mathcal{S}_0^2$ d'arité 2 contient les informations selon le schéma $r2(Titre, Critique)$ depuis 2001.

— Selon le principe de l'approche G*LAV*, à chaque requête q_{S0} sur le schéma des sources $\mathcal{S}_0$, une requête q_{G0} du schéma global lui est associé par la relation ζ, donc $\mathcal{M}_0 :=\{ \mathcal{M}_0^i,\ 1 \leq i \leq 2\}$ sera défini par ses composantes :

1. $\mathcal{M}_0^1:\{ (t, a, c) \mid r1(t, a, c)\}\ \zeta\ \{(t, a, c) \mid article(t, a, c), informatique(d), a \geq 1994\}$
2. $\mathcal{M}_0^2:\{ (t, r) \mid r2(t, r)\}\ \zeta\ \{(t, r) \mid article(t, a, c), reviewer(t,r), a \geq 1994\ \}$

2.3.3.3.2 Prototypes utilisant l'approche

2.3.3.3.2.1 XYLEME

Xyleme[11] est un système d'intégration physique (approche entrepôt) utilisant XML comme modèle de données. Les ontologies globales et locales sont exprimées à l'aide d'arbres, avec un mapping *GAV/LAV*. Le mapping est réalisé semi automatiquement par la génération de tables de correspondance entre les chemins de l'*ontologie globale* et les chemins des ontologies locales.

Dans Xyleme, l'approche choisie est celle d'une méthode d'intégration hybride, c'est-à-dire que la matérialisation des données est effectuée en utilisant une approche hybride constituant un mélange avec l'approche dite virtuelle ou médiateur et celle de type entrepôt mais aussi G*LAV*. A l'origine, l'objectif du projet était de mettre en place un moteur de recherche XML. Les données sources sont stockées dans le format XML (sans être transformées) et intégrées à travers un mécanisme de vues entre les DTD concrètes des sources et des DTD abstraites montrées à l'utilisateur. L'architecture fonctionnelle de Xyleme donnée par la figure 2.14.

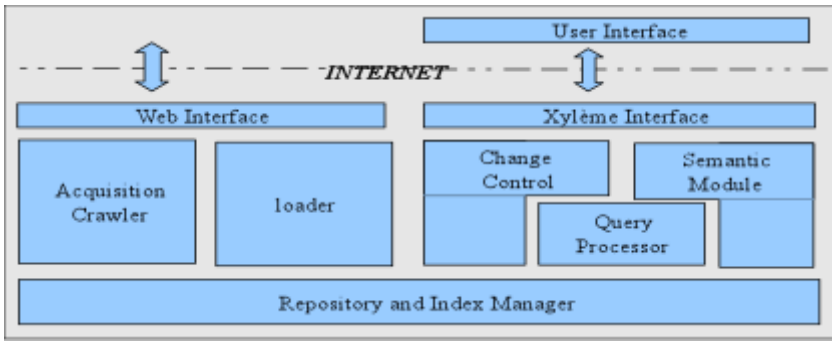


Figure 2-14 Architecture fonctionnelle de XYLEME

XML possède une dimension sémantique faible basée sur les balises. Pour un même type de contenu, il est possible de retrouver des définitions de balises différentes, d'où la nécessité d'y apporter de la sémantique. Pour résoudre ce problème Xyleme se base sur une structure de document XML idéale par rapport aux types de contenus recherchés (spécifiée dans le contexte de l'entreprise) ainsi que sur une correspondance entre cette structure idéale et les multiples structures réelles des sources XML.

Les rapprochements effectués entre le contenu et la structure du document XML idéal et ceux des multiples documents XML réels sont réalisés à l'aide de dictionnaires métiers et d'algorithmes d'intelligence artificielle.

[11] Http://www.xyleme.com/

Dans Xyleme, l'intégration est organisée autour de domaines, un domaine étant défini comme un ensemble de clusters. Ces domaines ne sont pas forcément disjoints et la structure de chacun est décrite par une DTD abstraite.

Une DTD abstraite, du point de vue structurel est un arbre de concepts qui a les fonctions suivantes:

- la définition d'un domaine sémantique ;
- la description de clusters de DTD concrètes ;
- une interface d'interrogation unique pour documents hétérogènes ;
- un schéma de structuration du résultat.

2.3.3.3.2.2 ActiveXML

ActiveXML a pour objet l'intégration de données et de services Web. Le système ActiveXML (ou AXML) est un système pair-à-pair fondé sur le modèle de documents intentionnels pour la gestion de données distribuées.

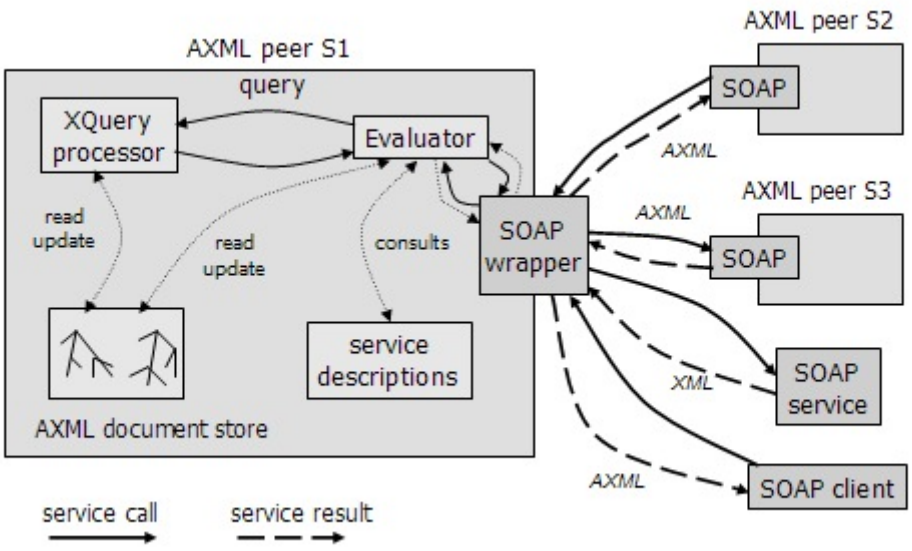


Figure 2-15 Architecture d'ActiveXML

L'entrepôt associé à un pair évolue avec la matérialisation des appels de services qu'il contient. En particulier, un pair peut appeler les services d'autres pairs AXML qui retournent des documents intentionnels (les paramètres d'un appel de services peuvent également être des documents intentionnels). Le paradigme repose sur la notion de P2P. L'originalité principale du modèle AXML (documents intentionnels partagés dans un environnement pair-à-pair (P2P)) est que la sémantique d'un document intentionnel est indépendante du processus de matérialisation des appels de service qu'il contient : un appel de service est à chaque

instant t considéré à la fois comme un appel de fonction et comme le résultat de cet appel à cet instant. Cette dualité appel/donnée peut être exploitée pour modéliser différentes techniques d'interrogation [Abiteboul et al., 2004], de réplication [Abiteboul et al., 2003] et d'échange [Milo et al., 2003] de données distribuées.

Plus précisément, chaque pair AXML contient un entrepôt de documents intentionnels et propose un ensemble de services définis sous forme de requêtes sur ces documents.

L'importance de cette approche se verra surtout dans la seconde partie de ce travail qui consiste à étudier les techniques d'intégration sémantique des applications avec l'utilisation des web services sémantiques.

2.3.3.3.2.3 XLive

XLive [Dang-Ngoc et al., 2004b] est un médiateur tout XML permettant d'exécuter des requêtes XQuery sur des sources de données hétérogènes.

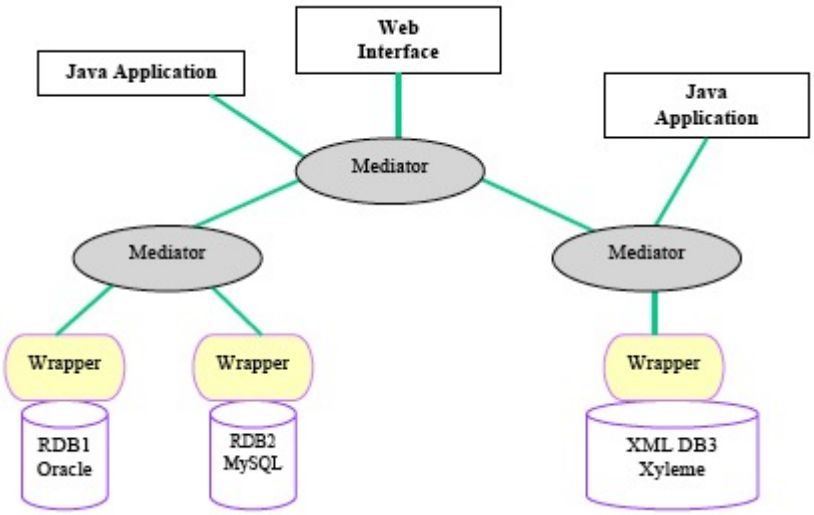


Figure 2-16 Architecture d'XLive

La sémantique de chaque source est décrite par une ontologie qui lui est associée et la sémantique des données intégrées est définie par des ontologies de domaine. Le mapping est effectué en associant à chaque ontologie une vue XML définie en XQuery. C'est un langage utilisé pour l'expression des requêtes aux médiateurs et adaptateurs (wrappers) et RDF pour l'annotation sémantique des documents à partir de termes d'une ontologie. Pour la définition des ontologies, c'est le langage OWL qui est utilisé.

La communication entre les adaptateurs et les médiateurs se fait par l'intermédiaire d'une interface commune, définie par une application java ou un service web XML/DBC. Avec cette technologie les requêtes sont définies en XQuery et les résultats sont retournés au

format XML. Pour son prédécesseur E-XMLMedia[12], basé sur l'approche *Global As View*, un entrepôt XML est construit pour le stockage des documents XML dans une base de données relationnelle, avec une interrogation basée sur XQuery et SQL : XML/DBC.

Le médiateur comporte: une console d'administration pour la configuration du médiateur, l'enregistrement de nouveaux adaptateurs avec en plus :

— un gestionnaire de métadonnées pour les mappings entre les schémas XML et les schémas relationnels;

— un analyseur/décomposeur pour la traduction des requêtes et l'élaboration de plans d'exécution distribués et l'optimisation;

— un évaluateur pour la délégation de sous-requêtes aux adaptateurs, collection des résultats ainsi que l'évaluation de la partie globale du plan (agrégats, jointures) en mémoire, reconstruction du résultat. Enfin un gestionnaire des connexions avec les adaptateurs.

Le médiateur de E-XmlMedia nommé e-XML Mediator se connecte aux sources de données via des Wrappers qui assurent : la traduction des données du format natif de chaque source vers XML, la traduction de la requête XML vers le langage de requête natif de la source. Des wrappers génériques permettent d'accéder à des sources SQL, e-XML Repository et HTML. Une interface ouverte permet de développer des wrappers spécifiques pour d'autres types de sources.

2.3.4 Approche dataweb basée sur XML

Dans cette section, nous allons nous intéresser à l'*approche dataweb qui s'appuie*sur ISYWEB, une architecture de système d'information pour le web proposée dans [Lo, 2002] , [Lo et Hocine, 2005]. L'approche s'articule autour d'un *entrepôt de documents XML* appelé *dataweb* [Hocine et Lo, 2000].

Nous présentons d'abord l'*approche dataweb* pour la résolution de l'hétérogénéité des données ainsi que le modèle basé sur XML sur lequel elle repose. Ensuite, nous situons l'*approche dataweb* dans la vague web sémantique actuelle. Depuis quelques années, des propositions sont faites visant à revoir la philosophie web sémantique pour la réorienter vers un concept de mise en ligne des données disponibles et un niveau d'abstraction qui descend du niveau de la page vers celle des données. Cette approche fait référence, à la philosophie dite des datahubs ou celle de la notion d'hyperdata.

[12] Http://www.e-xmlmedia.fr/

2.3.4.1 Dataweb basé sur XML pour l'intégration de données

L'approche dataweb est basée sur la notion d'unité d'information. Ce concept permet de modéliser l'ensemble des données d'un dataweb provenant d'une même source d'informations, et de définir ainsi un dataweb comme un ensemble d'unités d'informations. L'approche de construction d'un dataweb que nous allons exposer permet l'intégration dans une base de documents XML appelée source globale du dataweb des données issues de sources d'informations structurées et semi-structurées. Cette source globale est associée à un catalogue de méta-informations sur les données ; ce catalogue permet la manipulation du dataweb.

Selon la taxonomie consacrée aux types d'hétérogénéité vus dans le chapitre 1, nous avons constaté qu'il existe globalement deux niveaux d'hétérogénéité dans les systèmes d'informations: un niveau sémantique lié à la représentation du sens associé à la description des données et un niveau structurel relatif au format de représentation des données. Dans ce contexte, nous distinguons deux approches d'intégrations: l'approche entrepôt et l'approche virtuelle.

L'approche virtuelle laisse les données telles quelles dans les sources et procède à une migration des requêtes, mais elle ne facilite ni la mise en forme des données intégrées pour leur diffusion sur le web, ni l'ajout d'un module de recherche d'informations pertinentes [Lo, 2002]. Dans le contexte applicatif des données environnementales marqué par une collection des données en vue de leur exposition de manière ouverte sur le web pour répondre à des requêtes, l'approche virtuelle s'avère difficile à mettre en œuvre. Les données étant collectées pour suivre leur évolution de certaines caractéristiques ciblées dans l'environnement, l'approche la plus pertinente consiste à les migrer au même format sur un seul et même endroit. L'approche entrepôt adopte la même démarche en migrant les données au même endroit, facilitant ainsi le traitement personnalisé des données, l'archivage et leur interrogation.

En plus des avantages hérités de l'approche entrepôt pour résoudre le problème de l'intégration des données, l'*approche dataweb* propose d'utiliser le même format de représentation des données avec le langage XML. L'utilisation d'un même format de représentation du point de vue structurel permet de résoudre la problématique de l'hétérogénéité structurelle des données et facilite aussi leur échange. Le modèle proposé repose sur une source globale constituée d'un ensemble de documents XML qui est associé à un catalogue de méta-informations sur les données permettant la manipulation du *dataweb*.

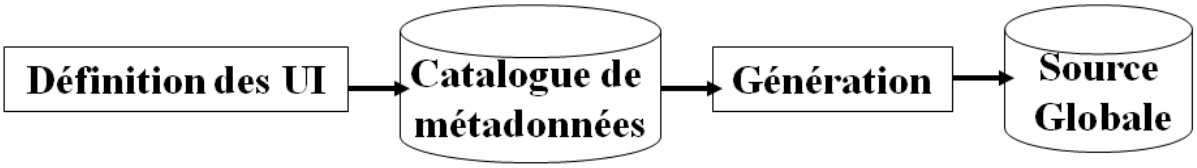


Figure 2-17 Processus de construction du dataweb (extrait de [Lo, 2002])

Pour représenter les entités composant le document la notion d'unité d'information est utilisée. Elle permet de modéliser l'information et la méta-information d'une source de données intégrée dans le *dataweb*. Un *dataweb* basé sur XML est ainsi défini comme un ensemble d'unités d'informations, une unité d'information étant constituée par un document et un méta-document décrit les informations contenues dans le document. Ce processus d'intégration permet ainsi de résoudre la problématique de l'hétérogénéité structurelle des données.

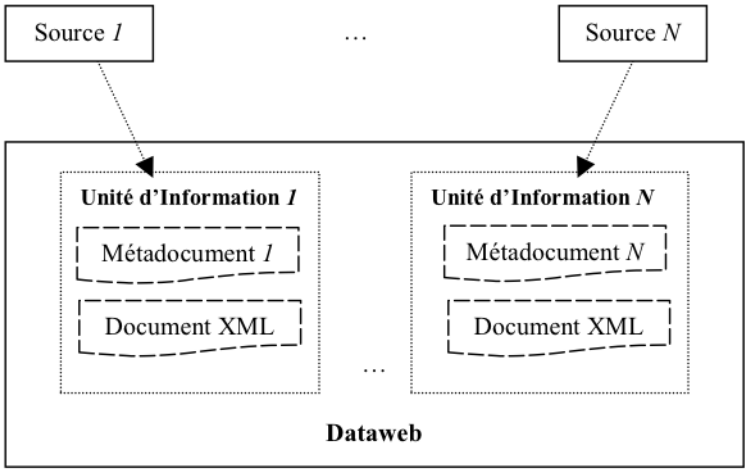


Figure 2-18 Schématisation des unités d'informations d'un dataweb (extrait de [Lo, 2002])

Pour prendre en compte l'aspect sémantique de l'hétérogénéité des données, l'introduction d'une base de concepts composée d'un thesaurus du domaine et d'un catalogue d'index décrivant le contenu sémantique des données est proposée. Le thesaurus du domaine structure les termes d'indexation en hiérarchie de concepts. Sa réutilisation permet la résolution du problème lié à la non expertise des utilisateurs par rapport au domaine.

Dans l'approche que nous proposons, nous substituons cette base de concepts par une *base de connaissances*. Le thesaurus est remplacé par une ontologie du domaine qui offre plus de possibilités.

Cette mise en relation des données et des connaissances est basée sur la notion d'unité élémentaire et d'unité sémantique. Une unité élémentaire est constituée par un nom et l'adresse de l'unité décrite dans le système de stockage des données via Xpath. Elle est un nœud feuille dans le document XML. L'ensemble des unités élémentaires qui n'ont de sens

que lorsqu'elles sont regroupées ensembles sont constituées sous la forme d'une unité sémantique. Ces éléments sont décrits en XML.

Cette approche a été appliquée dans le cadre du projet *SIC-Web AHPA (Système d'Informations et de Connaissances accessibles à travers le web - Anthroposystème et Hydrosystèmes Pyrénéens Atlantiques)*[13] du programme « Environnement Vie et Société (PEVS) »[14]. L'objectif du projet *SIC-Web-AHPA* était offrir un accès uniforme et transparent à l'ensemble des données (hydrologiques, géographiques, économiques, sociologiques, etc.) mises en commun dans le cadre du site atelier.

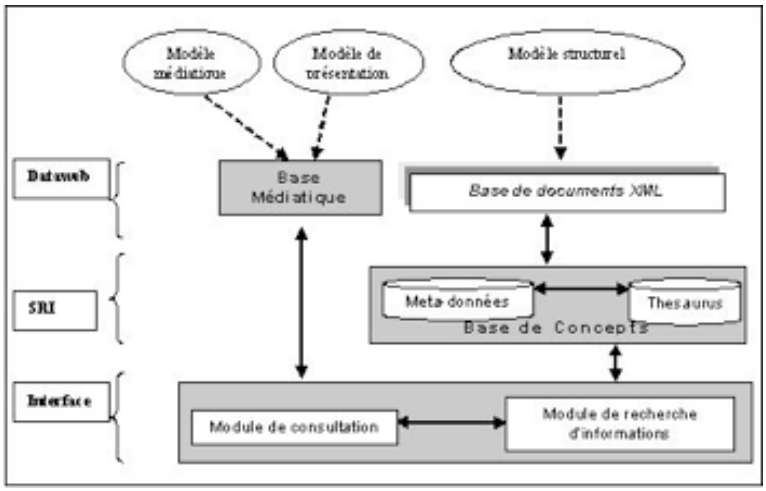


Figure 2-19 Architecture d'un dataweb basé sur XML

L'*approche dataweb* a aussi été appliquée dans le cadre du projet *SIC* Sénégal initié à l'Université de Saint-Louis depuis juin 2004.

Nous réutilisons cette démarche d'apport sémantique en substituant le catalogue d'index par une *base d'annotations* et au thesaurus une ontologie décrivant sémantiquement les données. Les unités élémentaires constituent les attributs des concepts des ontologies que sont les unités sémantiques dans l'*approche dataweb basé sur XML* décrite dans les travaux de Lo et Hocine [Lo, 2002], [Hocine et. Lo, 2000], [Lo et Hocine, 2000], [Lo et Hocine, 2005].

2.3.4.1.1 Composants du modèle

L'*approche dataweb* se résume en deux grandes phases. Une première phase vise à intégrer structurellement les données dans un *entrepôt de documents XML* et une seconde phase permet de constituer la base sémantique décrivant les données. Nous présentons donc le

[13] http://hpa.univ-pau.fr/
[14] http://www.cnrs.fr/cw/dossiers/doseau/recher/program/pevs.html

modèle de *dataweb basé sur XML* en premier et dans la deuxième partie le modèle de structuration sémantique associée par l'intermédiaire de la base de concepts.

2.3.4.1.2 Modèle pour l'intégration structurelle

Un *dataweb* étant décrit comme un ensemble d'unités d'informations, l'ensemble des opérations qui lui sont associées sont donc liées aux traitements des unités d'informations tel leur ajout suppressions ou mise à jour. Il peut donc être modélisé par le couple constitué par son nom ainsi que l'ensemble des unités d'information le constituant.

$$\mathcal{D} := \{\mathcal{N}om, \{\mathcal{U}I_i, i \in \mathcal{N}^*\}\} \text{ où :}$$

— $\mathcal{N}om$ est le nom du *dataweb* ;

— $\{\mathcal{U}I_i, i \in \mathcal{N}^*\}$ est l'ensemble des unités d'information.

Une unité d'information $\mathcal{U}I$ est définie par un document et un méta-document décrivant les informations qui lui sont associées. Elle est définie dans [Lo, 2002] comme le couple :

$$\mathcal{U}I_i := \{\mathcal{D}_i, \mathcal{MD}_i\} \text{ où :}$$

— $\mathcal{D}_i$ est un document XML issu d'une source de données du *dataweb* ;

— $\mathcal{MD}_i$ est un méta-document.

Un méta-document $\mathcal{MD}_i$ permet de décrire entre autres l'origine du document ainsi que la manière d'obtenir des données intégrées de ses sources. Il est défini par le n-uplet :

$$\mathcal{MD}_i := \{n, p, u, v, s\} \text{ où :}$$

— n est le nom de la source;

— p est le propriétaire de la source ;

— u est l'URL de la source ;

— v est la vue permettant d'obtenir les données à intégrer dans le *dataweb* ;

— s est la DTD du document XML d.

2.3.4.1.3 Modèle pour l'intégration sémantique

Le niveau sémantique de l'approche proposée dans [Lo, 2002] est constitué par l'ensemble des unités sémantiques définies pour représenter les concepts qui s'articulent autour d'unités dites élémentaires. Les unités élémentaires sont définies comme des nœuds du graphe XML qui contiennent l'information indexée. Une **unité sémantique (US)** est définie comme un sous-arbre de l'arbre XML dont les nœuds n'ont de sens que s'ils sont considérés dans le contexte du sous-arbre. Ces méta-informations sont extraites de manière automatique et validées par l'intermédiaire d'un expert du domaine.

L'unité sémantique résume sous une conceptualisation l'idée où la construction véhiculée par l'ensemble de ces unités élémentaires lorsqu'elles sont réunies ensemble. Comme on le retrouve dans [Lo, 2002] cette notion exprime le contexte local ou micro-contexte. Nous le définirons dans ce document.

L'unité élémentaire du point de vue structurelle est un nœud feuille d'un document XML exprimant la valeur d'un attribut donné. Sa position dans l'arbre est spécifiée grâce à un chemin Xpath. Donc, nous pouvons déduire de [Lo, 2002] le triplet permettant de représenter une unité élémentaire UE. Elle est définie par le n-uplet:

$UE_i := \{doc, path, valeur\}$ où :

— *doc* est le document d'origine de l'unité d'information;

— *path* est l'expression de son chemin d'accès dans le graphe;

— *valeur* est sa valeur.

2.3.4.2 Variante de l'approche

Une solution d'intégration de données proposées dans [Lima et al., 2003] est semblable à l'*approche dataweb*, avec la migration de toutes les données vers un format unique en XML. Elle permet ainsi de solutionner la problématique de l'hétérogénéité sémantique en migrant les données vers une base de documents XML, donc vers un *dataweb* en utilisant l'approche dite de « médiation » et lui associant une base de métadonnées. Cette approche n'utilise pas les ontologies mais ressemble en termes d'objectif à ceux visés par le projet *SIC*-Sénégal.

Elle permet aussi de prendre en compte la nature spatio-temporelle et descriptive des données. En d'autres termes, ce sont des données qui ont des attributs spatiaux, donc géographiques, collectées selon une période précise et des milieux bien définis, avec des attributs temporels. Ce qui pose donc un problème d'échantillonnage avec l'irrégularité des périodes selon les sources mais aussi l'existence certaine de périodes sans données.

En plus, l'existence de plusieurs sources (entreprises fournissant les données : ADRAO, SAED, DRDR-SL,...) pose la problématique de la distribution des données. On sait que chaque source est libre d'utiliser ses propres formats et schéma de stockage, mais possède des périodes et précisions de collections des données relatives.

C'est pourquoi dans [Lima et al., 2003], une architecture suivant quatre niveaux a été proposée. Elle consiste, au niveau le plus bas, d'avoir des sources de données de différents formats : xls, doc, ascci,...

Une interface permet de passer les données à un module au troisième niveau appelé « module d'intégration de données » dont le rôle est de convertir chaque type de données an format XML grâce à des adaptateurs, de vérifier les problèmes de contraintes, voir si une donnée a déjà été enregistrée ou non dans le *dataweb*. Nous avons aussi à ce niveau un deuxième module dit « module d'évaluation de la qualité des données » qui permet d'effectuer les calculs nécessaires afin d'interpoler sur les données manquantes.

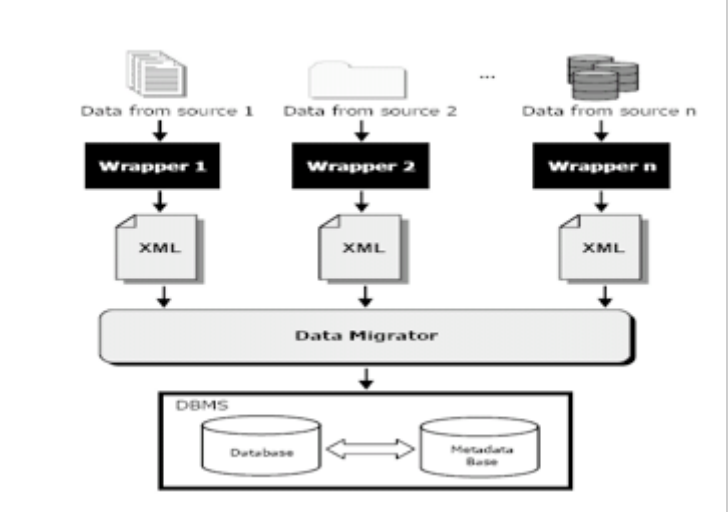


Figure 2-20 Architecture du modèle d'intégration de données

Il sert également à rejeter les données qui ne sont pas bonnes qualitativement selon des critères prédéfinis. Nous avons toujours dans ce niveau un dernier module dit « d'exécution des requêtes » qui va permettre de recevoir la requête de l'utilisateur, de l'envoyer au module d'évaluation de la qualité qui interroge le *dataweb* et évalue la requête selon la qualité des données. Cela est réalisé par une interaction entre le module « d'évaluation de la qualité » et celle de « requête ».

La base de métadonnées associée au *dataweb* est construite en considérant plusieurs standards :

— Content Standards for Digital Geospatial Metadata (CSDGM)[15];

— Federal Geographic Data Committee (FGDC)[16];

— WMO Core Metadata Standard[17];

— World Meteorological Organization (WMO)[18];

— Metadata to Scientific Workflows to Support Environmental Planning [Rocha, 2003] ;

[15] http://www.fgdc.gov/metadata/ontstan.html
[16] http://www.fgdc.gov
[17] http://www.wmo.ch/web/www/metadata/WMO-core-metadata-toc.html
[18] http://www.wmo.ch/web/www/metadata/WMO-core-metadata-toc.html

— Dublin Core Metadata Initiative (DC)[19].

2.3.4.3 Approches *dataweb*, « data web » et celle web sémantique

L'approche web sémantique vise à proposer une nouvelle version du web depuis l'année 2000, cependant différentes raisons freinent sa mise en ouvre. Dans cette partie, nous présentons l'approche web sémantique, les critiques qui lui sont adressées ainsi que les orientations proposées.

2.3.4.2.1 Web Sémantique

L'une des principales motivations des systèmes d'intégration de données est la séparation de la structure des données de la connaissance la décrivant. Dans le contexte du web, l'approche sémantique a préalablement été motivée au début par le besoin de séparation de la présentation de la localisation.

L'accessibilité de l'information quelque soit sa localisation a trouvé une solution à travers les formats de représentations accessibles à distance du début des années 1990. Cette phase est qualifiée de première génération du web. L'un des inconvénients rapidement reprochés au langage HTML est sa propension à trop faire la part belle à la présentation au profit des informations structurelles sur les éléments du document et leurs relations limitées à des balises n'offrant aucune possibilité à l'utilisateur de structurer ses propres informations. D'où, le besoin de séparer la structure de la présentation des données contenues dans le document HTML.

Cette démarche de séparation de la présentation de la structure a donné naissance à la norme XML et ses technologies à la fin des années 1990. Un langage séparant la structure de la présentation est conçu pour la description des données. Les balises n'étant pas prédéfinies, l'utilisateur peut définir et structurer ses propres balises en leur associant un modèle décrivant la grammaire de sa structure avec une DTD (Document Type Definition) ou un schéma XML (XML-Schema). Cela permet ainsi à toute application tierce de vérifier la conformité ou non d'un document en vue de son stockage, partage ou échange par rapport à une structure prédéfinie.

[19] http://purl.org/DC

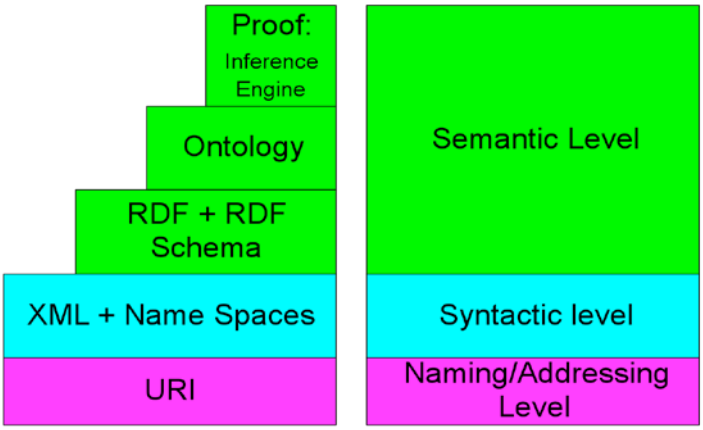


Figure 2-21 Architecture du web sémantique

De même, XML sert à l'interopérabilité du point de vue échange des données. Cependant, cette facilité d'emboîtage des balises n'a pas de signification standard, elle n'est compréhensible que pour l'utilisateur humain. En plus, la sémantique associée aux documents XML n'est pas accessible à la machine, n'ayant de sens que pour l'homme.

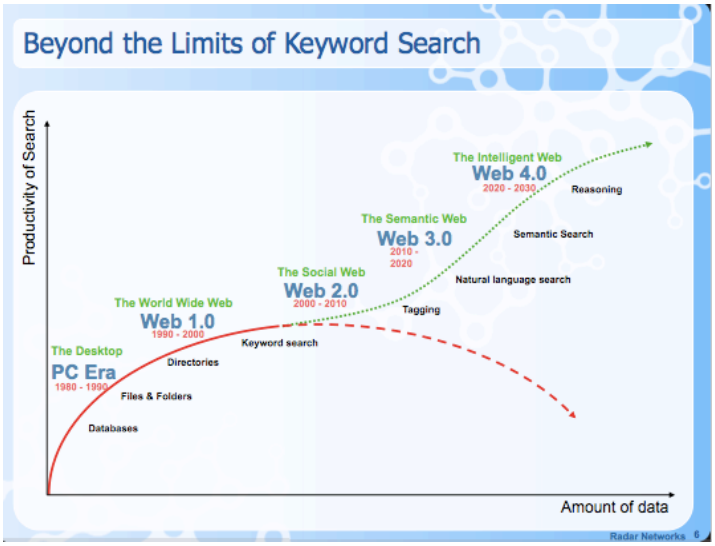


Figure 2-22 Limites de la recherche par mot-clé

L'abstraction à ce niveau est liée à la structure représentant les données, la nécessité de sémantique passe par l'utilisation de la structure pour en spécifier le sens, donc par une séparation explicite de la structure de représentation de son sens. C'est la philosophie actuelle du web.

Le web actuel a le défaut d'être conçu par des humains et pour être facilement accessible et compréhensible par des humains. Dans un contexte où la masse de données n'est pas facilement contrôlable par des utilisateurs humains, le web est de plus en plus utilisé par des machines telles que les moteurs de recherche, les robots. Une des questions majeures est

de localiser la bonne information, l'extraire et la mettre en relation avec d'autres informations localisées abordant ou intéressant le domaine d'application.

L'incompréhensibilité est liée à l'absence de sémantique clairement définie et formalisée pour être accessible. Pour cela, les langages RDF [Lassila et Swick, 1999] et RDF-Schema ont été définis pour la description sémantique des métadonnées, servant à fournir l'interopérabilité au niveau sémantique entre les applications. Des défauts de RDF et RDF-Schema ont fait sentir la nécessité d'ajout d'une couche ontologique. Le langage OWL[20] étendant le langage RDF-Schema a été introduit avec une sémantique formelle et un support efficace de raisonnement. Ces caractéristiques inclues l'approche globale de la première version du web sémantique.

Dans wikipédia[21], le web sémantique est défini comme désignant un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par le W3C[22].

Cette définition est assez simple pour une notion en réalité très complexe. Elle donne cependant les grandes lignes de l'approche web sémantique qui constitue une extension du web actuel visant à offrir une structuration des connaissances sur les sources disponibles. Cela permet à des non-humains d'utiliser et de combiner les données grâce à leur description sémantique pour répondre à des questions qui normalement nécessitent une compréhension de leur sens.

Cette notion est utilisée depuis 1994 par Tim Berners Lee, l'inventeur du web. De cette idée de base où l'on cherchait simplement à formaliser et structurer la connaissance décrivant les données afin de permettre à des machines de traiter l'information grâce à des informations dites sémantiques, un long chemin a été parcouru. Cette idée poursuit son évolution transformant progressivement le web en une base de données géante basée sur des standards prédéfinis.

La description des connaissances est structurée sous forme de métadonnées dont la description est standardisée, permettant ainsi d'utiliser les mêmes termes ou des termes dont les équivalences et relations sémantiques sont connues pour structurer l'information. Une interconnexion des différentes données devient alors disponible. La première version du web ne permettait qu'une mise en relation de l'information par le niveau des pages web avec l'url.

[20] http://w3.org/TR/owl-features/
[21] http://fr.wikipedia.org
[22] http://www.w3.org/

Dans le niveau d'abstraction du web sémantique, la mise en relation ne se limite pas à la relation hypertexte définissant un web de relations entre des pages, mais elle relie des données structurées dont les connaissances qui les décrivent sont mises en relation par une architecture normalisant les représentations. Cette standardisation est basée sur une structure homogénéisant ou structurant les métadonnées par l'intermédiaire des ontologies ainsi que les technologies permettant leur représentation, interrogation et diffusion.

Il existe actuellement plusieurs initiatives implantant cette philosophie actuelle avec des projets comme « web of data », « linked data »[23], « Giant Global Graph »[24] ou le « web 3.0 »[25].

2.3.4.2.2 Critiques du web sémantique

Il existe plusieurs critiques adressées au web sémantique. Sa lourdeur souvent associée à son système de normalisation et de standardisation est largement soulignée. Dans la perspective de la description des connaissances, il convient par exemple de s'accorder sur l'ensemble des termes à utiliser pour décrire les connaissances ainsi que les relations existantes entre ces termes. Ce consensus est bien difficile à mettre en œuvre dans l'échelle du web.

Rappelons que la problématique de base est la compréhension des données qui passe par celle de leur description, notamment au préalable celui du langage naturel qui est le plus utilisé dans les documents web. Des critiques du web sémantique s'appuient sur ce principe pour souligner que c'est de loin le cas. Les standards proposés comme RDF et OWL sont des langages pas très accessibles, ils sont plutôt orientés vers une manipulation par des machines. L'utilisation de ces langages permet de mesurer la difficulté de représenter un texte simple via RDF ou OWL. La question qui demeure donc est de savoir comment passer de la représentation textuelle à celle sémantique.

Le web sémantique devra être bâti sur un existant de plusieurs milliards de pages web et une masse de données colossale, d'où la nécessité d'annoter ces informations pour les prendre en compte dans le processus de structuration et mise en relation des connaissances. Il sera fastidieux de mettre en relation, du point de vue de l'unité d'information que constitue le nœud RDF, l'ensemble de ces informations. Il est bien possible, comme certaines solutions tentent de le faire, de scanner le web pour établir des triplets mais on tombe rapidement sur le coup de la complexité et l'aspect titanesque de ce processus. Une alternative est d'amener les

[23] http://linkeddata.org
[24] http://dig.csail.mit.edu/breadcrumbs/node/215
[25] http://fr.wikipedia.org/wiki/Web_3.0

diffuseurs de données à décrire leurs données et de les déclarer. C'est le fondement des initiatives sur les systèmes de datahubs où les sites web devraient être capables de signaler par eux-mêmes leurs changements auprès des différents moteurs de recherche.

D'autres critiques s'appuient sur le théorème de complétude de Kurt Gödel affirmant qu'un système ne peut à la fois être cohérent et complet. Il démontre théoriquement que tout système logique, aussi complexe soit-il, est incomplet dans le sens que le système ne peut prouver toutes les propositions vraies qu'il contient, sans sortir de lui-même pour avoir cette preuve. D'où, l'impossibilité d'être certain que le système logique global mis en œuvre ne mène pas à des contradictions.

Tim Berners Lee, souligne aussi dans un entretien accordé à BuissnessWeek datant de 2007 en affirmant *: Je ne pense pas que c'est un très bon terme, mais nous sommes coincés avec maintenant. Le mot web sémantique est utilisé par différents groupes pour signifier des choses différentes. Mais maintenant les gens commencent à comprendre que le web sémantique est le* ***« Data Web » (ou web de données)****. Je pense que nous aurions pu l'appeler le «* ***Data Web*** *»*[26].

2.3.4.2.3 Web sémantique deuxième version

Selon Nova Spivack [Spivack, 2007], la notion de « Data Web » est celle transformant le web d'un ensemble de serveurs de fichiers distribués à un ensemble de bases de données distribuées. Cette approche prône la mise à disposition des données sur le web de manière ouverte avec le passage de liens entre documents actuels à celui de liens entre les données contenues dans les documents. Spivack l'énonce en termes de notion d'hyperdata qui fait référence aux relations entre données par allusion à l'hypertexte dans les documents. Cette mise en relation des données fait face directement à deux problématiques de taille qui reviennent et qu'il faut résoudre : celle de l'hétérogénéité structurelle des données pour leur accès, leur échange, leur combinaison, et celle de l'hétérogénéité sémantique liée à leur description.

Du point de vue structurel, l'*approche dataweb* qui constitue la base de nos travaux propose l'alternative d'une représentation des données dans un *entrepôt de documents XML* ouvert sur le web. A cet entrepôt est associé un catalogue de métadonnées constitué par une base de concepts fournissant un ensemble d'unités élémentaires décrivant les attributs observés structurés sous la forme d'unités sémantiques. Leurs relations hiérarchiques sont

[26] I don't think it's a very good name but we're stuck with it now. The word semantics is used by different groups to mean different things. But now people understand that the Semantic Web is the Data Web. I think we could have called it the Data Web.

décrites grâce à l'utilisation d'un thesaurus du domaine. Donc, l'aspect de mise à disposition des données sur le web de manière accessible est bien retrouvé. Nous avons également une description sémantique des sources de données ainsi que des données et leurs relations. L'approche offre aussi un module de mise en ligne des données sur le web.

Nous pouvons ainsi avoir, dans un contexte de données environnementales, un système d'intégration permettant de mettre de manière ouverte des données avec une description sémantique sur le web. Du point de vue global, nous pouvons avoir un ensemble de relations trouvées et ajoutées par un expert du domaine sur le vocabulaire décrivant les données. Cette démarche s'inscrit donc bien dans le cadre de la démarche « data web » pour des données de même domaine décrite par un thesaurus préalablement existant.

2.4 Conclusion

Nous avons présenté dans ce chapitre la problématique de l'hétérogénéité des données et des approches d'intégration. Elle se décline en trois types : syntaxique, structurel et sémantique.

Pour résoudre la problématique de l'intégration structurelle des données, deux approches sont utilisées. La première consiste à utiliser un système d'adaptateur permettant de passer entre les formats de représentation des données. La seconde approche, que nous avons adoptée, consiste à homogénéiser structurellement les données, les transformant ainsi selon un format unique de représentation. Cela résout en partie le problème de la dimension organisationnelle (structurelle) de l'hétérogénéité.

A l'image de la problématique d'hétérogénéité relative au langage, il convient de restreindre le vocabulaire utilisé pour décrire sémantiquement les objets, à travers une grammaire avec une abstraction des règles hiérarchiques. Les ontologies comme conceptualisation d'un domaine conviennent bien à cette problématique.

Nous avons aussi présenté dans ce chapitre les approches d'intégration de données ainsi que le modèle formel des systèmes d'intégration. Deux approches sont utilisées globalement, celle dite entrepôt et celle dite médiateur. Pour la mise en relation du niveau des données et celui de leur description, toutes ces deux approches sont basées sur la définition de schémas. Nous avons abordé les différentes techniques permettant d'intégrer les différents schémas des sources et leur interrogation, c'est-à-dire comment répondre efficacement aux requêtes posées sur le schéma global. Des projets d'intégration les appliquant ont été présentés, parmi lesquels le système utilisé dans Xyleme ou dans [Lima et al., 2003] semblable à l'*approche dataweb*, avec la migration de toutes les données vers un format

unique en XML. Nous avons aussi présenté l'approche et le modèle de *dataweb* avec une analogie faite sur la démarche et les nécessités ouvertes par les défis et les perspectives du web sémantique.

Cette approche proposée dans [Lo, 2002] constitue la base de nos travaux. Elle propose une méthodologie qui permet, d'une part, de résoudre la problématique de l'hétérogénéité structurelle par la construction d'entrepôts de documents XML (ou *dataweb*) pour chaque *partenaire* et, d'autre part, l'association d'une couche sémantique avec une base de concepts. Cette démarche de sémantisation des données est corollaire à celle de l'utilisation et la conception des ontologies dans l'état de l'art que nous allons aborder dans le prochain chapitre.

Chapitre 3

Utilisation et conception des ontologies pour l'intégration de données

Sommaire

3.1 Introduction 68

3.2 Ontologie 68

3.2.1 Définitions 68

3.2.2 Composants d'une ontologie 70

3.3 Modèles formels d'ontologie 71

3.3.1 Structure d'ontologies à base lexicale 72

3.3.2 Modèle d'ontologie SOWA 74

3.4 Langages de représentation et d'interrogation 74

3.4.1 Extensible Markup Language 75

3.4.2 Ressource Description Framework/RDF-Schema 76

3.4.2.2 Niveau de définition des types de base RDF 79

3.4.2.3 Niveau de définition des types complexes 79

3.4.2.4 Niveau de définition des schémas 80

3.4.3 Ontology Web Language 82

3.4.4 Simple Protocol And RDF Query Language (SPARQL) 87

3.5 Extraction automatique d'ontologie à partir de données 88

3.5.1 Extraction d'ontologie par apprentissage automatique 89

3.5.2 Extraction d'ontologie à partir de sources XML 89

3.6 Approches d'intégration de données basées sur les ontologies 92

3.6.1 Approche mono-ontologie 93

3.6.2 L'approche multi-ontologie 94

3.6.3 L'approche hybride 95

3.7 Conclusion 96

3.1 Introduction

Le mot « Ontologie » se compose des racines grecques ontos (ce qui existe, l'existant) et logos (le discours, l'étude). En philosophie, on peut voir l'Ontologie comme une branche fondamentale de la Métaphysique qui s'intéresse à la qualité d'être, à la notion d'existence et aux catégories fondamentales de l'existant (Wikipédia). C'est une partie de la philosophie qui a pour objet l'étude des propriétés les plus générales de l'être, telles que l'existence, la possibilité, la durée, le devenir (Académie). C'est aussi l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe. L'Ontologie a donc des implications directes sur notre conception de la réalité (Wikipédia).

Dans ce chapitre, sont abordées les méthodologies de structuration et de formalisation des ontologies. Bien souvent, l'expert du domaine et l'informaticien chargé de la construction d'une ontologie ont leur propre approche de modélisation et de structuration avant de procéder à l'implémentation. Cela s'expliquerait probablement par la difficulté procéder à une conceptualisation d'un domaine. C'est ce qui nous amène à la question essentielle : peut-on disposer d'un modèle qui soit commun à un domaine? Cela ne semble pas être aussi évident que la modélisation des logiciels commune.

3.2 Ontologie

Les ontologies occupent une place très importante dans le contexte de notre travail. Il est nécessaire de préciser ce qu'elles sont et, en quoi elles peuvent participer à l'intégration de données.

3.2.1 Définitions

L'ontologie est un concept hérité de la philosophie, ce qui pourrait expliquer la difficulté et la nature complexe de sa définition. En informatique, la définition la plus usitée est sans doute celle de Th. Grüber qui définit une ontologie comme étant « la spécification d'une conceptualisation » [Grüber, 1993].

La conceptualisation est relative à l'idée que l'on se fait de quelque chose et la spécification est la description formelle de cette idée ou connaissance que nous en avons. Elle ne peut donc qu'être relative, c'est-à-dire ne refléter que le consensus ou le point de vue d'un ensemble restreint d'individus. Donc il ne s'agit pas d'une description formelle reconnue de tout le monde.

Cette définition renvoie à quelque chose d'essentiel qui est la nature partagée ou consensuelle des idées que l'on se fait des choses. Elle est par conséquent assez relative suivant le domaine d'application et peut difficilement refléter un consensus universel. Cette définition de Grüber semble ne pas préciser quelque chose d'essentielle qui est que cette conceptualisation repose sur un consensus partagé.

Borst [Borst 1997] propose d'ailleurs une version modifiée de cette définition où l'ontologie est définie comme «une spécification explicite et formelle d'une conceptualisation partagée ». Elle est pour un ensemble d'individus la spécification formelle et explicite d'un consensus partagé sur les connaissances. C'est ce qui montre son importance dans le domaine de l'intégration des connaissances.

Une ontologie permet pour un domaine d'application cible de se mettre d'accord sur l'ensemble des objets reconnus comme existants ainsi que leur manière d'être et leurs relations, standardisant ainsi les références sur ces objets ainsi que le descriptif qui leur est associé. Cette conceptualisation dans le cas d'une multiplicité d'objets partageant des traits communs identifiables permet aussi de désigner une représentation plus abstraite servant à résumer les traits communs. Cette abstraction est réalisée par l'intermédiaire d'un concept. La conceptualisation permet et facilite la communication de connaissances. D'une manière générale, Ushold et Gruninger [Uschold et Gruninger, 1996] divisent l'espace d'utilisation des ontologies en trois parties :

1. la communication entre personnes ayant des points de vue et des besoins différents ;
2. l'interopérabilité entre utilisateurs qui ont besoin de s'échanger des données et qui emploient des outils et vocabulaires différents ;
3. l'ingénierie des systèmes, où la capacité des ontologies à faire partager et réutiliser des connaissances.

L'utilité des ontologies dans le domaine de l'intégration de données se situe dans ces trois catégories identifiées par Ushold et Gruninger. Les ontologies peuvent être utilisées pour établir les liens sémantiques entre des éléments dans des sources différentes (OBSERVER [Mena et al, 2000]). Elles peuvent aussi servir de modèle d'interrogation lorsqu'elles sont utilisées pour spécifier le schéma global (SIMS [Arens et al, 1997], PICSEL [Rousset et Reynaud, 2003]). Cependant la méthode de conception des ontologies et la manière dont elles sont utilisées peuvent être différentes.

A l'image des thésaurus, la représentation de connaissances sous forme de consensus se fera via une restriction sur le vocabulaire globalement utilisé, et par conséquent sur un vocabulaire contrôlé et reconnu de tous. Leurs termes ou concepts permettent explicitement

d'identifier les objets existants. Ce consensus devra aussi refléter les connaissances sur la hiérarchie et les relations connues et convenues comme existantes entre ces objets.

Les ontologies sont censées donner une référence commune, pour un domaine donné à un ensemble d'individus. Elle sert à standardiser la nomination des objets ainsi que leur manière d'être, exprimant en partie leur sens.

Nous identifions deux types de composants pour une ontologie que sont les concepts et les relations.

3.2.2 Composants d'une ontologie

Les composants d'une ontologie sont l'ensemble des éléments servant à véhiculer les connaissances exprimant la conceptualisation partagée du domaine. De manière similaire à la taxonomie définie dans [Psyché et al., 2003], les connaissances traduites par une ontologie sont à véhiculer à l'aide des éléments suivants :

1. les concepts appelés aussi des classes constituent les nœuds ou éléments de base de la structure taxinomique de l'ontologie. Le concept constitue une abstraction de caractères communs à plusieurs individus. Selon Wikipédia, *« Dans le langage courant, un concept est défini comme une idée ou représentation de l'esprit qui abrège et résume une multiplicité d'objets empiriques ou mentaux par abstraction et généralisation de traits communs identifiables par les sens »*. Il peut donc abstraire le descriptif d'une tâche donnée, d'un processus, ou d'une idée résumant les traits communs d'individus comme c'est souvent le cas des concepts définis dans notre contexte environnemental.
2. les relations de diverses natures qui définissent et qualifient du point de vue sémantique les interactions existantes entre les concepts. Ces relations peuvent être de nature hiérarchique avec par exemple la subsomption ou de nature sémantique entre les concepts;
3. les fonctions qui définissent des types de relations particulières entre les concepts.
4. les axiomes qui sont utilisés pour inclure dans la structure ontologique des propositions toujours vraies. Leur inclusion dans une ontologie permet de définir la signification de composants d'ontologie, des contraintes sur les valeurs des attributs ou la vérification de l'exactitude d'informations indiquées dans l'ontologie ou de moyen de base pour l'inférence. Lorsque la structure de l'ontologie contient un ensemble symbolisant les axiomes, l'ontologie est qualifiée de lourde et dans le cas contraire d'ontologie légère.

3.3 Modèles formels d'ontologie

D'après le Petit Robert, la modélisation est « la mise en équation d'un phénomène complexe permettant d'en prévoir l'évolution ». En disposant donc d'une représentation d'un système réel, elle permet, par des hypothèses, de donner les états résultants dans un futur donné, souvent par l'utilisation de la simulation. En ingénierie des connaissances, la modélisation s'est positionnée dans cette perspective, mais en s'intéressant plus à la structuration des connaissances. Dans le domaine, les seules structures proposées sont celles dérivées des travaux de [Maedche et. al, 2002]. Différentes variantes sont proposées comme celle de [Hernandez et. al, 2006], mais pour la grande majorité des travaux, les concepteurs passent de la spécification à l'implémentation de l'ontologie en utilisant une des méta-ontologies (langages).

Les méthodes de gestion de projet sont applicables à l'ingénierie des ontologies, mais la part de formalisation indépendamment de l'implémentation n'est pas rigoureusement prise en compte. Cette confusion entre le modèle et son implémentation est assez répandue dans le domaine [Uschold et Gruninger, 1996]. Les ontologies peuvent être modélisées avec des langages de différents degrés de formalisation :

- hautement informelle : en langue naturelle,
- semi-informel : langue naturelle restreinte, contrôlée, structurée,
- semi-formel : dans un langage artificiel définit formellement,
- rigoureusement formel : dans un langage avec une sémantique formelle, des théorèmes, et des preuves pour vérifier les propriétés telles que la validité et la complétude.

Ce constat est également fait dans [Simperl et Tempich, 2006] où les auteurs affirment que : «*La majorité des interviewés n'ont pas perçu une différence claire entre la conceptualisation et les étapes nécessaires à l'implémentation. Après une description et une classification simple des résultats attendus, l'équipe d'ingénierie implante l'ontologie à l'aide d'un éditeur d'ontologie* »[27].

La question ici est surtout de savoir, si dans le processus de construction des ontologies, le modèle doit être établi en faisant abstraction de l'implémentation en reflétant uniquement au niveau structurel la conceptualisation attendue du domaine de cette dernière.

[27] The majority of the interviewees did not perceive a clear cut between the conceptualization and the implementation steps as necessary. After a lightweight description and classification of the expected outcomes the engineering team implemented the ontology with the help of a common ontology editor. This task was primarily performed by domain experts, who did not report any particular difficulties in getting familiar with or utilizing simple ontology editors

La difficulté est de pouvoir fournir une conceptualisation pouvant capturer toute la connaissance d'un domaine. Cette conceptualisation est souvent qualifiée de partielle car, en l'état de l'art, il est illusoire de croire pouvoir capturer dans un formalisme toute la complexité d'un domaine, selon F. Gandon[28].

3.3.1 Structure d'ontologies à base lexicale

Une structure d'ontologie déduite de celle proposée dans [Maedche et. al, 2002] est le quadruplet suivant :

$$\mathcal{S} := \{C;, \mathcal{R}, \mathcal{H}^C, \mathcal{A}^0\} \text{ où :}$$

1. C est l'ensemble des concepts ;
2. $\mathcal{R} \subseteq C \times C$ est l'ensemble des relations. $r = (c_1, c_2) \in$ R s'écrit aussi $r(c_1)=c_2$;
3. $\mathcal{H}^C$ hiérarchie (taxonomie) de concepts : $\mathcal{H}C \subseteq C \times C$, $\mathcal{H}^C(C_1, C_2)$ signifie que C_1 est un sous-concept de C_2 (relation orientée) ;
4. $\mathcal{A}^0$: ensemble d'axiomes, exprimés dans un langage logique adapté (logique de description, logique du 1er ordre).

Pour faire face au niveau lexical des ontologies, un lexique est introduit. Pour une structure d'ontologie $\mathcal{S} := \{C, \mathcal{R}, \mathcal{H}^C, \mathcal{A}^0\}$, un lexique L est défini comme le quadruplet

$$\mathcal{L} := \{\mathcal{L}^C, \mathcal{L}^{\mathcal{R}}, \mathcal{F}, \mathcal{G}\} \text{ où :}$$

1. $\mathcal{L}^C$ est l'ensemble dont les éléments constituent l'entrée lexicale des concepts ;
2. $\mathcal{L}^{\mathcal{R}}$ est l'ensemble dont les éléments constituent l'entrée lexicale des relations ;
3. $\mathcal{F} \subseteq \mathcal{L}^C \times C$ la relation de référence pour les concepts telle que :
   - $\forall\ l_c \in \mathcal{L}_C : \mathcal{F}(l_c) = \{c \in C / (l_c, C) \in \mathcal{F}\}$;
   - $\forall\ c \in C\ \mathcal{F}^{-1}(c) = \{l_c \in \mathcal{L}^C / (l_c, c) \in \mathcal{F}\}$;
4. $\mathcal{G} \subseteq \mathcal{L}^{\mathcal{R}} \times \mathcal{R}$, la relation de référence pour les relations telle que :
   - $\forall\ lr \in \mathcal{L}C : \mathcal{G}(l_r) = \{r \in \mathcal{R} / (l_r, \mathcal{R}) \in \mathcal{G}\}$;
   - $\forall\ r \in \mathcal{R}: \mathcal{G}^{-1}(r) = \{l_r \in \mathcal{L}^{\mathcal{R}} / (l_r, r) \in \mathcal{G}\}$;

Une ontologie est formalisée comme un couplet $(\mathcal{L}, \mathcal{S})$ où la composante structurelle $\mathcal{S}$ permet une structuration et conceptualisation rigoureuse du domaine et la composante lexicale $\mathcal{L}$, le vocabulaire contrôlé sur cette conceptualisation. Il existe deux types d'ontologies, selon la composante de la structure $\mathcal{S}$:

[28] http://interstices.info/display.jsp'id=c_17672

— Si la composante $\mathcal{A}^0$ est absente dans la structure $\mathcal{S}$, on dit que l'on a une ontologie légère. Une ontologie légère ne se compose que des concepts, des types atomiques et les relations hiérarchiques « *is-a* » entre les concepts. Il est plus facile de trouver un consensus lors de leur spécification [Kiryakov et al., 2003]. Ce type d'ontologie peut être modélisé en utilisant les méthodologies issues du génie logiciel comme UML (Unified Modeling Language) et la modélisation par diagramme Entité/Relation.

— Si la composante $\mathcal{A}^0$ est présente dans la structure, on dit que l'on a une ontologie lourde ou riche. Une ontologie riche est une ontologie légère avec les contraintes de cardinalité, une taxonomie de relations, des axiomes/héritages sémantiques comme les logiques de description [Ducourneau et. al, 1998], [Kayser 1997], les logiques de proposition, etc. Parmi les méthodologies de modélisation utilisées pour les ontologies lourdes on distingue: celles issues de l'intelligence artificielle basées sur les frames et sur la logique de premier ordre ou celle de la modélisation basée sur la logique de description.

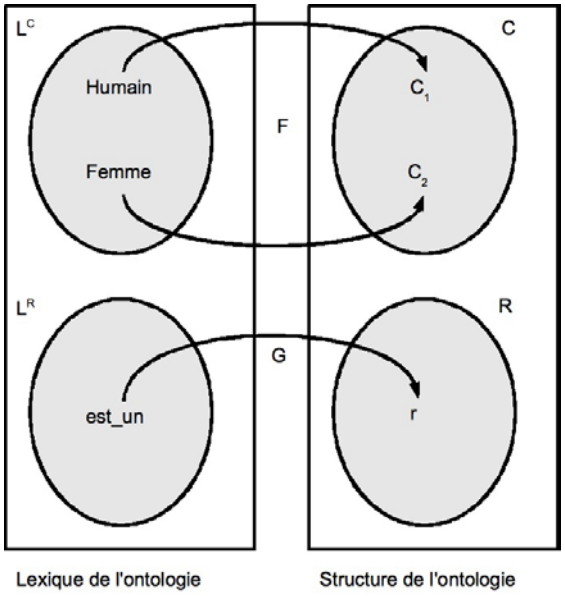


Figure 3-1 Illustration graphique des composantes d'une ontologie

Illustrons cette formalisation par l'exemple suivant. Soit une $\mathcal{S} := \{\mathcal{C}, \mathcal{R}, \mathcal{H}^C, \mathcal{A}^0\}$ une structure d'ontologie légère tel que : $\mathcal{C}=\{c_1, c_2\}$ et $\mathcal{R}=\{r\}$ où $r(c_1)=c_2$ et $\mathcal{A}^0=\emptyset$; l'ontologie étant légère. *Soit* $\mathcal{L} := \{\mathcal{L}^C, \mathcal{L}^R, \mathcal{F}, \mathcal{G}\}$ le lexique associé tel que $\mathcal{L}^C = \{\text{'Femme'}, \text{'Humain'}\}$ et $\mathcal{L}^R = \{\text{'est\_un'}\}$, $\mathcal{F}(\text{'Femme'})=c_1$, $\mathcal{F}(\text{'Humain'})=c_2$.

Une structure dérivée proposée dans [Hernandez et. al, 2006] consiste à rajouter dans la structure de l'ontologie trois composantes supplémentaires représentant, la signature des relations associatives, la signature des relations d'attribut, l'ensemble des relations d'attribut

et l'ensemble des types de données. Ainsi l'ontologie sera modélisée selon la structure suivante :

$$S := \{C , \mathcal{R} , \mathcal{A} , \mathcal{T}, \mathcal{H}^C , \sigma_{\mathcal{R}} , \sigma_{\mathcal{A}} , \mathcal{A}^0\}$$

— $\sigma_{\mathcal{R}} : \mathcal{R} \rightarrow C \times C$ est la signature d'une relation associative ;

— $\sigma_{\mathcal{A}} : \mathcal{A} \rightarrow C \times \mathcal{T}$ est la signature d'une relation d'attribut.

Nous allons reprendre le même modèle qui convient bien aux données du *SIC*, ainsi que l'idée que nous avons introduit des ontologies. Il permet par sa composante lexicale de capturer le vocabulaire contrôlé du partenaire et de structurer le domaine de chaque partenaire grâces aux mécanismes proposées dans sa composante structurale.

Un modèle formel d'ontologie PLIB (Parts LIBrary : norme ISO 13584) a été également proposé dans [Pierra et al., 2005].

3.3.2 Modèle d'ontologie SOWA

Bien que rangée dans les langages des ontologies, la structure d'un modèle de graphes conceptuels [Sowa, 1999] est intéressante dans ce cadre. Elle permet de modéliser une ontologie. Cette approche est d'autant plus intéressante qu'étant l'une des rares à intégrer le niveau des instances dans la structure de l'ontologie. C'est une problématique à revoir par rapport à la frontière avec les bases de connaissances. La structure est le quintuplet :

$$S := \{\mathcal{TC}, \mathcal{TR} , \sigma , I, \tau \}$$ où :

1. $\mathcal{TC}$ et $\mathcal{TR}$ sont respectivement, les ensembles hiérarchiquement structurés des types de concepts et des relations ;
2. σ est l'application qui permet d'ordonner les relations en prenant en compte leurs arguments ;
3. I est un ensemble de marqueurs individuels (instances de concept) ;
4. $\tau : I \rightarrow \mathcal{TC}$ est une application définie tel que $\{ \forall\ i \in I, \tau\ (i) = t \in \mathcal{TC}\}$

3.4 Langages de représentation et d'interrogation

De nombreux langages ont été utilisés pour la spécification d'ontologies, la plupart étant basés sur XML et les logiques de description. Dans cette section, nous présenterons brièvement RDF/RDFS et OWL qui sont les plus utilisés dans ce contexte.

RDFS et OWL sont aussi des méta-ontologies, une méta-ontologie étant une ontologie de l'ontologie définissant donc un ensemble de primitives de représentation d'une ontologie. XML est le support sur lequel s'appuient RDF et OWL pour définir des structures de données

et les relations qui les lient. C'est un des langages permettant de structurer les connaissances. Cependant, il fournit seulement une structure syntaxique pour des documents et ne donc permet pas une interprétation sémantique des données.

3.4.1 Extensible Markup Language

XML est un méta-langage proposé par le W3C en 1988 avec comme principale objectif la séparation de la structure logique des documents de leur structure physique (ou de présentation). Il a été proposé comme une extension du langage HTML. C'est un langage conçu pour la description des données. Les balises XML n'étant pas prédéfinies, l'utilisateur peut écrire ses propres balises pour structurer ses données en faisant abstraction de la présentation des données. Un ensemble de mécanismes a été développé pour spécifier le chemin d'accès et d'identification des données avec Xpath[29], un outil pour l'accès aux données comme XQuery[30].

Pour tenir compte de l'aspect présentation des données, le standard XSL (eXtensible StyleSheet Language) a été proposé ; il permet de transformer un document XML en un autre format de représentation.

Un document XML peut avoir trois parties : un prologue, un certains nombre d'éléments et un épilogue.

Le prologue comporte une déclaration XML définissant la version et format d'encodage des données et une référence en option à des documents de structuration externe spécifiant la grammaire servant de modèle à la structuration du document.

```
<?xml version="1.0" encoding="UTF-16" ?>
<!DOCTYPE book SYSTEM "monDTD.dtd">
```

Les éléments XML permettent de définir ce dont parle le document donc les données. Un élément comporte une balise d'ouverture, le contenu ou les données que l'on désire représenter et une balise de fermeture. L'utilisateur peut choisir le nom des balises.

Le contenu d'un élément peut être du texte, d'autres éléments ou encore rien du tout. L'élément peut aussi avoir un ou plusieurs attributs. L'attribut permet de définir une propriété de l'élément concerné. L'attribut en XML a un nom et une valeur.

[29] http://www.w3.org/TR/xpath
[30] http://www.w3.org/TR/xquery/

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<CATALOGUE>
        <ALBUM CD>
                <TITRE>Xalle</TITRE>
                <ARTISTE>Titi</ARTISTE>
                <PAYS>SENEGAL</PAYS>
                <PRIX>2500</PRIX>
                <ANNEE>2002</ANNEE>
        </ALBUM CD>
        <ALBUM CD>
                <TITRE>Alsamaday</TITRE>
                <ARTISTE>Youssou Ndour</ARTISTE>
                <PAYS>SENEGAL</PAYS>
                <PRIX>2000</PRIX>
                <ANNEE>1988</ANNEE>
        </ALBUM CD>
</CATALOGUE>
```

3.4.2 Ressource Description Framework/RDF-Schema

RDF (Ressource Description Framework) [Lassila et Swick, 1999] est un langage, recommandé par le W3C et fondé sur les notions de ressources et de relations entre ressources. RDF est aussi un langage de formalisation des relations entre les différents termes d'un vocabulaire ou ontologie. En RDF, les concepts et les objets sont identifiés par des URIs (Uniform Resource Identifiers), et les ressources décrites en termes de propriétés et de valeurs de propriétés, permettant ainsi de représenter un document comme un graphe orienté étiqueté. Il introduit un document structuré en RDF avec une syntaxe composée des triplets RDF constitués par l'association (Sujet, Predicat, Objet). RDF fournit un langage pour exprimer des déclarations à propos de ressources, en utilisant des propriétés et des valeurs.

Cependant, il nous faut aussi définir le vocabulaire (l'ensemble des termes) que nous souhaitons utiliser dans ces déclarations, le structurer en une hiérarchie de classes et spécifier quelles propriétés s'y appliquent. Par exemple, pour décrire l'activité de conduite on pourrait définir des termes tels que virage à droite, arrêt, accélérer, décélérer, ou encore angle du volant, etc.... Afin de disposer d'un vocabulaire propre aux utilisateurs du domaine et donner du sens aux informations stockées sous forme de triplets, il a été alors proposé l'ajout des RDFS, RDF-Schéma pour définir les métadonnées, les ontologies.

RDFS (RDF Vocabulary Description Language : RDFSchema) est un langage permettant de décrire de telles classes et propriétés, et de spécifier quelles classes et propriétés doivent être utilisées ensemble.

Figure 3-2 Exemple de schéma d'un triplet RDF

Cependant, le couplet RDF/RDFS peut, dans certains cas, se révéler comme une approche insuffisante. Cette dernière, en plus du manque de mécanisme d'inférence, compte un certain nombre de limites comme :

— rdfs:range définit le domaine de valeurs d'une propriété quelle que soit la classe concernée. Par exemple, il ne permet pas d'exprimer que les vaches ne mangent que de l'herbe alors que d'autres animaux mangent également de la viande ;

— RDFS ne permet pas d'exprimer que deux classes sont disjointes. Par exemple, les classes des hommes et des femmes sont disjointes ;

— RDFS ne permet pas de créer des classes par combinaison ensembliste d'autres classes (intersection, union, complément). Par exemple, on veut construire la classe Personne comme l'union disjointe des classes des hommes et des femmes ;

— RDFS ne permet pas de définir de restriction sur le nombre d'occurrences de valeurs que peut prendre une propriété. Par exemple, on ne peut pas dire qu'une personne a exactement deux parents ;

— RDFS ne permet pas de définir certaines caractéristiques des propriétés : transitivité (par exemple : estPlusGrand-Que), unicité (par exemple : estLePèreDe), propriété inverse (par exemple : mange est la propriété inverse de estMangéPar).

Le besoin de disposer d'un vocabulaire s'est avérée indispensable pour décrire ces métadonnées ; il a été alors proposé l'ajout des RDFS, RDF-Schéma pour définir les métadonnées, les ontologies.

Il y a quatre niveaux de modélisation dans RDF: un niveau physique, un niveau de définition des types de base avec trois types d'objets, un niveau de définition des types complexes et au niveau de définition des schémas.

Dans ce niveau une description de métadonnées RDF est un triplet $(s; p; o)$ signifie *que « le sujet s a comme valeur pour la propriété p l'objet o »*, s et p sont des URLs et o est une URL ou une valeur. Formellement, nous avons :

— un ensemble URLs : $\mathcal{U}$;

— un ensemble de littéraux (chaînes de caractères) : $\mathcal{V}$;

— un ensemble de déclarations, avec les opérateurs mathématiques sur les ensembles.

Un exemple de déclaration RDF est le triplet suivant *: (Http ://lil.univ-littoral.fr/ sicweb, dc :author, #Moussa).*

Ici *dc* est un préfixe qui désigne *«dublin code»* pour décrire des ressources. Cette ligne permet de spécifier que l'auteur est Moussa qui est ici une ressource interne à la base de déclaration, raison pour laquelle nous avons un« # » devant, qui est un identificateur d'objet local. Ensuite nous pouvons spécifier la page du *SIC* par exemple *«Http://lil.univ-littoral.fr/sicweb »* par le triplet suivant :

(#sicweb, #homepage, Http ://lil.univ-littoral.fr/sicweb)

Il est aussi possible de signifier que la ressource Moussa est le directeur du projet *sic*web par un triplet RDF *(#sicweb, #directeur, #Moussa)* et de spécifier qu'il possède un nom donné *(#Moussa, #nom, "Moussa Lô").* C'est ce qui nous donne le graphe RDF est représenté par la figure 3.3.

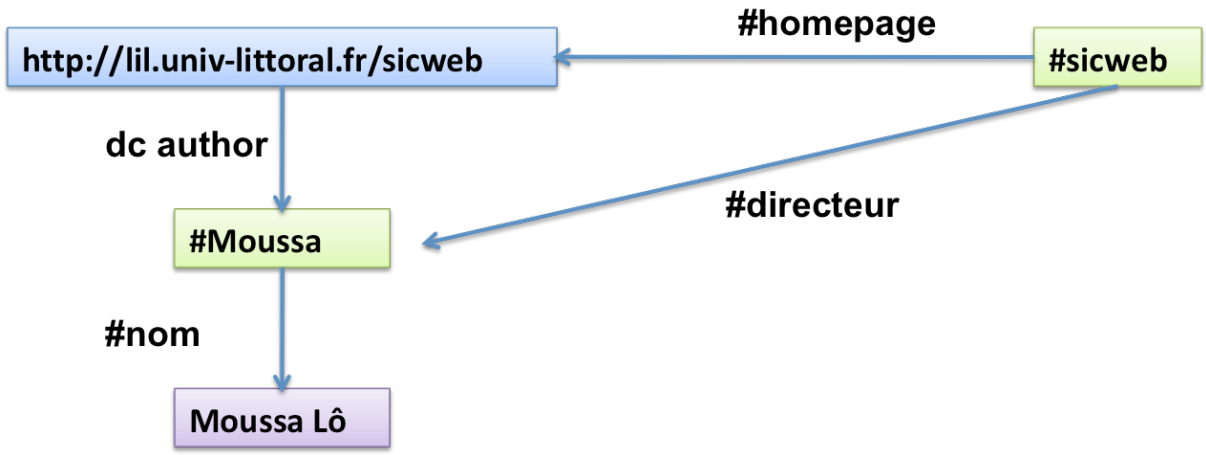


Figure 3-3 Exemple RDF

Une des propriétés intéressante au niveau RDF est la réification : elle permet de définir les triplets RDF comme des objets. Si on affirme que *« Moussa »* est l'auteur de la page web *« Http://lil.univlittoral.fr/sicweb»*, comment écrire ou introduire la déclaration consistant à se prononcer sur le triplet. Par exemple pour dire que *« Sall croit que « Moussa » est l'auteur de la page web « http ://lil.univ-littoral.fr/ sicweb ».*

Pour cela RDF permet de donner un nom au triplet et ensuite utiliser cet objet en passant par son nom en utilisant des propriétés prédéfinies dans la recommandation RDF du W3C :

(#statement1, RDF:subject, #Moussa)

(#statement1, RDF:predicate, dc :author)

(#statement1, RDF:object, Http ://lil.univ-littoral.fr/ sicweb)

(#Sall, #croit, #statement1)

3.4.2.2 Niveau de définition des types de base RDF

Dans ce niveau, trois types d'objets sont définis. Ils permettent de distinguer les URLs qui sont des ressources et/ou des propriétés et/ou des déclarations (statement) :

— un ensemble d'URLs : U;

— Ressources (RDF: Ressource) : Tout objet décrit par des expressions RDF est appelé ressource. Une ressource peut être une page Web entière, une partie d'une page web, une collection complète de pages, un objet qui n'est pas directement accessible par le web. Les ressources sont toujours nommées par des URIs avec des ancres « ids » optionnelles. Tout objet peut avoir une URI : L'extensibilité des URIs permet l'introduction d'identificateurs pour toute entité imaginable ;

— Propriétés (RDF:Property) : Une propriété est un aspect, une caractéristique, un attribut, ou une relation spécifique utilisée pour décrire une ressource. Chaque propriété possède une signification spécifique, définit ses valeurs permises, les types de ressources qu'elle peut décrire, et les relations qu'elle entretient avec les autres propriétés ;

— Déclarations (RDF:Statement) : Une ressource spécifique associée à une propriété définie ainsi que la valeur de cette propriété pour cette ressource est une déclaration RDF. Ces trois parties individuelles d'une déclaration sont appelées, respectivement, le sujet, le prédicat, et l'objet. L'objet d'une déclaration (c'est-à-dire, la valeur de la propriété) peut être une autre ressource ou un littéral ; c'est-à-dire, une ressource (spécifiée par une URI) ou une simple chaîne ou autre type de données primitif défini par XML.

3.4.2.3 Niveau de définition des types complexes

RDF permet dans ce niveau, de définir les types complexes avec les listes, séquences, énumérations,…. Un container est une ressource de type RDFs: containers (le «s»' de RDFs désigne RDF Schéma). La classe containers a trois sous-classes :

— RDF:Bag : multi-ensemble de ressources ;

— RDF:Sequence : séquence de ressources ;

— RDF:Alt : énumération de ressources.

L'appartenance à une collection est encodée par des propriétés de type *« RDF:_n »* où n est un entier.

Par exemple, nous pouvons définir les membres du LIL :

(#students, RDF:type, RDF:Bag)

(#students, RDF:_1, #Antoine)

(#students, RDF:_2, #Sall)

Ces deux lignes permettent de déclarer une collection de type *« RDF:Bag »* dans laquelle on met deux instances *#antoine et #sall.*

Nous pouvons refaire la même chose pour une collection sur les membres du LIL :

(#lilmembers, RDF:type, RDF:Bag)

(#lilmembers, RDF:_1, #Basson)

(#lilmembers, RDF:_2, #Boubeffa)

(#lilmembers, RDF:_3, #Bourgin)

Dans RDF, une collection est une ressource, nous avons la possibilité d'avoir des collections de collections : *(#lilmembers, RDF:_6, #students).* Ce qui permet de spécifier que les éléments de la collection students sont aussi des membres du LIL. Tandis qu'une liste est une ressource de type *RDF:List.*

3.4.2.4 Niveau de définition des schémas

RDF a été conçu comme étant un ensemble de classes et de propriétés prédéfinies. Dans le cas de XML, ceci nous amène ainsi à mettre en place les schémas XML. . RDF permet de représenter des déclarations de propriétés sur des ressources, mais pas d'exprimer des connaissances sur les propriétés ou sur les types de ressources. Par exemple on voudrait bien avoir des réponses s aux questions : Quelles sont les propriétés autorisées sur un type de ressources ? Quelles sont les valeurs autorisées pour une propriété ? Quels sont les liens entre les types de ressources (généralisation / spécialisation) ? RDF ne permet pas d'y apporter des réponses. Par conséquent, la solution consiste à définir le vocabulaire du domaine des ressources.

L'idée a été émise de mettre en place les RDF Schéma. RDF Schéma permet d'étendre ce vocabulaire avec des classes et types de propriétés spécifiques à une application ou un domaine, permettant ainsi de décrire des classes et des propriétés. Il faut remarquer que RDF-Schéma ne fournit pas de vocabulaire mais permet de définir un vocabulaire.

RDF-Schema est un « système de typage » pour RDF, comparable à l'approche orientée objet : nous avons les notions de Classes : Une classe est un type (ou une catégorie) qui regroupe plusieurs instances (ressources) partageant des caractéristiques communes. Une classe est identifiée par une URI. Pour préciser qu'une URI est une classe, il faut écrire que cette ressource a pour *« RDF:type » « RDFs:Class».* Par convention, un nom de classe commence

toujours par une majuscule et une instance par une minuscule. Un type de propriété est une ressource de type *«RDF:Property»*. RDFS permet de restreindre le domaine et le co-domaine d'un type de propriété :

— *RDFs:subpropertyOf : sous-propriétés ;*

— *RDFs:domain : domaine d'une propriété, c'est-à-dire les types de ressources sur lesquelles peux porter une ressource ;*

— *RDFs:range : co-domaine, c'est-à-dire l'ensemble des valeurs autorisées pour la propriété.*

Dans la suite de l'exemple précédent, nous aurons la définition du schéma suivant :

— (#Group, RDF:type, RDFs:Class) pour définir la classe nommée #Group ;

— (#ResearchGroup, RDF: type, RDFs:Class) pour spécifier que #ResearchGroup est une classe, une ontologie ;

— (#ResearchGroup, RDFs:subclassOf, #Group) permet de dire que #Research-Group est une sous-classe de la classe #Group. Donc la notion de généralisation et spécialisation d'une classe est possible ;

— (#memberOf, RDF:type, RDF:Property) pour définir une propriété #memberOf ;

— (#memberOf, RDFs:domain, #Student) pour spécifier que #memberOf est une propriété de la classe #Student ;

— (#memberOf, RDFs:domain, #Researcher) pour spécifier que #memberOf est une propriété de la classe #Researcher ;

— (#memberOf, RDFs:range, #Group) pour bénéficier que la valeur de #memberOf est un co-domaine, l'ensemble des valeurs que la classe #Group peut avoir ;

(#*sic*web, RDF:type, #ResearchGroup), on utilise RDF:type pour définir que la classe #*sic*web est une instance de la classe #ResearchGroup RDFS permet donc la représentation d'ontologies simples (sans capacité de raisonnement intégrée).

RDFS permet de définir une relation hiérarchique comme l'héritage, les propriétés des attributs. Cependant, il ne permet pas de définir des méthodes de classe, mais uniquement de les décrire. La structure n'est pas non plus fixée : d'autres propriétés peuvent être ajoutées.

Sous forme de schéma, nous aurons une partie méta-schéma, la partie au centre étant l'ontologie et la partie la plus basse, les instances, ceci est montré par la figure 3.4.

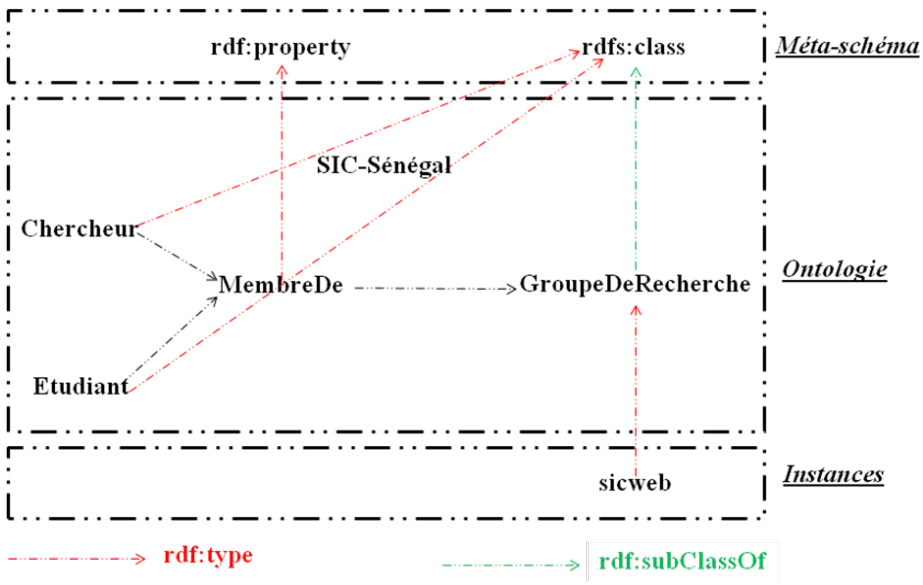


Figure 3-4 Exemple de Schéma-RDF

RDF/RDFS est un modèle de métadonnées puissant largement accepté et utilisé mais trop limité pour formuler des contraintes sémantiques plus riches et raisonnées. C'est ainsi que le W3C a proposé le langage OWL.

3.4.3 Ontology Web Language

OWL (Web Ontology Language) [McGuinness et Harmelen, 2004] est inspiré de DAML (DARPA) et OIL (EEC). OWL incorpore les dispositifs de base de RDFS (schema de RDF) : définitions de classe, de propriété, de domaine, de sous-classe et de secondaire-propriété. OWL, produit du Working Draft W3C de février 2003 est une extension de RDFS, en vue de faciliter le partage, l'intégration d'ontologies et de meta-données, contrairement à RDFS qui considère la définition d'une ressource comme étant l'union de ses descriptions.

OWL est un langage basé sur RDF et sur une sémantique formelle définie par une syntaxe rigoureuse. Ce langage enrichit le modèle des RDF-Schemas en définissant un vocabulaire riche pour la description d'ontologies complexes.

RDF-Schema définit le plus petit nombre de concepts et de propriétés nécessaires à la définition d'un vocabulaire simple. Il s'agit essentiellement : des notions de classe, ressource, littéral et les propriétés de sous-classe, de sous-propriété, de champ de valeur, de domaine d'application. OWL est un langage beaucoup plus riche qui, aux notions définies par RDF-Schema, ajoute les propriétés de classe équivalente, de propriété équivalente, d'identité de deux ressources, de différences de deux ressources, de contraire, de symétrie, de transitivité, de cardinalité, etc., permettant de définir des rapports complexes entre des ressources[31].

[31] http://websemantique.org/Vocabulaire

Une propriété relie une classe à une classe ou à un littéral. Les types de données de XML-Schema sont supportés pour les propriétés littérales. En outre, le langage OWL fournit des constructions pour définir des rapports entre les classes telles que l'équivalence et les expressions d'ensemble (par exemple, A est l'union de C et B prive de D). Des propriétés peuvent également être qualifiées comme fonctionnelles, symétriques, ou transitives ; deux propriétés peuvent être déclarées en tant qu'inverse. Les cardinalités sont prolongées et des restrictions peuvent être imposées aux valeurs de propriété. En conclusion, l'information et des annotations, versionages peuvent être définies pour des ontologies [Gardarin et Dang-Ngoc, 2004].

Il existe trois variantes d'OWL:

— OWL Lite est destiné à des tâches qui exigent une taxonomie ou classification hiérarchique. Cette version reprend une restriction de RDFS enrichit avec de nouvelles primitives, il ne supporte qu'un sous-ensemble des constructions du langage OWL;

— OWL DL Contrairement à la version OWL Lite contient toutes les primitives OWL avec des contraintes garantissant la décidabilité du langage ;

— Contrairement à OWL DL, OWL Full n'imposant pas de restrictions est la version indécidable d'OWL.

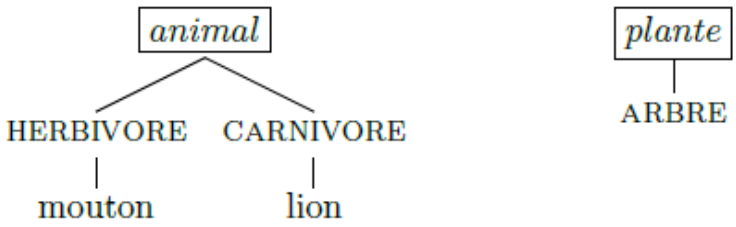


Figure 3-5 Exemple de structure ontologique

Considérons l'exemple de représentation d'une ontologie réduite de la flore et des animaux et la représentation graphique de l'architecture des concepts et instances de l'ontologie.

De la structure de gauche de la figure 3.5 on peut distinguer un système de représentation logique permettant de spécifier que les animaux comptent deux groupes : celui des herbivores et celui des carnivores. Nous spécifions deux instances pour chaque groupe d'animaux. La structure montre aussi que les arbres sont des plantes.

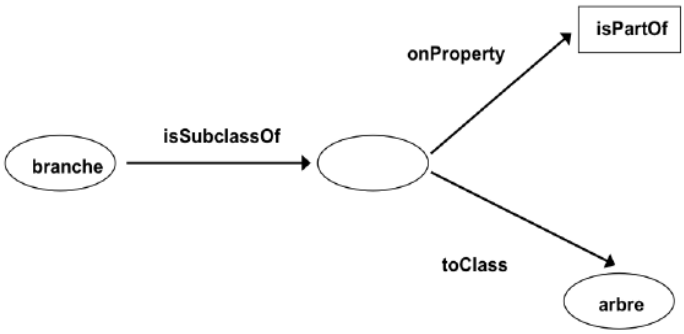
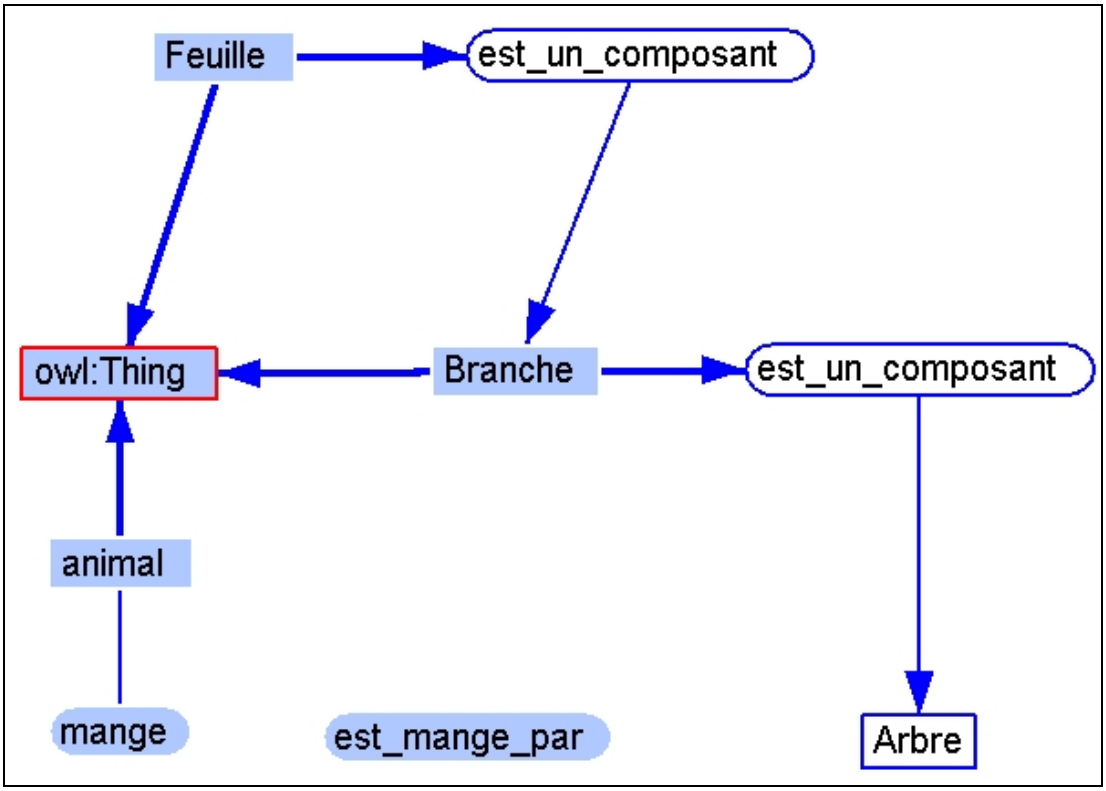


Figure 3-6 Cas d'utilisation de multiples ontologies

En utilisant les ontologies, il est possible de spécifier que les branches font parties des arbres et qu'aussi une branche à des feuilles. La figure 3.6 donne un exemple de présentation schématique spécifiant que les branches sont une composante des arbres et les feuilles sont une composante des branches.

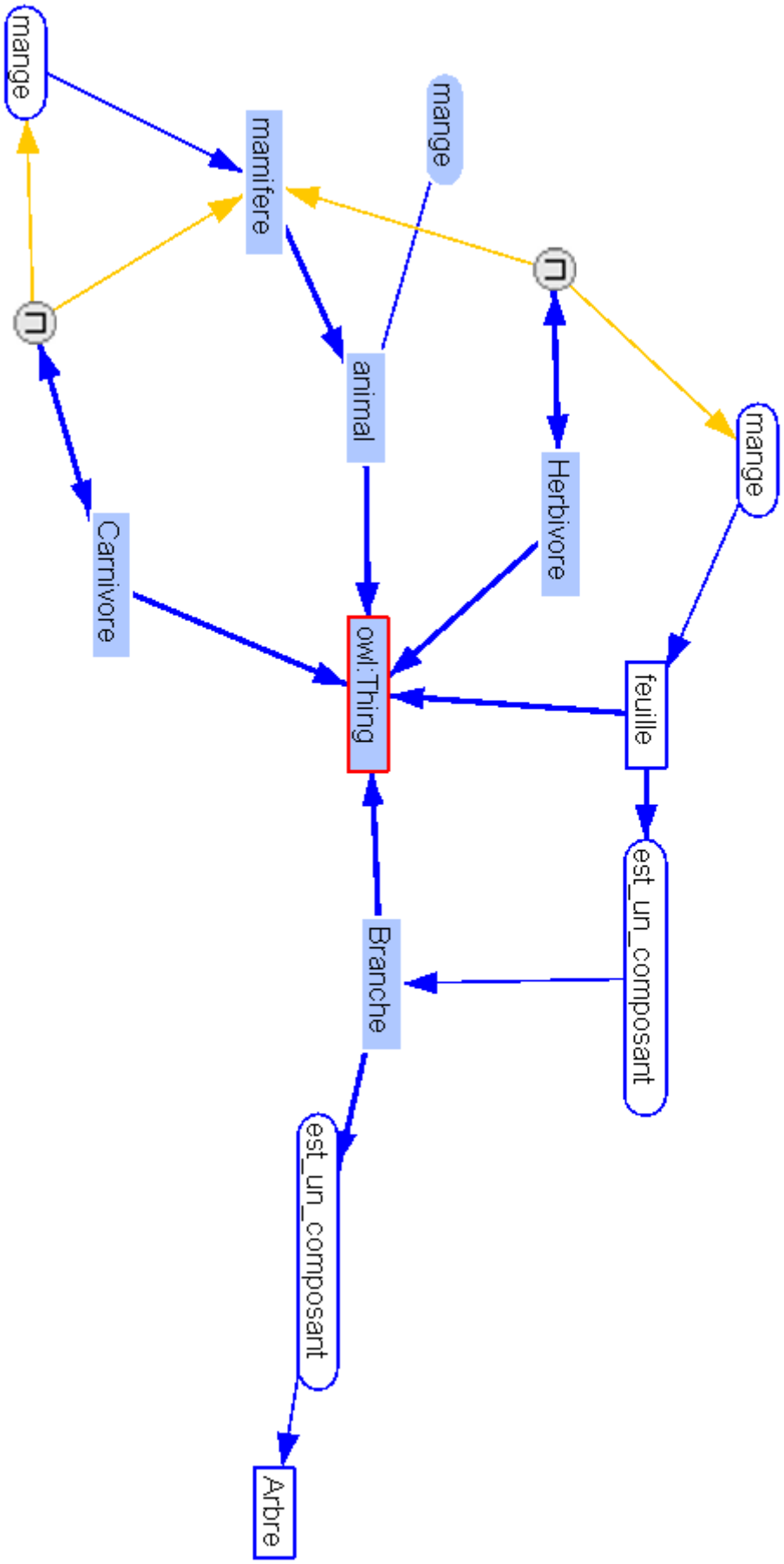
Ainsi nous pouvons, comme le montre la représentation graphique ci-dessous spécifier que qu'une *« branche »* est un composant d'un *« arbre »* qui a lui-même une *« feuille »* comme composant. De même que spécifier la propriété qu'un *« animal » « mange »* et que la relation inverse de la propriété *« mange »* est *« mangé-par »*.



Cette représentation schématique est ainsi représentée en OWL suivant le code ci-dessous :

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns="http://www.owl-ontologies.com/Ontology1193075118.owl#"
  xmlns:vcard="http://www.w3.org/2001/vcard-rdf/3.0#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:jms="http://jena.hpl.hp.com/2003/08/jms#"
  xmlns:rss="http://purl.org/rss/1.0/"
  xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 <owl:Ontology rdf:about=""/>
 <owl:Class rdf:about="http://www.w3.org/2002/07/owl#Thing"/>
 <owl:Class rdf:ID="Feuille">
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:hasValue>
     <owl:Class rdf:about="#Branche"/>
    </owl:hasValue>
    <owl:onProperty>
     <owl:ObjectProperty rdf:ID="est_un_composant"/>
    </owl:onProperty>
   </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
 </owl:Class>
 <owl:Class rdf:ID="Branche">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:hasValue>
     <owl:Thing rdf:ID="Arbre"/>
    </owl:hasValue>
    <owl:onProperty rdf:resource="#est_un_composant"/>
   </owl:Restriction>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="animal">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
 </owl:Class>
 <owl:ObjectProperty rdf:ID="est_mange_par"/>
 <owl:ObjectProperty rdf:ID="mange">
  <rdfs:domain rdf:resource="#animal"/>
  <owl:inverseOf rdf:resource="#est_mange_par"/>
 </owl:ObjectProperty>
</rdf:RDF>
```

Le langage OWL permet de définir une restriction sur une collection, nous savons par exemple que les carnivores sont des mammifères, mais dans cette classe, ce sont ceux qui mangent d'autres mammifères. Cela peut se définir en faisant une restriction dans la classe *carnivore* spécifiant sa définition et préciser que le concept de *carnivore* est une collection de mammifères qui mangent des mammifères. Reprenons l'exemple précédent et rajoutons ces nouveaux concepts et propriétés.


mange
mamifere
mange
animal
Herbivore
mange
Carnivore
owl:Thing
feuille
est_un_composant
Branche
est_un_composant
Arbre

Il suffit de rajouter au code précédent celui-ci :

```
…
 <owl:Class rdf:ID="Carnivore">
 <owl:intersectionOf rdf:parseType="Collection">
  <owl:Class rdf:about="#mamifere" />
  <owl:Restriction>
   <owl:onProperty rdf:resource="#mange" />
   <owl:hasValue rdf:resource="#mamifere" />
  </owl:Restriction>
 </owl:intersectionOf>
 </owl:Class>

 <owl:Class rdf:ID="Herbivore">
 <owl:intersectionOf rdf:parseType="Collection">
  <owl:Class rdf:about="#mamifere" />
  <owl:Restriction>
   <owl:onProperty rdf:resource="#mange" />
   <owl:hasValue rdf:resource="#feuille" />
  </owl:Restriction>
 </owl:intersectionOf>
 </owl:Class>

 <owl:Class rdf:ID="mamifere">
       <rdfs:subClassOf rdf:resource="#animal"/>
 </owl:Class>
…
```

3.4.4 Simple Protocol And RDF Query Language (SPARQL)

SPARQL[32] est un protocole et un langage de requête similaire au SQL des bases de données. Il a été développé pour des requêtes sur des graphes RDF. Cette spécification permet à l'utilisateur de s'abstraire du format de représentation des données sources.

Une requête SPARQL est un n-uplet (GP, DS, SM, R) où :

— GP est un motif de graphe (motif de la requête) ;

— DS est un ensemble de données RDF;

— SM est un "transformateur de solution" : Projection, Distinct, Order, Limit, Offset

[32] http://www.w3.org/TR/rdf- sparql-query/

— R est un format de résultat : SELECT, CONSTRUCT, DESCRIBE, ASK

Exemple de requête SPARQL :

```
PREFIX ugb :<http://www.ugb.sn/dess#>
SELECT ?etudiant
WHERE {
        ?etudiant ugb:inscrit ?x.
        ?x ugb:siteweb Http://www.ugb.sn
}
Le résultat de la requête précédente en binding(liste des valeurs sélectionnées pour
chaque réponse rencontrée) en format XML
<head>
        <variable name="etudiant"/>
</head>
<results ordred="false" distinct="false">
        <result>
                <binding name="etudiant">
                        <uri>Http://www.ugb.sn/data.rdf#niang</uri>
                </binding>
        </result>
        <result>
                <binding name="etudiant">
                        <uri>Http://www.ugb.sn/data.rdf#geuye</uri>
                </binding>
        </result>
</results>
```

3.5 Extraction automatique d'ontologie à partir de données

Il existe différentes méthodologies de construction d'ontologie, mais c'est généralement celle d'Uschold et Grüninger qui est utilisée. Nous nous intéressons ici à la construction automatique des ontologies à partir de documents ou données semi-structurées en s'aidant de la réutilisation d'ontologies.

3.5.1 Extraction d'ontologie par apprentissage automatique

Dans le domaine de l'apprentissage automatique, il existe quelques résultats propres à un domaine, par analyse de corpus de textes. Dans ce cas, le système construit une liste des principaux termes récurrents et tente de les relier en utilisant un dictionnaire ou un glossaire et une base de règles grammaticales. Il fournit en sortie une ontologie des termes du domaine, mais ce résultat reste partiel : l'utilisateur doit encore le corriger et l'affiner.

Nous avons aussi des approches comme [Kietz et al., 2000], [Navigli et al., 2003] et aussi dans [Maedche et Staab, 2001] qui utilisent des techniques statistiques et d'apprentissage automatique pour construire une ontologie. Ces méthodologies peuvent être classées en deux groupes. Les méthodes de construction basées sur des textes non structurés comme la méthodologie TERMINAE ([Aussenac-Gilles et al., 2000(a)]), [Aussenac-Gilles et al., 2000(b)]) repose sur l'analyse de corpus linguistiques. D'un autre côté nous avons celles transformant un thésaurus en ontologie comme [Miles et al., 2003], [Wielinga et. al ,2001] ou [Clark et. al , 2000]. D'autres approches de cette classe réutilisent une ontologie existante ou une hiérarchie de concepts comme WordNet [Fellbaum, 1998], GermaNet[33], SemCor [Miller et al., 1993] comme *base de connaissances*.

3.5.2 Extraction d'ontologie à partir de sources XML

XML est actuellement devenu un standard reconnu en terme d'interopérabilité pour échange des données au point de vue syntaxique. Cependant l'utilisation du langage XML pour représenter les données présente rapidement des limites lorsque la nécessité s'impose d'intégrer des sources de données différentes. Cette difficulté est due à l'hétérogénéité sémantique pouvant exister entre les sources sur les choix du vocabulaire permettant de représenter les connaissances sur les données. Avec les ontologies, il est possible de représenter formellement le modèle du domaine de connaissance partagé avec les concepts, les attributs, les relations et les instances [Rodrigues et al., 2006].

Il n'existe cependant encore à ce jour aucun standard même si un nombre de travaux assez abondants sont menés dans le domaine du mapping des données ou schéma de données XML vers des ontologies.

Des approches comme celle de [Klein, 2002] proposent un algorithme de mapping avec une procédure de transformation directe des données XML en données RDF, en annotant les documents XML via des spécifications RDFS externes.

[33] http://www.sfs.uni-tuebingen.de/lsd/

Nous avons l'initiative du W3C nommée GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [Hazaël-Massieux, 2005], [Hazaël-Massieux et Connolly, 2005], signifiant selon l'encyclopédie en ligne Wikipédia[34] littéralement « récolte de descriptions de ressources à partir des dialectes de langages ». Un langage de marquage permet d'extraire des données RDF à partir de documents XML ou XHTML en utilisant des algorithmes de transformation explicitement liés, typiquement représentée en XSLT[35]. C'est un système de mapping explicite, pour un fichier XML. L'auteur devra implémenter un espace de nom, un attribut et un chemin pointant vers le chemin pour accéder au fichier contenant l'algorithme à utiliser pour l'extraction de ses données. Il est maintenant embarqué dans Jena, une première version de « Jena [Carroll et al., 2003] GRDDL Reader »[36] est actuellement disponible. Dans Jena et dans un environnement java, elle permet d'utiliser cet outil.

Cette approche est intéressante d'autant plus que nous disposons sur le plan applicatif de données organisées sous forme tabulaire. En outre, il est proposé un mécanisme permettant de passer directement des feuilles aux données RDF. Cependant, cette liberté laissée à l'utilisateur de spécifier l'algorithme permettant d'extraire ses métadonnées RDF et éventuellement le langage pour l'interprétation pose évidemment un réel problème de sécurité[37]. Ce risque est limité mais qu'en est-il du passage à l'échelle et aussi des risques après déploiement ? Pour ne pas prendre de risques, nous avons préféré utiliser un métier propre au *SIC*.

Il existe aussi des approches comme celle dans [Clark et. al , 2000] pour incorporer les paradigmes RDF et XML. Ils ont développé un modèle d'intégration pour XML et RDF en intégrant la sémantique et les règles d'inférences de RDF à XML, de sorte que les requêtes puissent bénéficier des possibilités de raisonnement de RDF. Cependant, cette approche ne résout pas la problématique de l'interrogation de sources hétérogènes, avec des syntaxes et modèles différents.

Lakshmanan et Sadri [Lakshmanan et Sadri ,2003] proposent également une infrastructure pour l'interopérabilité entre sources de données XML structurant sémantiquement l'information véhiculée par les données en utilisant un vocabulaire spécifique commun. Cependant, l'approche proposée se fonde sur la disponibilité d'une ontologie standard spécifique à l'application devant servir de schéma global. Des approches similaires sont proposées dans [Reif et. al, 2005] et [Battle, 2004]. [Reif et. al, 2005] propose

[34] http://fr.wikipedia.org/wiki/GRDDL
[35] http://www.w3.org/TR/xslt
[36] http://jena.sourceforge.net/grddl/index.html
[37] http://www.yoyodesign.org/doc/w3c/grddl/#sec

une approche dans le contexte d'un projet nommé WESA (Web Engineering for Semantic Web Applications) pouvant être utilisée pour générer des métadonnées RDF à partir de schémas de documents XML. Cette phase de construction se passe en deux étapes : d'abord, durant la phase de conception du schéma XML, un mapping avec les ontologies doit être défini. Ensuite pour chaque document XML, les règles de mapping définies dans l'étape précédente sont appliquées pour générer la représentation RDF. Cette méthodologie requiert cependant que les règles de mapping d'XML-Schema à une ontologie OWL soient générées manuellement.

Dans [Matthias et al., 2004] est proposée une méthodologie de mapping d'XML à RDF et aussi d'XML-Schema à OWL. Cependant, cette méthodologie ne traitant pas de la création d'un modèle OWL dans le cas où aucun schéma XML n'est disponible, elle ne convient pas dans notre contexte.

Dans [Bohring et Auer, 2005], les auteurs proposent une méthodologie de construction automatique d'ontologie à partir de données XML en procédant à un mapping entre XML et OWL. Cette approche se base sur les propriétés proposées par RDFS permettant la création d'hiérarchie de classe et leurs propriétés pour exploiter la structure des documents XML afin de générer automatiquement les classes qui correspondent aux concepts de l'ontologie. Comme nous le verrons, la grande difficulté de partir des documents XML est l'extraction automatique des relations autres que hiérarchiques. En effet, seules les relations hiérarchiques de composition sont explicitement représentées. Les auteurs de cette approche ne s'occupent que de l'extraction des relations structurelles en considérant par exemple que tout nœud fils terminal est lié à son père par une relation de type « sous-partie-de ». Un ensemble de règles de correspondance entre XML-Schema et OWL permet de spécifier que tout nœud terminal et attribut de nœud est traduit dans l'ontologie comme un « owl:DatatypePrpoperties ».

Dans [Isabel et al., 2004] aussi est proposée une méthodologie analogue dans un contexte d'intégration de documents XML avec la génération automatique d'une ontologie OWL pour chaque document puis la fusion de ces dernières pour constituer une *ontologie globale* à la source.

Dans [Rodrigues et al., 2006] est proposé un framework java « JXML2OWL » incorporant un mécanisme permettant de mettre en correspondance un schéma XML en une ontologie OWL existante. Contrairement aux autres approches où un mapping direct à partir d'XML-Schéma est proposé vers OWL, ici, les auteurs proposent à l'utilisateur de spécifier en premier lieu les mappings, ensuite une feuille de style générée sert à transformer les instances d'un schéma XML en instances d'une ontologie existante.

Dans [Bohring et Auer, 2005] est proposé un mécanisme permettant de procéder à un mapping d'XML-Schema ou un schéma de fichier XML généré vers une ontologie OWL. C'est une approche très semblable à celle que nous utilisons en dehors du fait que nous n'utilisons pas de schéma XML. Un ensemble de règles de mapping entre chacune des possibilités de représentation dans le langage XSD (XML-Schema Definition) à celle du vocabulaire OWL est proposée. Le tableau 3.1 montre les correspondances proposées résumant globalement le consensus utilisé pour le mapping. Nous ne nous écartons pas de cette approche de base. La seule particularité de notre méthodologie est de passer à l'identification des types d'éléments XSD cités dans ce tableau dans les sources XML et de passer à leur mapping en OWL.

Enfin, une approche XR (XML -> RDF)[Visscher, 2005] propose un cadre permettant l'extraction de données RDF à partir de XML. Afin de définir une transformation pour un document XML donné, sa structure définie en XML-Schéma est utilisée pour définir des expressions Xpath. RDF peut alors être utilisé pour transformer automatiquement tout format XML d'un document conforme à ce document de base.

XSD	OWL
xsd :elements, ayant des éléments fils ou au moins un attribut	**owl:Class, couplé avec owl:ObjectProperties**
xsd :elements, n'ayant ni fils ni attribut	**owl:DatatypeProperties**
xsd :complexType	**owl:Class**
xsd :SimpleType	**owl:DatatypeProperties**
xsd :minOccurs, xsd :maxOccurs	**owl:minCardinality, owl:maxCardinality**
xsd :sequence, xsd :all	**owl:intersectionOf**
xsd :choice	**combinaison de owl:intersectionOf, owl:unionOf et owl:complementOf**

Tableau 3-1Tableau des mapping d'XML-Schéma à OWL ([Bohring et Auer, 2005])

3.6 Approches d'intégration de données basées sur les ontologies

Les ontologies sont utilisées dans de nombreux travaux pour résoudre les problèmes d'intégration. C'est le cas dans [Cui et al., 2001], [Chong et al., 2002], [Buccella et Cechich, 2003] et [Cruz et al., 2004]. Cette utilisation des ontologies permet d'abstraire et de capturer la connaissance du domaine des données cibles. Lorsqu'il s'agit d'intégrer sémantiquement plusieurs sources de données, il conviendra de faire des choix sur la manière de prendre en

compte la structure sémantique décrivant les connaissances de chacune. Dans une approche comparative aux groupes sociaux (c'est le cas dans la nouvelle génération du web), il arrive que chaque groupe utilise son propre langage pour décrire les mêmes objets que les autres. Dans ce cas, afin de pouvoir faire coopérer ou mettre en commun les connaissances des différentes communautés, trois alternatives se présentent. Il faut :

— utiliser un langage externe aux groupes servant de pivot ;
— ou bien évoluer vers un même langage partout ;
— ou entre les groupes sociaux, définir les équivalences binaires de chaque terme utilisé.

Comparativement, c'est le même principe que la taxonomie de l'utilisation des ontologies pour l'intégration de plusieurs sources de données disposant chacune d'une ontologie comme définie dans [Wache et al., 2001]. Nous allons dans cette partie les étudier afin de justifier le choix d'architecture des ontologies dans le contexte de l'*approche dataweb sémantique*.

3.6.1 Approche mono-ontologie

Dans l'approche dite « single ontology », une *ontologie globale* est utilisée pour toutes les sources, proposant ainsi un vocabulaire partagé pour la spécification de la sémantique. L'un des avantages de ce cas d'utilisation des ontologies est de pouvoir disposer d'un vocabulaire partagé. Il présente également l'avantage d'unifier les requêtes, dans le cas où la méthode d'intégration utilise un processus d'unification des requêtes.

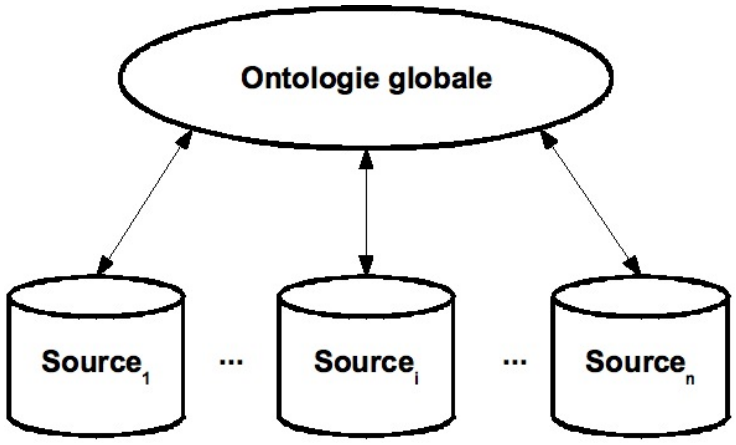


Figure 3-7 Cas d'utilisation d'une *ontologie globale*

Selon [Wache et al., 2001] un inconvénient à utiliser cette approche est qu'elle semble ne bien s'adapter qu'aux problèmes d'intégration où les sources d'informations à intégrer ne présentent qu'une vision d'un seul domaine. Dans le cas contraire, cette approche devient problématique, chaque source référençant exactement la même ontologie, sans possibilité d'extension ou d'adaptation.

Selon [Bellatreche et al., 2004] une des limitations de ces systèmes est aussi qu'une fois l'ontologie partagée définie, chaque source doit utiliser le vocabulaire commun. L'ontologie partagée est globale, et en conséquence, chaque source locale a moins d'autonomie.

Le processus utilisant ce système est lourd à réaliser avec des sources nombreuses et très hétérogènes. Il nécessite aussi une re-génération de l'ontologie de haut niveau à chaque modification d'une composante, ce qui est assez coûteux.

3.6.2 L'approche multi-ontologie

Dans l'approche « multiple ontology », les ontologies sont considérées comme étant des représentations approximatives représentant le point de vue d'une communauté ou d'un individu [Bouquet et al., 2003], et à les faire correspondre (mapping) l'une à l'autre. Nous avons donc une ontologie par source de données, et chacune est indépendante.

Cette approche permet une plus grande flexibilité, permettant d'utiliser des ontologies évoluant de manière autonome et mises à jour fréquemment. De surcroît, elle n'est pas bloquante si l'une d'elles vient à disparaître et il est facile de supprimer une source (il suffit de supprimer les correspondances (mapping) avec l'ontologie locale). Mais l'appariement (matching) entre représentations autonomes est difficile. Il requiert parfois l'utilisation de ressources linguistiques spécialisées, et nécessite jusqu'à présent l'intervention d'experts des domaines : actuellement, aucun outil ne permet la résolution automatique de l'appariement. Cette approche est utilisée dans OBSERVER (Ontology Based System Enhanced with Relationships for Vocabulary hEterogeneity Resolution).

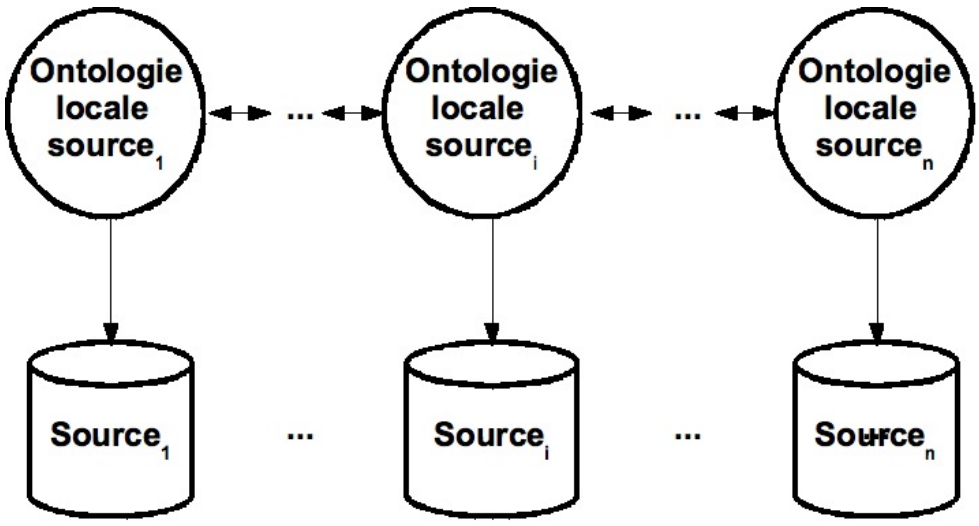


Figure 3-8 Cas d'utilisation de multiples ontologies

3.6.3 L'approche hybride

Dans l'approche dite « hybride » comme alternative aux deux approches précédentes, des correspondances (mappings) entre les ontologies locales (une ontologie par source) sont établies. Pour chacune d'elles, on établit aussi des correspondances avec une unique ontologie de plus haut niveau « upper level ontology ».

Comme pour l'approche précédente, sources et ontologies peuvent être développées de manière entièrement autonomes. On procède simplement à une mise à jour lorsqu'une ou plusieurs ontologies évoluent. Cette approche est surtout intéressante si toutes les ontologies sont conformes à un certain standard.

Dans [Wache et al., 2001], il est proposé que toutes les ontologies locales soient décrites à l'aide d'un vocabulaire partagé (qui peut être une ontologie [Stuckenschmidt et al., 2000]) comprenant les termes basiques du domaine : de nouveaux termes, plus complexes, pourront être créés à partir de combinaisons des premiers, à l'aide d'opérateurs spécifiques. Cela requiert donc de commencer par créer le vocabulaire commun, ainsi que les règles de combinaison des termes. L'un des inconvénients est lié au fait que les ontologies ne peuvent être réutilisées facilement. Mais elles doivent être redéveloppées à partir de rien. En effet, toutes les ontologies se réfèrent au vocabulaire partagé [Wache et al., 2001].

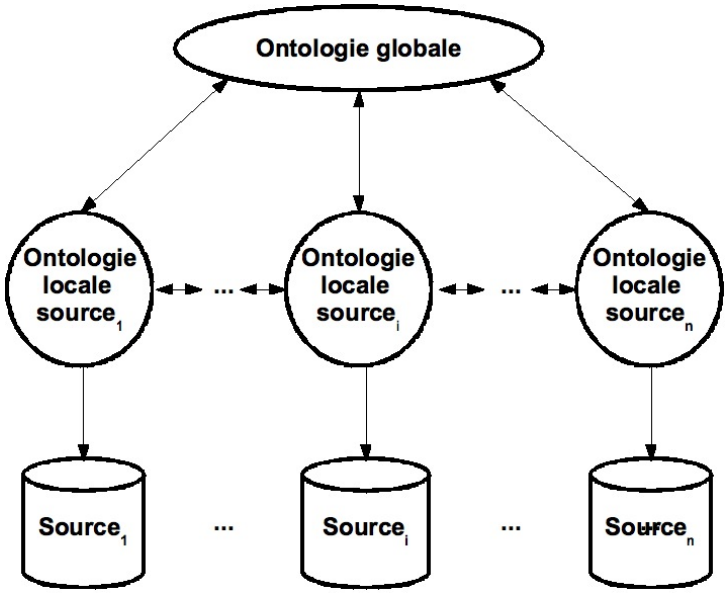


Figure 3-9 Architecture de l'approche hybride

Les difficultés de ces approches restent liées à deux phases. D'une part, il faut souvent procéder à une phase d'adaptation où l'on exploite l'aspect réutilisation des ontologies. D'autre part, il faut nécessairement les mettre en correspondance et aussi analyser le pourquoi

de ce rapprochement. Nous devrons sûrement réfléchir à chacune de ces problématiques dans le cadre du *SIC*-Sénégal. Cette approche a été adoptée dans COIN [Goh, 1997], MECOTA [Wache et.al, 1999] et BUSTER [Stuckenschmidt et al., 2000].

3.7 Conclusion

Ce chapitre aborde dans un premier temps la définition et l'utilisation des ontologies. Ensuite, il présente la formalisation des modèles d'ontologie. Le modèle d'ontologie le plus adapté à notre contexte est le modèle des ontologies à base lexicale dont nous avons examiné la structuration. Par ailleurs, nous avons étudié les langages permettant de représenter les modèles d'ontologie et de les interroger.

Nous avons aussi rendu compte de l'extraction de la sémantique formelle permettant de manipuler des données. En outre, trois approches d'intégration basées sur les ontologies ont été exposées. Il s'agit des approches dites mono-ontologie, multi-ontologie et de l'approche hybride, qui se trouve plus adaptée à notre contexte.

Il apparaît ainsi que la démarche d'utilisation des ontologies pour résoudre la problématique de l'hétérogénéité sémantique des données est un choix largement partagé dans la communauté d'intégration des données. Plusieurs architectures s'offrent comme option à l'utilisation des ontologies dans les systèmes d'intégration. Dans le contexte de l'intégration de données environnementales avec une contrainte de caractère propriétaire et privé des données, il convient d'opter pour l'architecture hybride en associant une ontologie aux sources de chaque *partenaire* avec l'utilisation d'une *ontologie globale* pour la médiation des sources. Cela pose aussi un autre problème, qui est la construction de ces ontologies, dans un contexte où la masse de données est relativement importante. Il convient de la rendre la plus automatique possible.

Dans la partie qui va suivre, nous présentons notre approche d'intégration de données environnementale basée sur l'extension de l'*approche dataweb*.

Deuxième partie

Contributions

Chapitre 4

Du modèle de dataweb au dataweb sémantique basé sur XML

Sommaire

4.1 Introduction 100

4.2 De l'intégration des données vers l'intégration des connaissances 100

4.2.1 Architecture globale du système d'intégration 100

4.2.2 Un processus d'intégration en trois phases 103

4.2.2.1 Intégration structurelle des données au sein des *partenaires* 103

4.2.2.2 Intégration sémantique des données au sein des *partenaires* 104

4.2.2.3 Médiation entre les différents *partenaires* 104

4.2.3 Les différentes couches de l'architecture 104

4.2.3.1 Sources natives du *partenaire* 104

4.2.3.2 Représentation structurée des données *partenaires* 105

4.2.3.3 Représenter la sémantique des données *partenaires* 105

4.2.3.4 Le niveau médiateur 105

4.3 Composants et modèle du système d'intégration 106

4.3.1 Un modèle de système d'intégration par partenaire 106

4.3.2 Un modèle d'ontologie spécifique par partenaire 108

4.3.3 Un modèle formel de base d'annotations 109

4.3.4 Un modèle de base de connaissances 110

4.3.5 Un modèle formel de *dataweb sémantique* 110

4.3.6 Un modèle formel de système d'intégration par partenaire 110

4.3.7 Un modèle formel de système d'intégration globale 111

4.4 Conclusion 111

4.1 Introduction

Ce chapitre présente l'*approche dataweb sémantique*. Cette approche apporte un support sémantique à l'intégration structurelle des données, les préparant ainsi à la phase d'intégration sémantique grâce à l'extraction d'ontologie à partir du vocabulaire les décrivant.

L'intégration sémantique s'appuie donc sur l'utilisation d'une ontologie à laquelle est associée une base d'annotations faisant le lien avec les parties de documents décrits par les concepts. Ensuite, pour la médiation entre les différents *dataweb*, nous préconisons l'utilisation d'une *base de connaissances* en ajoutant aux composants déjà cités une *ontologie générique* contenant les concepts que le *partenaire* est susceptible de partager avec les autres. Nous formons ainsi un système d'intégration structurelle et sémantique des données d'un *partenaire* d'une part, et des différents systèmes d'autre part.

Dans ce chapitre, après avoir présenté les couches et niveaux de l'approche d'intégration, nous allons nous intéresser au modèle des différents composants du système d'intégration. La présentation de l'approche est l'objet de la partie 4.2 et le modèle formel des différents composants celui de la partie 4.3.

4.2 De l'intégration des données vers l'intégration des connaissances

L'approche adoptée pour réaliser l'intégration est ascendante. Elle permet une intégration selon une architecture incrémentale qui se présente en couches superposées. Les données d'une couche supérieure sont obtenues à partir des données de la couche inférieure adjacente et chaque couche se situe à un certain niveau d'abstraction du système d'intégration. Ainsi, le processus général d'intégration se déroule en quatre phases dont chacune est dépendante de la précédente.

4.2.1 Architecture globale du système d'intégration

Nous proposons une approche d'intégration prenant en compte ces différents aspects et composée de trois phases :

— une première phase de pré-intégration permet la construction des entrepôts de documents XML,

— une deuxième phase est celle de la construction des ontologies à base lexicale OWL *partenaires*,

— une troisième phase de médiation entre les différents systèmes d'intégration basée sur les ontologies. De manière globale c'est un système de médiation décentralisé basé sur des moteurs de recherche sémantique auxquels sont associés une *base de connaissances* [Sall et al., 2009]. Cette *base de connaissances* est constituée par une *base d'annotations* décrivant les données du *partenaire*, une *ontologie générique* et une ontologie spécifique.

Pour résoudre la problématique de l'hétérogénéité structurelle, nous avons proposé l'introduction d'une phase préalable de pré-intégration par une représentation de l'ensemble des données *partenaires* sous XML puis une restructuration de chaque document. Cela constitue ainsi un *entrepôt de documents XML* ou *dataweb* [Lo, 2002]. Les données sont ainsi chargées, extraites puis transformées de leur format initial sous forme de table de données au format XML. Nous avons ainsi les deux premières phases « Extract and Transform » du processus ETL (Extract, Transform and Load - extraction, transformation et chargement). L'extraction est la phase d'identification et de collecte des informations pertinentes pour la construction de l'entrepôt. La transformation vise généralement, entre autres, à disposer d'une structuration et d'une mise en évidence de certaines caractéristiques des données, comme c'est le cas dans notre contexte.

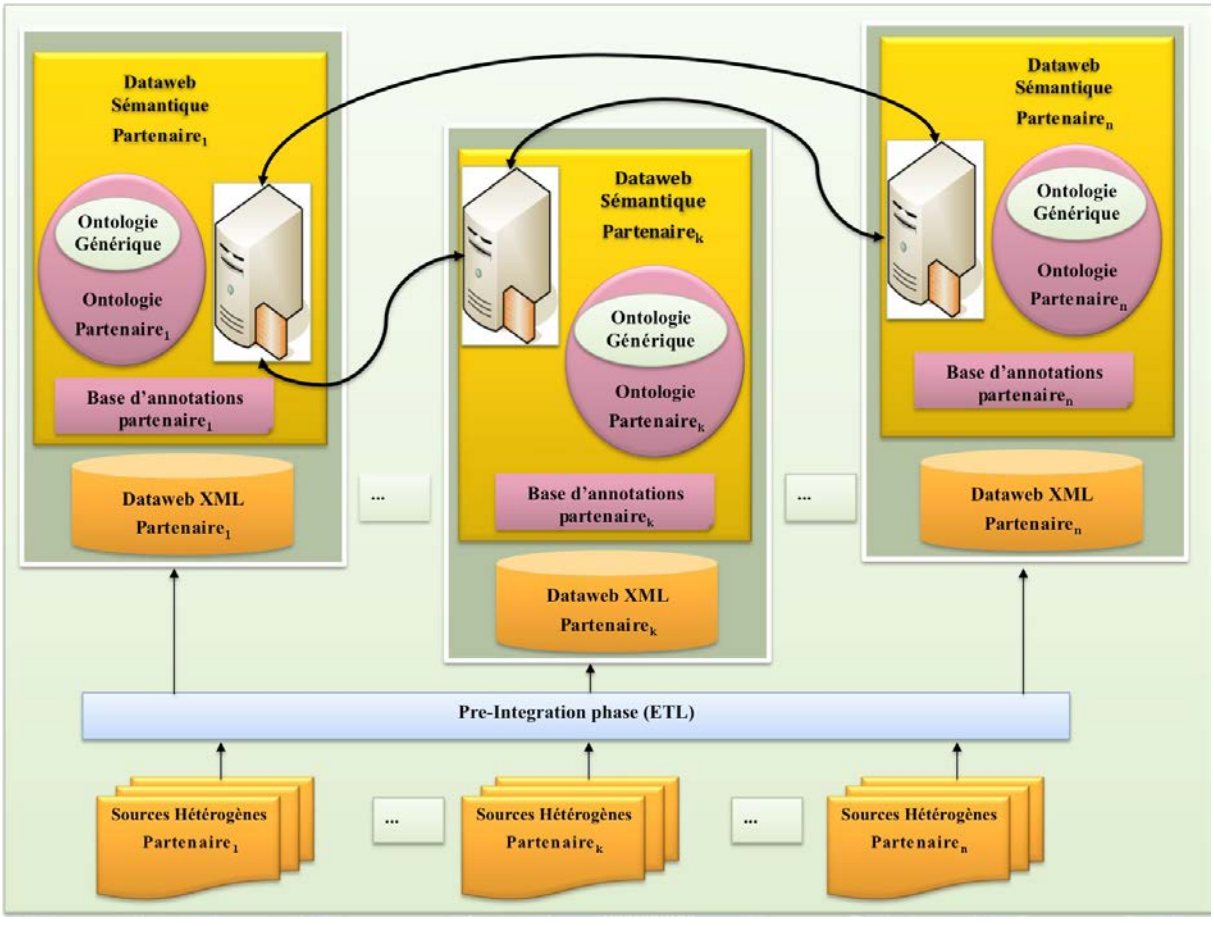


Figure 4-1 Architecture globale du système d'intégration

Afin de résoudre la problématique de l'hétérogénéité sémantique, notre approche consiste à utiliser le vocabulaire contrôlé décrivant les données de chaque *partenaire* comme un support pour la construction d'une ontologie dite *partenaire*. Cette phase d'intégration sémantique est basée sur la réutilisation d'ontologie. Dans le contexte applicatif du projet *SIC*-Sénégal, l'ontologie est basée sur le thésaurus AGROVOC de la FAO (Food and Agriculture Organisation) compte tenu de la nature agricole (environnementale) des données manipulées.

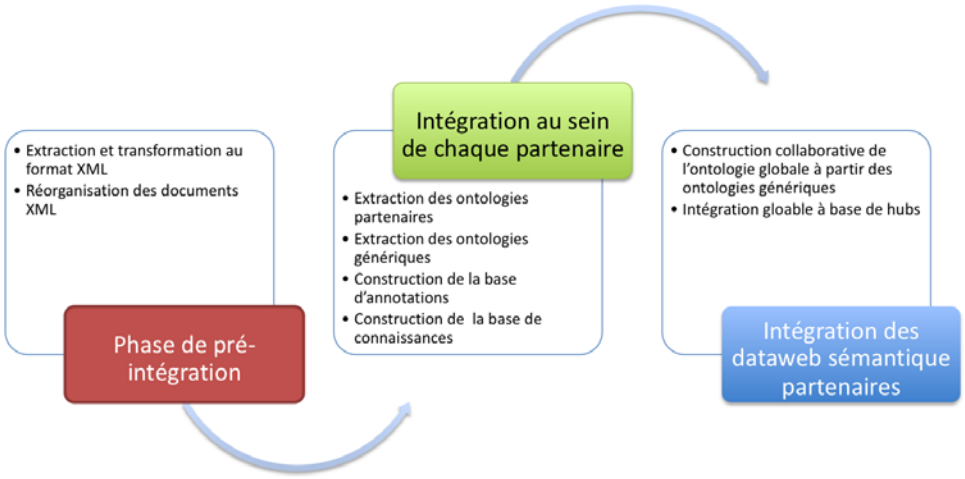


Figure 4-2 Découpage du processus d'intégration des données et des partenaires

Nous avons proposé un processus d'identification et d'extraction automatique des concepts des ontologies *partenaires*. Ces concepts sont, après extraction, subsumés à ceux de l'ontologie basée sur le thésaurus AGROVOC de la FAO. Ce processus de subsomption permet d'associer une signification plus forte à chaque concept et l'inférence de connaissances sur et entre les concepts candidats des ontologies *partenaires*.

Cette phase est une contribution à la phase non moins importante d'extraction des connaissances à partir des données du système d'intégration sémantique proposé. Elle est dite processus automatique ECD (Knowledge Discovery in Databases, KDD) [Fayyad et al., 1996], [Fayyad et al., 2001]. Dans la littérature, c'est un processus permettant l'extraction à partir de données d'une information inconnue auparavant et potentiellement utile [Frawley et al., 1991]. L'information inconnue dans notre exemple est celle des concepts et la sémantique déduite de l'ontologie basée sur le thésaurus AGROVOC de la FAO à partir du vocabulaire contrôlé de chaque *partenaire*. Elle est importante pour une compréhension et une combinaison des données des *partenaires*. L'occurrence de ces concepts candidats dans l'ontologie basée sur le thésaurus AGROVOC de la FAO justifie la validité et pertinence de la

méthodologie d'identification utilisée. La combinaison de ces concepts dans le contexte du projet *SIC-Sénégal* valide la notion de *partenaire* ou partenariat choisie désignant une mise en commun des hétérogénéités pour atteindre une même finalité. Il s'agit notamment de la finalité de l'intégration participative des données distribuées et propriétaires sur la vallée du fleuve Sénégal.

4.2.2 Un processus d'intégration en trois phases

La figure 4.3 présente les 4 niveaux principaux du processus général d'intégration. Nous les détaillons dans la section qui suit.

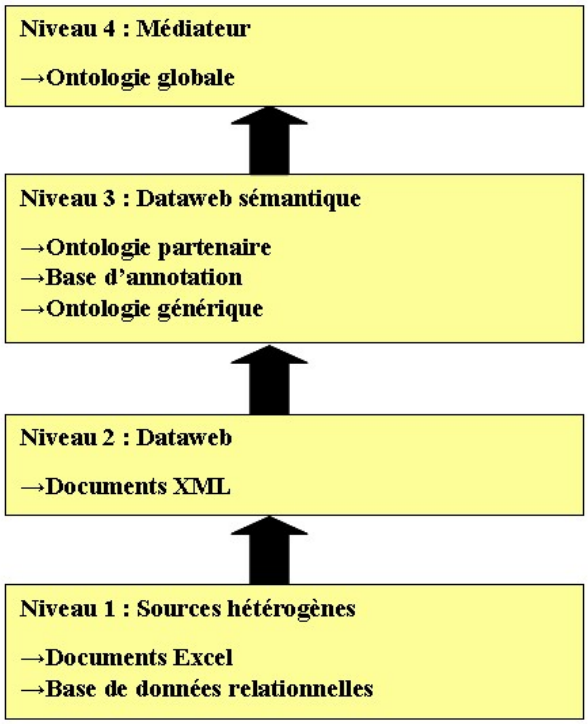


Figure 4-3 Description des 4 niveaux du processus d'intégration

4.2.2.1 Intégration structurelle des données au sein des *partenaires*

La première phase est la résolution des problèmes liés à l'hétérogénéité structurelle des sources des *partenaires*. Cela se fait selon une *approche dataweb* [Lo, 2002], consistant en la réalisation, pour chaque *partenaire*, d'un *entrepôt de documents XML* intégrant les sources de données dudit *partenaire*. Ceci permet de disposer d'un format de représentation commun pour toutes les données du *partenaire* ainsi que d'un vocabulaire contrôlé pour ledit *partenaire*.

4.2.2.2 Intégration sémantique des données au sein des *partenaires*

La deuxième phase est celle de l'intégration des données de chaque *partenaire* du point de vue sémantique. Nous la réalisons en ajoutant une couche sémantique au système d'intégration. Il s'agit de construire une ontologie OWL, dite *partenaire*, à partir des documents XML de l'entrepôt de chaque *partenaire*. Elle a pour but de décrire la sémantique de leurs données, et de définir l'ensemble des correspondances entre cette ontologie et les documents XML par l'intermédiaire d'une *base d'annotations*. Simultanément, nous construisons une ontologie dite générique contenant l'ensemble des concepts de l'ontologie *partenaire* que ce dernier désire partager avec les autres.

4.2.2.3 Médiation entre les différents *partenaires*

La troisième phase est celle de la médiation entre les différents *partenaires* afin qu'ils puissent partager leurs données. Cela nécessite la définition d'un vocabulaire commun à tous les *partenaires*. Nous construisons, à partir des ontologies génériques des différents *partenaires*, une *ontologie globale*, commune à tous ces *partenaires*, à l'effet de recevoir des requêtes et d'offrir une interface commune pour l'interrogation des sources.

4.2.3 Les différentes couches de l'architecture

La figure 4.3 montre une vue sur la représentation des 4 couches de l'architecture du système d'intégration. Nous détaillons chacune de ces couches dans la section qui suit.

4.2.3.1 Sources natives du *partenaire*

Le premier niveau de l'architecture est constitué par les sources natives du *partenaire*. Ce sont les données concernées par l'intégration. Dans le contexte des données environnementales, elles sont généralement stockées dans des documents contenant des structures tabulaires et des bases de données relationnelles. Les travaux effectués actuellement ont été faits sur des échantillons de données provenant uniquement de tableaux.

Ces données présentent plusieurs particularités liées, d'une part, à leur nature environnementale et donc spatio-temporelles, et d'autre part, à une représentation sous des structures différentes selon les *partenaires*. Une structuration préalable de ces données est donc nécessaire pour quelles puissent être intégrées.

4.2.3.2 Représentation structurée des données *partenaires*

Les entrepôts de données *partenaires* sont constitués par un ensemble de documents XML, obtenus après une phase d'intégration structurelle. Nous avons donc à ce niveau un ensemble de documents XML qui représentent les données du *partenaire* dans le système d'intégration.

Les entrepôts permettent de résoudre la problématique de l'hétérogénéité structurelle des données des *partenaires*, du moins celles de nature tabulaire, en disposant d'un même format de représentation et aussi du vocabulaire contrôlé de chaque *partenaire*. XML convient à la transcription du format d'origine des données tabulaires, en permettant de refléter fidèlement la structure et la hiérarchie de chaque tableau. Mais cela ne suffit pas pour associer une description assez forte aux données.

Sans cette description, il sera difficile, par exemple, pour une machine de trouver les bonnes ressources, d'analyser et d'inférer sur les données afin de donner des déductions, dégager les bonnes informations et combiner les différents résultats. Par conséquent, il s'avère nécessaire d'ajouter une couche sémantique à l'effet de décrire et de structurer le contenu des sources *partenaires* et ainsi de réaliser une intégration sémantique des données.

4.2.3.3 Représenter la sémantique des données *partenaires*

Ce niveau représente la couche sémantique du système d'intégration. Il est constitué par un *dataweb sémantique* contenant l'ontologie *partenaire*, et permet de décrire la sémantique des données, représentées ici par les documents XML, en vue de pallier aux problèmes liés à l'hétérogénéité sémantique des données. Chaque *dataweb sémantique* intègre aussi une *base d'annotations* à l'effet d'identifier les sources XML originelles des concepts dans l'ontologie *partenaire*.

Cela s'explique par le fait que l'ontologie *partenaire* sert après de support pour l'expression des requêtes qui seront ensuite exécutées sur les sources de données, c'est-à-dire sur les documents XML. De l'ontologie *partenaire*, est extrait l'ensemble des concepts nécessaires pour la construction de l'*ontologie générique* du *partenaire*, inclut aussi dans le *dataweb sémantique*.

4.2.3.4 Le niveau médiateur

Le niveau médiateur représente le point d'entrée du système puisqu'il contient l'*ontologie globale* utilisée pour l'interrogation du système. Il assure l'interrogation du système global alors que les différentes ontologies *partenaires*, assurent celle des sources. Ces

ontologies sont exploitées au moyen d'un moteur de recherche sémantique qui permet de les interroger : ce dernier assure le support d'interrogation du système de médiation.

Comme nous venons de le décrire, l'architecture du système d'intégration repose sur des couches superposées dont chacune se situe à un certain niveau d'abstraction du système. Dans la suite, nous présentons les approches que nous adoptons pour construire les différentes composantes de ces couches, notamment :

— Le niveau *dataweb* correspondant à la création d'entrepôts de documents XML, situés au deuxième niveau ;

— Le niveau *dataweb sémantique* permettant la construction d'une ontologie *partenaire*, d'une *base d'annotations* et d'une *ontologie générique*, situées au troisième niveau ;

Le système d'intégration des *dataweb sémantique*s nécessitant la construction d'une *ontologie globale* constitue le quatrième niveau.

4.3 Composants et modèle du système d'intégration

Dans cette section nous présentons les modèles des différents composants du système d'intégration. Autour du noyau constitué par la notion de *dataweb* se superpose la couche contenant la *base de connaissances*.

Chaque *dataweb sémantique* interagit avec les autres systèmes d'intégration en gravitant autour d'une *ontologie globale*. Nous adoptons une démarche de modélisation « bottom-up » pour les composants du système d'intégration.

4.3.1 Un modèle de système d'intégration par partenaire

Nous présentons ici le modèle du système d'intégration que nous utilisons chez chaque *partenaire*. Un modèle d'intégration est défini comme nous l'avons vu dans le premier chapitre par un triplet constitué par le schéma global, l'ensemble des schémas sources et le modèle de mapping entre ces deux composantes. Nous avons utilisé une ontologie comme schéma global et qui est construite à partir des documents XML. Le mapping entre ce schéma global et les schémas des sources est obtenu pendant la phase de construction de cette ontologie.

Nous modélisons, de manière formelle, un système d'intégration de données *partenaire* I_{part}, qui est ici un *dataweb*, selon le triplet :

$$I_{part} := \{G_{part}, S_{part}, \mathcal{M}_{part}\} \text{ où :}$$

— $\mathcal{G}_{part}$ est le schéma global ou médiateur représentant le domaine d'intérêt du système d'intégration appelée ici une ontologie. Elle est donc définie par le couplet $(\mathcal{S}_{ontopart}, \mathcal{L}_{part})$ où $\mathcal{S}_{ontopart} := \{\mathcal{C}_{part}, \mathcal{R}_{part}, \mathcal{A}_{part}, \mathcal{T}_{part}, \mathcal{C}_{sub}, \mathcal{H}_{part}^{C}, \sigma_{\mathcal{R}}, \sigma_{\mathcal{A}}\}$ est la structure de l'ontologie et $\mathcal{L}_{part}$ son lexique. Il est exprimé avec des contraintes d'intégrité dans un langage $\mathcal{L}_{\mathcal{G}part}$ sur un alphabet $\mathcal{A}_{\mathcal{G}part}$ comprenant un symbole pour chaque élément de $\mathcal{G}_{part}$;

— Spart est l'ensemble des schémas des sources décrivant la structure des sources qui sont dans notre contexte des documents XML. Il est exprimé dans un langage $\mathcal{L}_{\mathcal{S}part}$ sur un alphabet $\mathcal{A}_{\mathcal{S}part}$,

— $\mathcal{M}_{part}$ est le mapping établissant une relation entre $\mathcal{G}_{part}$ et $\mathcal{S}_{part}$. Nous utilisons une approche d'intégration en mode *GAV*. Selon les principes de cette approche, le mapping $\mathcal{M}$ du système d'intégration est dans ce cas obtenu en associant à tout élément $\mathcal{C}_g$ du schéma global $\mathcal{G}$, une requête q_S sur le schéma des sources $\mathcal{S}$. Le mapping est un ensemble d'assertions qui pour tout concept c_g de l'ontologie *partenaire* G associe une vue sur les sources XML, de la forme :

$$c_g \zeta q_S$$

En résumé, le mapping dans une approche *GAV* est un ensemble de triplets $(\mathcal{AG}, Q(\mathcal{AS}), \zeta)$ où ζ est ici assimilable à la notion de correspondance « ≅ » et $\mathcal{AG}$ assimilable à la composante lexicale de l'ontologie comprenant l'ensemble des concepts constituant sa taxonomie. Notre démarche d'ingénierie suit une approche ascendante, étant donné que les composantes du schéma médiateur sont des vues créées à partir des schémas sources. Cela va de pair avec le principe de construction de l'ontologie servant de schéma médiateur en partant d'une déduction de ses composantes à partir des entités composant le schéma des sources XML. Cela introduit une contrainte, le mapping ne se fera pas directement avec le schéma, mais ce sera d'une vue ou élément de l'ontologie *partenaire* vers un ensemble d'entités bien situées dans les documents XML. Comme défini dans [Sall et Lo, 2007], la stratégie de mapping est basée ici sur l'homonymie dans cette phase de structuration. Nous définirons, donc chaque $\mathcal{M}_{part}^{i}$ comme étant constitué d'un ensemble de règles de la forme :

$$ri : \mathcal{G}_{part}^{j} \in \{\mathcal{C}_{part} \cup \mathcal{A}_{part}\} \zeta \{locUrl=axeXPathEntite_{k}\}$$ où :

- $\mathcal{C}_{part}$ et $\mathcal{A}_{part}$ sont les ensembles de concepts et de relations d'attribut de l'ontologie médiateur ;

- G_{part}^{j} est un élément du schéma global pouvant être soit un concept, soit une relation d'attribut de l'ontologie locale ;
- *locUrl* est l'url de base identifiant la source XML dans l'entrepôt associé à ce système d'intégration ;
- $axeXPathEntite_k$ est l'axe par *XPath*, du chemin de localisation de l'entité k concerné dans le document XML.

Donc, le mapping associe à un élément G_{part}^{j} du schéma global qui est ici l'ontologie *partenaire*, une vue ou entité sur un fragment de source XML spécifié par son chemin vers le document XML concerné dans l'entrepôt associé à ce système d'intégration en utilisant l'axe de son chemin de localisation Xpath.

L'interrogation des sources adressée au schéma global se fera par un dépliement de la requête en plusieurs pour les envoyer aux différentes vues concerné, c'est un processus complexe, nous reviendrons sur le schéma et la méthodologie adoptée.

4.3.2 Un modèle d'ontologie spécifique par partenaire

La structure d'une ontologie *partenaire* O_{part} à laquelle est associé un lexique est le huit-uplet :

$$S_{part} := \{C_{part}, \mathcal{R}_{part}, \mathcal{A}_{part}, \mathcal{T}_{part}, C_{sub}, \mathcal{H}_{part}^{C}, \sigma_{\mathcal{R}}, \sigma_{\mathcal{A}}\} \text{ où :}$$

– C_{part}, $\mathcal{A}_{part}$, $\mathcal{T}_{part}$, C_{sub} sont respectivement les ensembles contenant, les concepts de l'ontologie, les relations d'attribut, les types de données, les concepts subsumeurs de l'ontologie réutilisée;

– $R_{part} \subseteq (C_{part} \times C_{part})$ est l'ensemble des relations associatives. Il permet de définir les types de relations reliant les concepts de l'ontologie dans $(C_{part} \times C_{part})$;

– $\mathcal{H}_{part}^{C}$ hiérarchie (taxonomie) de concepts : $\mathcal{H}_{part}^{C} \subseteq (C_{part} \times C_{part}) \cup (C_{part} \times C_{sub})$, $\mathcal{H}_{part}^{C}(C_1, C_2)$ signifie que C_1 est un sous-concept de C_2, pour les relations de subsomption entre les concepts de l'ontologie et éventuellement la subsomption d'un concept de l'ontologie à un concept synonyme dans l'ontologie de référence;

– $\sigma_{\mathcal{R}} : \mathcal{R}_{part} \rightarrow C_{part} \times C_{part}$ est la signature d'une relation associative ;

– $\sigma_{\mathcal{A}} : \mathcal{A}_{part} \rightarrow C_{part} \times \mathcal{T}_{part}$ est la signature d'une relation d'attribut, $\mathcal{T}_{part}$ est composé des types simples et d'une partie de l'ensemble des concepts subsumeurs C_{sub} pour les attributs ayant un co-domaine faisant référence à un concept de l'ontologie de référence.

Pour le niveau lexical des ontologies, un lexique est introduit. Il est le vocabulaire contrôlé du *partenaire* et contient l'ensemble de ses relations et labels de concepts. Pour une structure d'ontologie $\mathcal{S}_{part}$, un lexique $\mathcal{L}_{part}$ est défini comme le quadruplet :

$$\mathcal{L}_{part} := \{\mathcal{L}_{part}^{C}, \mathcal{L}_{part}^{\mathcal{R}}, \mathcal{F}_{part}, \mathcal{G}_{part}\} \text{ où :}$$

- $\mathcal{L}_{part}^{C}$ est l'ensemble dont les éléments constituent l'entrée lexicale des concepts ;
- $\mathcal{L}_{part}^{\mathcal{R}}$ est l'ensemble dont les éléments constituent l'entrée lexicale des relations ;
- $\mathcal{F}_{part} \subseteq \mathcal{L}_{part}^{C} \times C_{part}$ la relation de référence pour les concepts tels que :
  - $\forall\ l_c \in L_{part}^{C} : F_{part}(l_c) = \{c \in C_{part} / (l_c, C_{part}) \in F_{part}\}$
  - $\forall\ c \in C_{part}\ \ F_{part}^{-1}(c) = \{l_c \in L_{part}^{C} / (l_c, c) \in F_{part}\}$
- $\mathcal{G}_{part} \subseteq \mathcal{L}_{part}^{\mathcal{R}} \times \mathcal{R}_{part}$, la relation de référence pour les relations telles que :
  - $\forall\ l_r \in \mathcal{L}_{part}^{C} : \mathcal{G}_{part}(l_r) = \{r \in \mathcal{R}_{part} / (l_r, \mathcal{R}_{part}) \in \mathcal{G}_{part}\}$
  - $\forall\ r \in \mathcal{R}_{part}\ \ \mathcal{G}_{part}^{-1}(r) = \{l_r \in \mathcal{L}_{part}^{\mathcal{R}} / (l_r, r) \in \mathcal{G}_{part}\}$

4.3.3 Un modèle formel de base d'annotations

La mise à disposition seule d'une couche sémantique pour une source de données ne suffit pas. Il faut également mettre en relation la couche sémantique et les données. C'est dans ce cadre que la *base d'annotations* intervient en spécifiant les relations qui existent entre les éléments de la couche sémantique et les données. Dans notre contexte, le schéma global interne à une source de données est une ontologie. Nous avons choisi pour alléger le système de ne pas utiliser de schémas locaux, pour ainsi faire jouer pleinement son rôle de pont au schéma global entre les données existantes.

C'est là où la notion d'hyperdata citée dans l'introduction entre en jeu permettant implicitement de mettre en relation les données via le schéma global. Dans notre démarche, nous n'imposons pas le respect ou l'utilisation d'une grammaire déterminée, ce qui participe à la flexibilité et l'ouverture, sinon ce serait fort handicapant au vu de la masse importante de documents.

Nous utilisons une approche d'intégration en mode *GAV* (*Global As View*). Selon les principes de cette approche, le mapping $\mathcal{M}_{part}^{i}$ du système d'intégration est obtenu en associant à tout élément C_g du schéma global $\mathcal{G}$, donc à tout concept de l'ontologie *partenaire*, une requête q_S sur le schéma des sources $\mathcal{S}_{part}$.

Une *base d'annotations* d'un *partenaire* est alors modélisée par un ensemble de règles de mapping r_{part}^{i} , de la forme :

$$r_{part}^{i} : \{ c_{part}^{i} \in C_{part}^{i} \} \rightarrow \{ locUrl=axeXPathEntite_{part}^{k} \} \text{ où :}$$

— OS_{part}^{i} est un concept de l'ensemble C_{part}^{i} des concepts de l'ontologie médiateur ;

— $locUrl$ est l'url de base identifiant la source XML dans l'entrepôt associé à ce système d'intégration et $axeXPathEntite_{k}$ part est l'axe par XPATH, du chemin de localisation de l'entité k concerné dans le document XML.

4.3.4 Un modèle de base de connaissances

La *base de connaissances* est constituée de la *base d'annotations* ainsi que de l'*ontologie générique* et de celle spécifique au *partenaire*. Ainsi, une *base de connaissances* BC_{part}^{i} sera modélisée par le triplet :

$$BC_{part}^{i} := \{OG, OS_{part}^{i}, BA_{part}^{i}\} \text{ où :}$$

— OG est l'*ontologie générique* ayant dans sa taxonomie l'ensemble des concepts partagés par tous les *partenaires* et ayant chacun un subsumeur dans l'ontologie externe de référence ;

— OS_{part}^{i} est l'ontologie locale spécifique au *partenaire* ;

— BA_{part}^{i} part est la *base d'annotations* du *partenaire*.

4.3.5 Un modèle formel de *dataweb sémantique*

Le *dataweb sémantique* est constitué par l'*entrepôt de documents XML* auquel on ajoute la *base de connaissances* pour un apport sémantique. Il est donc modélisé par un couple constitué par le *dataweb* et la *base de connaissances partenaire*. Ainsi un *dataweb sémantique* est un couplet

$$DS_{part}^{i} := \{ D_{part}^{i}, BC_{part}^{i} \} \text{ où :}$$

— D_{part}^{i} est le *dataweb partenaire* ou *entrepôt de documents XML*;

— BC_{part}^{i} est la *base de connaissances* du *partenaire*.

4.3.6 Un modèle formel de système d'intégration par partenaire

Nous avons un modèle d'intégration interne à un *partenaire* assez particulier. D'habitude, dans les systèmes d'intégration, on établit les correspondances entre un schéma global fédérateur et les schémas locaux par un ensemble de règles de mapping. Cependant, dans notre approche, l'utilisation de XML et l'approche d'ingénierie ascendante des ontologies permettant d'établir directement les correspondances avec les parties XML décrites

directement par les concepts. Nous modélisons, de manière formelle, un système d'intégration de données *partenaire* I_{part}^i, par

$$I_{part}^i := \{OS_{part}^i, \mathcal{BA}_{part}^i\} \text{ où :}$$

— OS_{part}^i est le schéma global ou médiateur représentant le domaine d'intérêt du système d'intégration, appelée ici une ontologie. Elle est donc définie par le couplet $(S_{part}^i, \mathcal{L}_{part}^i)$ où $S_{part}^i := \{C_{part}^i, \mathcal{R}_{part}^i, \mathcal{A}_{part}^i, \mathcal{T}_{part}^i, C_{sub}^i, \mathcal{H}_{part}^C, \sigma_{\mathcal{R}}^i, \sigma_{\mathcal{A}}^i\}$ est la structure de l'ontologie et $\mathcal{L}_{part}^i$ son lexique. Il est exprimé avec des contraintes d'intégrité dans un langage $\mathcal{LG}_{part}^i$ sur un alphabet $\mathcal{AG}_{part}^i$ comprenant un symbole pour chaque élément de $\mathcal{G}_{part}^i$;

— $\mathcal{BA}_{part}^i$ est la *base d'annotations* établissant les correspondances entre chaque concept de l'ontologie *partenaire* OS_{part}^i et les nœuds décrits dans les documents des *dataweb* $\mathcal{D}_{part}^i$.

4.3.7 Un modèle formel de système d'intégration globale

Le système d'intégration globale à ce niveau sera formalisé par le triplet

$$I_{global} := \{OG_{global}, OS_{global}, \mathcal{M}_{global}\} \text{ où :}$$

– OG_{global} est le schéma global ou médiateur, nous utilisons ici une *ontologie globale* comme médiateur. Il exprimé avec des contraintes d'intégrité dans un langage $\mathcal{L}_{OGglobal}$ sur un alphabet $\mathcal{A}_{OGglobal}$ comprenant un symbole pour chaque élément de OG_{global};

– OS_{global} est l'ensemble des schémas des sources décrivant la structure des sources, au niveau structurel il s'agit de l'ontologie associée au « *dataweb* » *partenaire* avec la composante ontologique de son modèle d'intégration de données. Il est exprimé dans un langage d'ontologie $\mathcal{L}_{OSglobal}$ sur un alphabet $\mathcal{A}_{OSglobal}$ comprenant un symbole pour les éléments de chacune des ontologies *partenaires* ;

– $\mathcal{M}_{global}$ est le mapping établissant une relation entre l'*ontologie globale* jouant le rôle de schéma global et médiateur OG_{global} et OS_{global}.

4.4 Conclusion

L'objectif de ce travail est de permettre à plusieurs organismes et experts *partenaires* de partager leurs sources de données hétérogènes. Dans ce contexte, est proposée une approche d'intégration qui peut être résumée en trois phases [Sall et Lo, 2007] dont nous avons proposé les modèles dans cette partie.

Une première phase consiste à effectuer une intégration structurelle des données avec la réalisation, pour chaque *partenaire*, d'un *entrepôt de documents XML* (*dataweb*) intégrant les sources de données dudit *partenaire*.

Une deuxième phase permet l'intégration sémantique des données de chaque *dataweb* avec une *base de connaissances* construite à partir des données *partenaires* associées à chaque *dataweb*. Nous avons également présenté le modèle de la *base de connaissances* ainsi que de ses composantes.

Enfin, dans la troisième, le traitement du problème de l'intégration se résume à l'intégration des différents *dataweb sémantique*s en utilisant un système à base de hubs avec une *ontologie globale* construite de manière collaborative.

Dans ce chapitre, nous avons aussi présenté le modèle de l'*approche dataweb sémantique* essentiellement basé sur le modèle de *dataweb*. Ce dernier est constitué par un *dataweb* et une *base de connaissances*, dont les modèles ont été exposés. Le modèle de la *base de connaissances* est composé du modèle de l'ontologie *partenaire*, l'*ontologie générique* ainsi que celle d'annotations. Le modèle de l'ontologie est basé sur celle des ontologies à base lexicale.

Partant du vocabulaire contrôlé décrivant les données de chaque *partenaire* pour la construction de son ontologie, l'approche de modélisation à base lexicale s'est avéré la plus adaptée. De surcroît, cette phase de modélisation globale permet de formaliser les composants et leurs interactions facilitant ainsi le contrôle et l'implémentation des composants de leurs caractéristiques.

Cette approche d'intégration requiert deux préalables : (1) l'existence d'un *entrepôt de documents XML* pour chaque *partenaire* à partir duquel une ontologie sera générée (2) l'existence d'une ontologie du domaine servant de référence pour la construction des ontologies locales. Notre point de vue est que l'extraction d'une sémantique fiable à partir des données ne peut se faire sans l'existence d'une ontologie normalisant préalablement le domaine développé par les experts.

Dans le chapitre qui suit, nous nous intéresserons au processus de construction de chacune des composantes du système d'intégration.

Chapitre 5

Intégration sémantique et structurelle

Sommaire

5.1 Introduction 114

5.2 Nature et structure des données sources initiales 115

5.3 Construction d'un entrepôt pour chaque *partenaire* 116

5.3.1 Extraction et transformation des données sources 117

5.3.2 Restructuration et nettoyage des données sources 119

5.3.2.1 Restructuration ciblant les caractéristiques spatiales 123

5.3.2.2 Restructuration ciblant les caractéristiques temporelles 124

5.3.2.3 Restructuration ciblant les unités de mesure 125

5.4 Processus d'extraction des ontologies *partenaires* 126

5.4.1 Extraction et subsomption des concepts candidats 128

5.4.1.1 Extraction des concepts candidats 129

5.4.1.2 Subsomption des concepts candidats 132

5.4.2 Extraction et inférence des relations sémantiques 133

5.4.2.1 Extraction de relations à partir du *dataweb partenaire* 133

5.4.2.1.1 Les relations de subsomption 133

5.4.2.1.2 Les relations associatives 134

5.4.2.1.3 Les relations d'attribut 134

5.4.2.2 Extraction de relations sémantiques par inférence 135

5.5 Construction des bases d'annotations 136

5.6 Construction des ontologies génériques aux sources 137

5.7 Un système à base de hubs pour une médiation entre *partenaires* 138

5.7.1 Architecture du système 139

5.7.2 Adaptation du système par rapport à notre contexte 140

5.8 Construction de l'*ontologie globale* 141

5.9 Conclusion 142

5.1 Introduction

Ce chapitre présente le processus de construction d'un *dataweb sémantique* pour chaque organisme fournisseur de données en vue de l'homogénéisation structurelle et l'intégration sémantique chez tous les *partenaires*. Ce processus prépare les données intégrées *partenaires* à la phase d'intégration entre les sources.

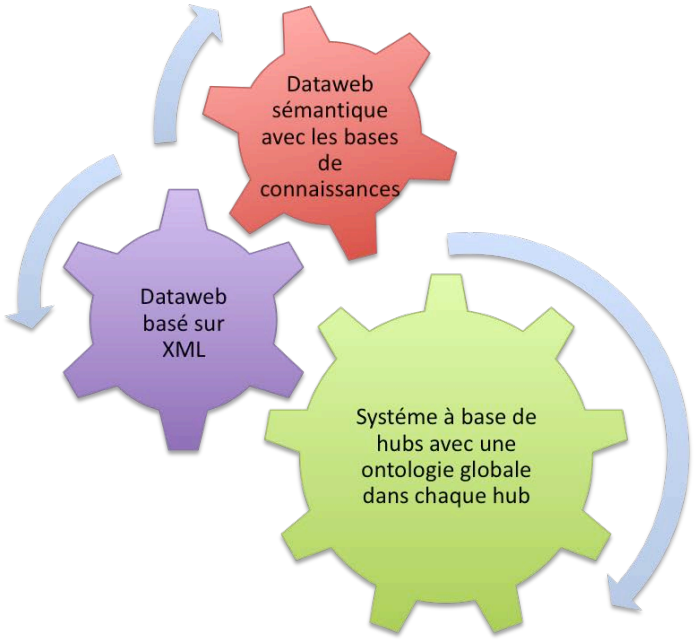


Figure 5-1 Trois composants de base du système d'intégration

La transformation des données sources initiales sous un format commun (pré-intégration) permet de transformer l'ensemble des données sources représentées avec des structures diverses en XML pour homogénéiser leur structuration. Comme le montre la figure 5.1, nous avons les trois composants modulaires qui forment l'*approche dataweb sémantique* permettant d'arriver à une intégration sémantique et structurelle des données.

La première section expose la nature des données cibles ainsi que leur structure générale.

La deuxième section présente les trois phases permettant la construction des *dataweb partenaires* que sont l'extraction, la transformation et la restructuration des données structurées XML en vue de leur chargement dans le *dataweb* structurellement homogène.

La troisième partie expose le processus d'extraction des concepts à partir des données sources des *dataweb* en vue de la construction d'une ontologie pour chaque *partenaire* enrichie par l'utilisation d'une *ontologie générique* par rapport à la tâche d'intégration de données environnementales.

Les dernières parties présentent les méthodologies utilisées pour la construction des liens entre les concepts extraits et les parties de document décrites ainsi que le système de médiation globale entre les *dataweb* basé sur la constitution d'une troisième ontologie dite globale à partir d'ontologies génériques des *partenaires*.

5.2 Nature et structure des données sources initiales

Une étude des tableaux de données issues des échantillons de données montre qu'ils sont de nature *«individus × variables »*. C'est-à-dire présentant l'évolution l'observation sur des propriétés communes à des individus, comme le montre la structure générale présentée par la table 5.1.

C	P_1			…			P_i			…			P_k		
	a^{11}	…	$a^{1\alpha}$	…	…	…	a^{i1}	…	$a^{i\beta}$	…	…	…	a^{k1}	…	$a^{k\gamma}$
I_1	$v^{11}{}_1$	…	$v^{1\alpha}{}_1$	…	…	…	$v^{i1}{}_1$	…	$v^{i\beta}{}_1$	…	…	…	$v^{k1}{}_1$	…	$v^{k\gamma}{}_1$
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
I_n	$v^{11}{}_n$	…	$v^{1\alpha}{}_n$	…	…	…	$v^{i1}{}_n$	…	$v^{i\beta}{}_n$	…	…	…	$v^{k1}{}_n$	…	$v^{k\gamma}{}_n$

Tableau 5-1 Structure d'un tableau « individus * variables » général

Dans cette structure, les I_n sont les individus dont le type ou la nature est représenté par la variable dite privilégiée(en langage statistique) C, et ici on s'intéresse aux valeurs v_n^{β} sur les attributs a^{β} des propriétés $\mathcal{P}_i$. La structure générale d'un document XML issue de l'extraction des données sous format XML est illustrée par la figure 5.2.

Les données des sources *partenaires* sont à l'origine représentées sous plusieurs formats différents. On retrouve aussi bien des formes de données structurées que semi-structurées. Dans nos travaux, nous nous intéressons aux données présentées sous forme de table de données. Ce sont des tableaux montrant pour la plupart l'évolution temporelle de certaines caractéristiques d'un des éléments faisant l'objet d'étude. Donc, comme spécifié dans [Lo, 2002], il y a une acquisition et conservation de données ou de documents.

Cependant cette information n'est pas mise à jour mais archivée et ensuite restituée à l'utilisateur final. C'est l'aspect conservatoire des systèmes d'information s'intéressant à des données environnementales (systèmes d'informations environnementaux [Dzeakou et. al, 1998]) avec en plus un aspect observatoire exprimant le fait que les usagers de tels systèmes peuvent effectuer des observations (mesures, synthèse, recherche, extraction, etc.) [Dzeakou et Derniame, 1998].

Dans le langage des tableaux de données, ces observations portent sur des individus en ciblant certaines de leurs propriétés ou attributs.

Comme souligné dans la présentation de l'article [Lechevallier, 2005][38], un tableau de données est une représentation matricielle. Il montre la mise en correspondance ou la mise en relation d'un ensemble I d'individus avec l'ensemble $\mathcal{P}$ des variables.

$$\mathcal{T}: I \chi \mathcal{P} \rightarrow \mathcal{V}$$

L'ensemble des individus constitue l'unité de base sur laquelle les mesures sont réalisées selon un ensemble de variables permettant de les décrire. Ainsi, chaque individu I_i, aura une valeur $\mathcal{V}_i$ qui lui sera affectée pour une description $\mathcal{P}_i$. Les variables, selon les tableaux, indiquent des propriétés aussi bien qualitatives que quantitatives des individus.

Du point de vue de la sémantique des tableaux, nous avons dans les échantillons des tableaux pouvant être de type hétérogènes. C'est autant dire que les variables peuvent être de nature différente avec une mise en correspondance de type « individus _ variables »[39] comme le montre le tableau 5.1.

Selon cette sémantique, il se pose deux cas possibles. Dans le premier cas, l'ensemble des individus est un singleton. Dans le second, nous avons un ensemble d'individus non singleton complètement disjoints de l'ensemble des variables ; comme dans les tableaux de proximité, est, l'objet des relevés.

5.3 Construction d'un entrepôt pour chaque partenaire

La construction d'un *dataweb partenaire* est effectuée en deux phases. Une première phase permet par la pré-intégration de transformer l'ensemble des données sources initiales sous format XML et une deuxième phase rend possible la restructuration des données. Le rôle de la restructuration consiste dans notre contexte à réorganiser et nettoyer chaque document XML issu de la transformation par une normalisation et une extraction des propriétés spatio-temporelles.

[38] http://www.infres.enst.fr/rdc05/resumes.html
[39] http://www.tn.refer.org/hebergement/analyse/chap1_3.html

1. **Algorithme** : Génération d'un document XML à partir d'une forme tabulaire
   Données : Une forme tabulaire de la forme du tableau 5.1
   Résultat : Un document XML
2. **pour** une forme tabulaire **faire**
3. Créer un document XML;
4. Insérer un nœud racine $\mathcal{N}_{\mathcal{R}}$ ayant comme nom le titre du tableau;
5. Insérer un nœud fils à $\mathcal{N}_{\mathcal{R}}$ ayant pour label C
6. **pour tous les** individus I_n **faire**
7. **pour** chaque propriété $\mathcal{P}_i$ **faire**
8. Créer un nœud ayant pour label $\mathcal{P}_i$;
9. **pour tous les** attributs a_{ij} de $\mathcal{P}_i$ **faire**
10. Insérer à $\mathcal{P}_i$ un attribut a_{ij} ayant pour valeur value v_n^{ij};
11. Insérer $\mathcal{P}_i$ comme un nœud fils de I_n;
12. Insérer I_n comme un nœud fils de C ;

Algorithme 1 : Génération d'un document XML à partir d'une forme tabulaire

5.3.1 Extraction et transformation des données sources

L'extraction et la transformation des données sources visent à extraire les données de leur forme tabulaire initiale pour les transformer en XML. L'une des tâches les plus importantes dans ce processus est de pouvoir identifier les nœuds et les attributs à partir de la structure tabulaire. Nous avons développé un algorithme pour cela.

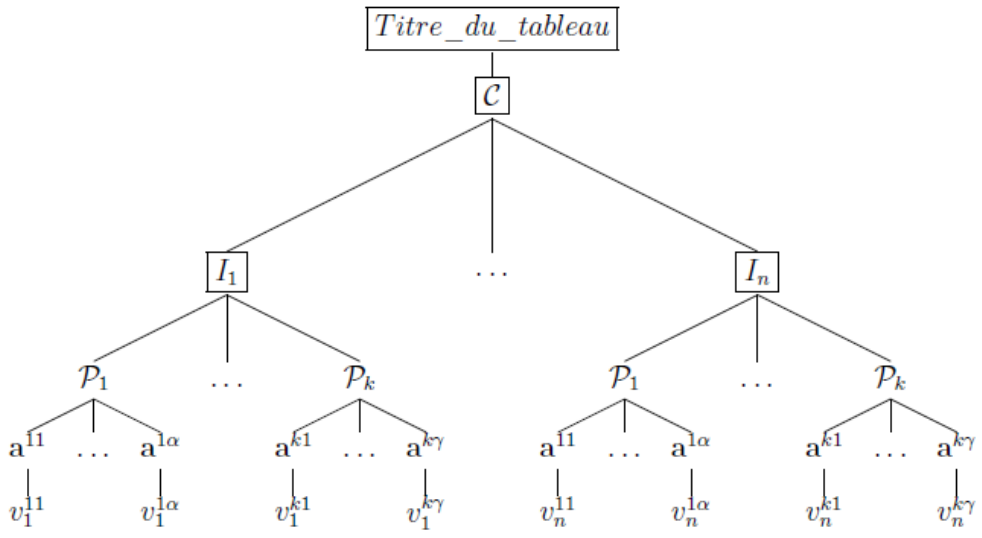


Figure 5-2 Structure XML résultant de la transformation du tableau 5.1

```
<?xml version="1.0" encoding="UTF-8" ?>
<Titre_du_tableau_de_donnees>
        < C>
                <I1>
                        <P1>
                                <a11>v112 </a11>
                                        ...
                                <a1_>v1_2 </a1_>
                        </P1>
                                ...
                        <Pi>
                                <ai1>vi12 </ai1>
                                        ...
                                <ai_>vi_2 </ai_>
                        </Pi>
                                ...
                        <Pk>
                                <ak1>vk12 </ak1>
                                        ...
                                <ak>vk2 </ak>
                        </Pk>
                </I1>
                ...
                <In>
                        <P1>
                                <a11>v11n </a11>
                                        ...
                                <a1_>v1_n </a1_>
                        </P1>
                        ...
                        <Pi>
                                <ai1>vi1n </ai1>
                                        ...
                                <ai_>vi_n </ai_>
                        </Pi>
                                ...
                        <Pk>
                                <ak1>vk1n </ak1>
                                        ...
                                <ak>vkn </ak>
                        </Pk>
                </In>
        </C>
<T itre_du_tableau_de_donnees>
```

Figure 5-3 Document XML résultant de la transformation du tableau 5.1

Les données considérées sont généralement stockées sous forme de table de données résultant de l'évolution des observations sur le terrain des caractéristiques d'une population d'individus dont les propriétés en commun sont les objets cibles d'une étude.

Les tableaux précédents ont été valorisés par des experts de leur domaine. Donc les noms de colonnes et titre de chaque tableau véhiculent des termes du vocabulaire *partenaire*, propre au domaine et de manière implicite les relations entre ces dernières. XML convient de façon particulière à la représentation de ces données où chaque case sera transformée en nœud du document et chaque ligne du tableau un sous-arbre à part entière comme le montre la figure 5.2.

Les formes tabulaires de ce genre seront transformées automatiquement comme montré dans [Sall et al., 2009] autour de leur titre qui devient le nœud racine du document structuré en XML résultant de leur transformation. Du point de vue lexical, ces documents XML véhiculent l'idée de vocabulaire contrôlé ou d'abstraction sur le vocabulaire et de hiérarchisation avec les relations de compositions entre les nœuds.

Cette relation de composition qui relève de la subsomption automatiquement extractible découle de la forme tabulaire initiale des données. Sur une même colonne, les descriptions dans chaque ligne représentent une particularité dont la nature relève de celle décrite dans sa première ligne. Par conséquent dans le cas du tableau général 5.1 la variable privilégiée est C, et les individus I ne constituent que des constructions ou particularités de C. L'algorithme 1 détaille les instructions permettant de réaliser la transformation sous le format XML.

5.3.2 Restructuration et nettoyage des données sources

Dans cette phase, l'objectif visé est de normaliser les documents des *dataweb*. Ainsi, après étude des données, l'expert devra fournir les caractéristiques à restructurer. Cette restructuration peut prendre plusieurs natures différentes dont les stratégies de résolution que nous avons mise en place sont l'objet de cette section.

La restructuration peut porter sur le nettoyage des noms de labels afin d'en éliminer des occurrences de caractères non nécessaires ou ne pouvant pas passer le validateur XML. De manière générale, on cherche à corriger les occurrences de certains labels dont des acronymes sont à corriger dans les documents XML. A l'image du processus de nettoyage des données en datamining, l'opération de nettoyage a principalement pour finalité de procéder à l'imputation des données manquantes, le filtrage du bruit dans les données et de corriger les incohérences [Han et Kamber, 2001]. Dans nos travaux, nous n'avons pas cherché de

solutions par rapport à cette problématique des données manquantes, la démarche et l'approche d'intégration pouvant s'en abstraire.

Les données de nature environnementales ont la propriété d'être toujours relatives à la région où elles ont été collectées, le moment et souvent portent une mesure normalisant les quantités observées. Ces trois caractéristiques imposent le besoin de rechercher les occurrences temporelles et spatiales dans les connaissances décrivant les données.

1. **Algorithme :** Restructuration d'un document XML par identification des propriétés spatiales et temporelles
   Données : *Une forme tabulaire de la forme du tableau 5.1*
   Résultat : *Un document XML*
2. **pour** *un fichier structuré XML* **faire**
3. Normaliser C;
4. Traiter les caractéristiques spatio-temporelles de C;
5. Parsing du fichier XML;
6. **pour tous** *les individus* I_n **faire**
7. Normaliser I_n;
8. Traiter les caractéristiques spatio-temporelles de I_n;
9. **pour toutes** *les propriétés* $\mathcal{P}_i$ *associées à* I_n **faire**
10. Normaliser $\mathcal{P}_i$;
11. Traiter les caractéristiques spatio-temporelles de $\mathcal{P}_i$;
12. **pour tous** *les composants* a^{ij} de $\mathcal{P}_i$ faire
13. Normaliser a^{ij};

Algorithme 2 : Réorganisation d'un document XML issu d'une forme tabulaire

Dans les tableaux de données, le plus souvent les unités temporelles et spatiales sont exprimées sous forme de phrases et non considérés comme des colonnes à part. C'est le cas, dans notre contexte d'application. Pour traiter ces deux étapes, il est donc nécessaire de disposer d'un dictionnaire des unités spatiales, relevant l'ensemble des occurrences de nom de localité possibles. Une autre alternative est le traitement manuel avant la phase de transformation des documents en XML.

Pour les unités temporelles précisant à quelle période les données ont été collectées, nous avons des occurrences possibles par exemple sous forme de saison agricole dans le contexte subsaharien de l'application ou comme une simple date. L'observation et l'étude des

données peuvent permettre de déterminer une règle générale de spécification de ces types d'occurrences. Dans le cadre d'application, nous avons pu dégager une heuristique pour extraire ces caractéristiques temporelles.

Pour les unités de mesures, il suffira de constituer comme pour les noms de localités un dictionnaire des formes d'occurrence des différentes unités de mesure. Elles seront ainsi extraites, des noms de labels et mises en exergue comme montré dans la partie sur la restructuration ciblant les unités de mesure.

L'algorithme 2 montre les grandes lignes permettant la restructuration et le nettoyage. Un parcours de l'ensemble des nœuds du document XML, normalise le document avec un nettoyage avec l'étape de normalisation, puis une restructuration des occurrences spatio-temporelles est effectuée.

Soit l'exemple du tableau 5.2 présentant plusieurs formes d'hétérogénéité que nous allons aborder. Nous allons nous baser dessus pour illustrer les exemples des différentes restructurations.

Régions	Propriété(an1/an2)			
	attribut1(ha)	attribut2(ha)	…	attributn(ha)
Label_Spatial1	val1	val2	…	valn
…	…	…	…	…
…	…	…	…	…

Tableau 5-2 Exemple de tableau illustratif pour les restructurations

D'un point de vue applicatif certains tableaux contiennent des occurrences de caractères inutiles ou comme « % » qui ne peuvent passer le validateur XML. Nous identifions également dans ce processus l'occurrence des unités de mesure sous différentes formes comme des unités de mesure de poids ou de surface comme l'hectare(ha) que nous prendrons comme exemple pour illustrer la restructuration suivant les unités de mesure.

La deuxième étape est celle de l'extraction et de l'identification des caractéristiques spatio-temporelles dans les documents XML pour les mettre en exergue. Nous procédons à une simple recherche sous forme de « matching » dans les noms de balise pour rechercher l'occurrence de certains mots avant d'insérer un nouveau nœud fils pour celui dans lequel s'est faite l'extraction.

Pour les caractéristiques temporelles, ce sont des périodes dans le contexte sahélien qui s'étalent sur deux ou plusieurs années sous forme de saisons agricoles. Pour résoudre la problématique de leur extraction, notre approche consiste à étudier les données afin de

chercher une règle d'écriture de ces périodes. Par exemple, dans notre contexte applicatif, nous avons étudié les données pour pouvoir mettre en place une heuristique permettant d'identifier ces caractéristiques temporelles dans les noms de balises XML des données. Il apparaît en effet que les saisons sont représentées par les *partenaires* dans les échantillons de données sous la forme *« debut » _ « fin »* soit la date début de la saison et la date de fin séparées par un « tiret bas ». Pour chaque document XML, ce traitement est effectué sur tous ses composants en commençant par le titre du tableau.

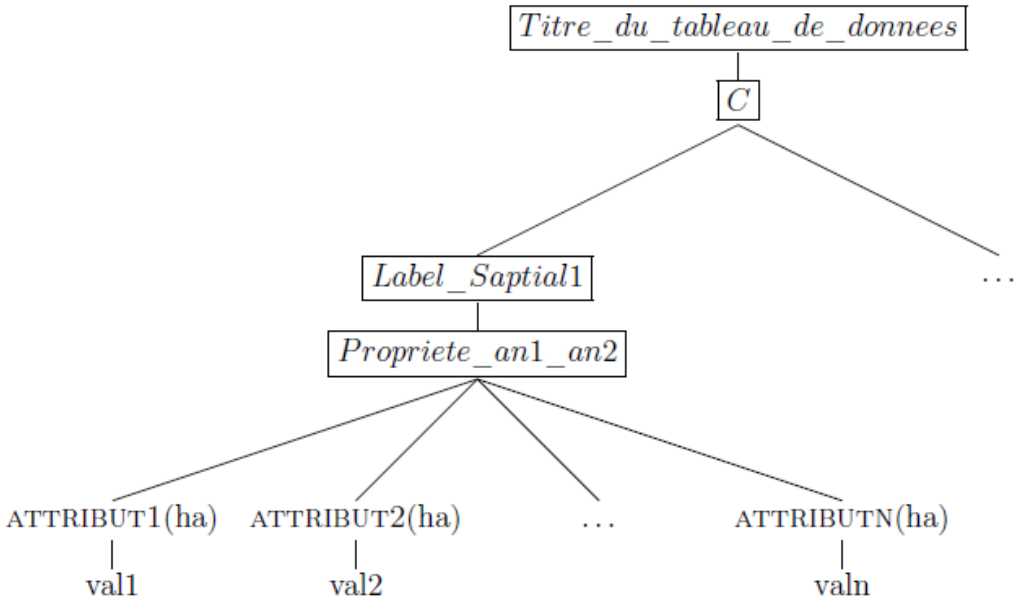


Figure 5-4 Une vue sur la structure XML résultant de l'XMLisation du tableau 5.2

Ainsi donc, chaque document issu de la pré-intégration est passé dans le module de réorganisation. En sortie nous avons un document qui répond aux normes nécessaires pour faciliter une interrogation intégrant les dimensions spatiales et temporelles. L'algorithme 2 montre les grandes lignes de ce processus de réorganisation.

En considérant que la normalisation consiste à reformater les a^{ik}, $\mathcal{P}_j$ et I_i selon les critères définis par l'expert du domaine, la restructuration est constituée par un parsing des fichiers XML, la normalisation des éléments XML ainsi que l'identification et la mise en évidence des caractéristiques spatio-temporelles. Nous allons discuter des cas de chacun des ces composantes.

Soit l'exemple du tableau 5.2 dont une partie de l'arbre XML est représenté ci-dessous. Nous avons dans le tableau plusieurs unités spatiales que sont *« Label_Saptial1 »,...* et une unité

temporelle qu'est la saison *« an1_an2 »* est extraite de la propriété *« Propriété(an1/an2) »*. Nous avons également l'unité de mesure de surface « (ha) » ou « hectare».

5.3.2.1 Restructuration ciblant les caractéristiques spatiales

Dans le cas d'une restructuration suivant les caractéristiques spatiales, si le label correspond exactement à la caractéristique recherchée comme c'est le cas d'une des localités énumérées par l'expert dans le tableau 5.2, alors le nœud sera complètement restructuré. Il sera remplacé par un nœud ayant comme label *« localite »* avec un attribut « nom » ayant comme valeur le nom de la localité trouvée. Dans le cas où la caractéristique trouvée est incluse dans son label, alors un attribut *« localite »* avec comme valeur l'occurrence spatiale lui est ajouté.

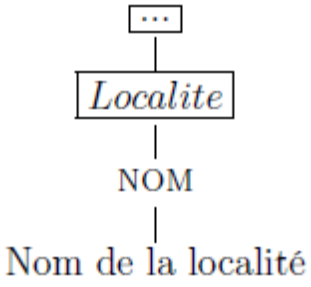


Figure 5-5 Forme générale du nœud inséré pour la restructuration spatiale

L'expert du domaine devra fournir une liste complète des entités nommées servant de support pour rechercher les occurrences afin de les mettre en exergue. Cette recherche des caractéristiques spatiales est importante dans le contexte des données environnementales. Les requêtes sur les données de cette nature en vue de leur combinaison doivent prendre en compte leur aspect localisation lors de leur collecte.

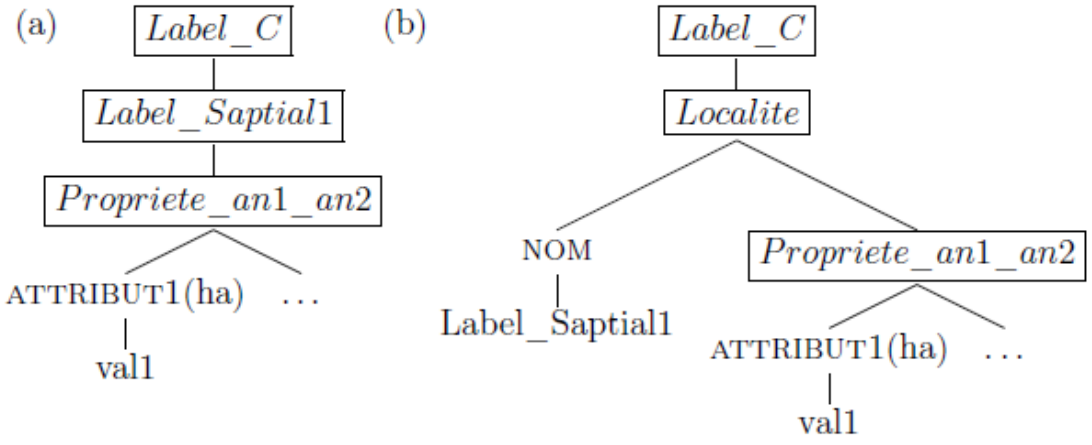


Figure 5-6 Exemple de restructuration ciblant les caractéristiques spatiales

La figure 5.5 donne la forme générale de la restructuration que nous proposons. C'est le cas du nœud « Label_Saptial1 » dans la figure 5.6. Il est restructuré et devient la valeur de l'attribut nom d'un nouveau nœud « localite ». L'item (a) montre le nœud original et le (b) le nouveau nœud obtenu après restructuration spatiale.

5.3.2.2 Restructuration ciblant les caractéristiques temporelles

Dans le cas où la restructuration porte sur les attributs temporels, il s'agit alors d'identifier les saisons dans les labels. Dans cette phase, nous proposons d'identifier les occurrences temporelles indiquant le moment dans lequel les données ont été recueillies sur le terrain. Lorsqu'il faut répondre à une requête avec une combinaison et inférence au niveau sémantique, il convient de ne prendre en compte que les données ayant le même contexte temporel. Dans notre cas nous avons étudié les données afin d'identifier un modèle d'occurrence des caractéristiques temporelles.

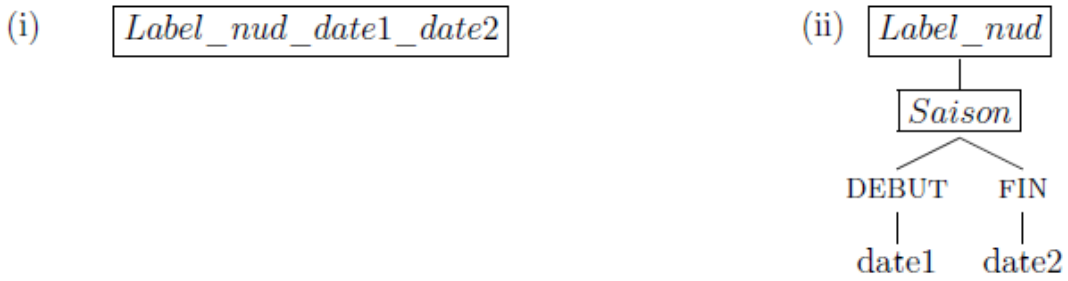


Figure 5-7 Forme générale de restructuration ciblant les caractéristiques temporelles

Selon la stratégie définie par l'expert étudiant les données, les données de nature environnementales sont généralement exprimées sous forme de saison. Pour cette raison, nous proposons de restructurer le nœud comportant cette caractéristique pour en extraire la valeur et insérer un nouveau nœud nommé saison avec comme attributs les années de début et de fin de la saison à laquelle les données ont été collectées.

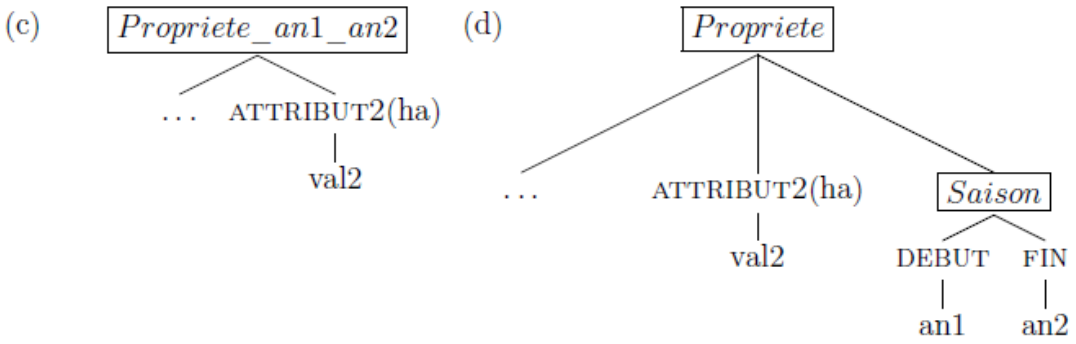


Figure 5-8 Exemple de restructuration ciblant les caractéristiques temporelles

Il convient d'attribuer au nœud cible un nouveau fils *saison_agricole* avec un attribut *debut* et un attribut *fin* qui auront comme valeur la date de début de la saison et la date de fin.

Prenons l'exemple *de Propriete_an1_an2*, la période *_an1_an2* sera extraite pour constituer le nouveau nœud fils et l'ancien nœud va porter comme label restructuré la chaîne de caractères *Propriete*. Dans la figure 5.8 l'item (a) montre le nœud original et le (b) le nouveau nœud obtenu après restructuration spatiale.

5.3.2.3 Restructuration ciblant les unités de mesure

Le troisième cas est celui de l'extraction des caractéristiques unitaires, i.e les unités de mesures. Elles doivent être mises en exergue après avoir identifié une unité de mesure. L'unité de mesure est extraite du label du nœud et deux attributs *valeur* et *unité* lui seront rajoutés.

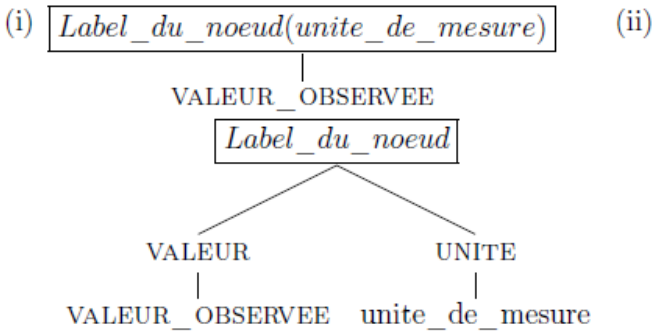


Figure 5-9 Forme générale de restructuration ciblant les unités de mesure

Prenons l'exemple du nœud *« attribut1(ha) »* et de *« attribut2(ha) »* où l'on peut détecter l'occurrence de l'unité de mesure de surface « hectare » sous la forme *« (ha) »*.

L'occurrence identifiée de l'unité de mesure des superficies infestées permet de restructurer les nœuds comme *attribut1(ha)* pour lui rajouter deux attributs supplémentaires que sont l'unité de mesure en hectare et la valeur du relevé comme illustrés par la figure 5.10.

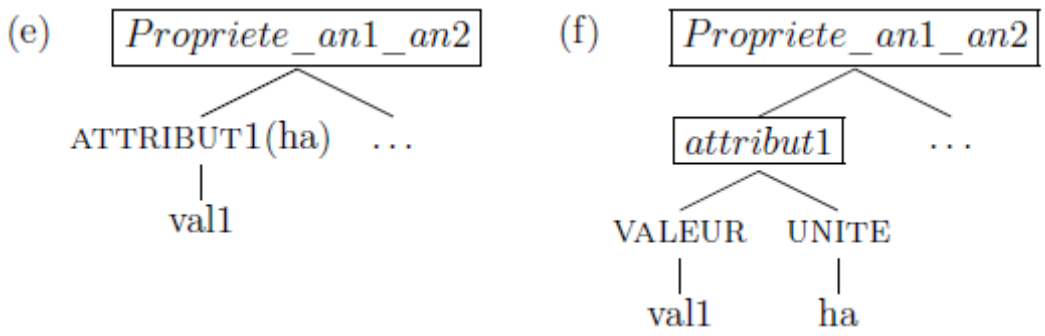


Figure 5-10 Exemple de restructuration ciblant les unités de mesure

L'identification de *Label_Saptial1* comme unité spatiale permet de restructurer le nœud initial.

Globalement, à partir de la structure arborescente fournie en début de cette section, la forme suivante, grâce aux informations fournies par l'expert et aux heuristiques mises en place.

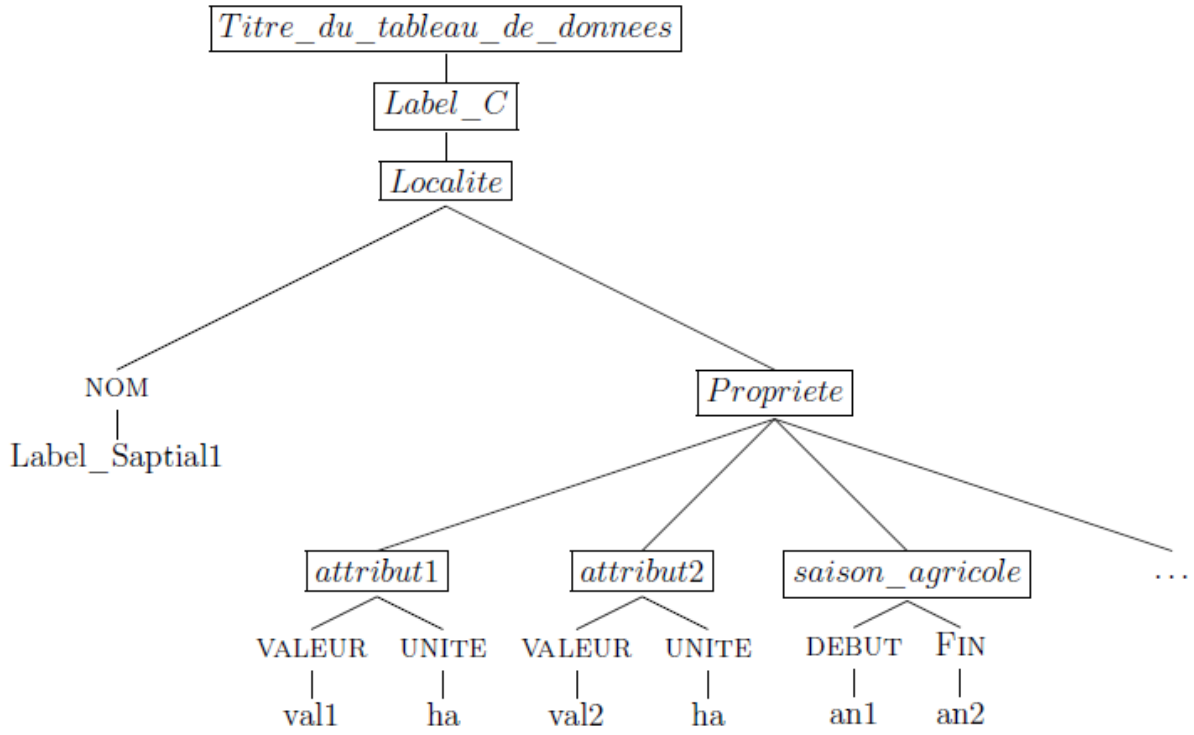


Figure 5-11 Une vue de la structure résultant de la restructuration de la figure 5.4

5.4 Processus d'extraction des ontologies partenaires

Afin de permettre l'intégration sémantique des différents *dataweb partenaires*, nous associons une ontologie OWL à chaque *partenaire*. Nous nous intéressons ici à la construction de cette ontologie.

Nous proposons une méthodologie de construction semi-automatique d'ontologie OWL à partir d'un *entrepôt de documents XML* par la réutilisation d'une ontologie existante avec des concepts plus générique du domaine d'application. Contrairement aux méthodologies qui souvent opèrent un passage vers XML-Schema pour la construction d'ontologies, nous mettons en œuvre une approche permettant d'extraire automatiquement chaque Figure concept candidat, ses attributs et composants ainsi que les relations de cardinalité entre les concepts et leurs attributs.

La figure 5.12 montre un extrait d'une ontologie *partenaire*. Cette ontologie laisse apparaître trois concepts *partenaires* dont deux trouvent des subsumeurs dans l'ontologie *AOS*

de référence. La réutilisation d'une ontologie plus générique couvrant le domaine comme *AOS* de la FAO dans le contexte des données environnementales nous permet de rajouter des relations sémantiques telles que la subsomption. Une section est spécialement réservée à l'ontologie *AOS*, pour le moment nous le prenons comme un représentatif des ontologies génériques standardisant un domaine bien défini.

```
<rdf :RDF (...) >
        (...)
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_x">
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_a"/>
</owl :Class>
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_y">
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_b"/>
</owl :Class>
        (...)
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_n">
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_a"/>
        (...)
</owl :Class>
</rdf :RDF>
```

Figure 5-12 Exemple d'extrait d'un format ontologie OWL d'un partenaire

Nous considérons qu'un document XML capitalisant les connaissances d'un domaine contient un ensemble de micro-contextes qui seront exprimés par les relations hiérarchiques et sémantiques entre les nœuds non-terminaux et leurs composantes. Cet ensemble de micro-contextes forme le contexte global du document. D'ailleurs, un document XML permet pour une restriction sur un vocabulaire donné de fournir une représentation hiérarchisée sur ces connaissances. Pour leur donner un sens, notre démarche consiste à rechercher et à introduire les relations d'ordre, de dépendance sémantique sur cette hiérarchie. Il existe cependant deux approches pour exprimer le sens qui, dans ce contexte, se confond par la notion de vecteur sémantique (étant donné que nos concepts sont formés d'un ensemble de termes hiérarchisés et ont des relations avec d'autres composantes de par l'architecture du document XML).

Dans une première approche, le vecteur pointe vers un objet du monde réel et dans ce cas le sens du vecteur sémantique devient une bijection entre les deux espaces : l'espace linguistique des mots et l'espace du monde réel. Ce phénomène est appelé dans la communauté linguistique la référence d'un mot. Ce type d'expression du sens ne nous convient pas pour la simple raison qu'il n'existe pas d'outil d'expression sémantique exprimant cette bijection avec l'espace du monde réel. Cependant, nous allons réutiliser cette

notion de référence non pas par une bijection vers le monde réel mais à un concept subsumeur dans une ontologie. Bien sûr, la bijection va d'une partie de l'ensemble des concepts de l'ontologie subsumeur de référence aux concepts candidats avec comme unique critère l'homonymie aux sens morphosyntaxiques.

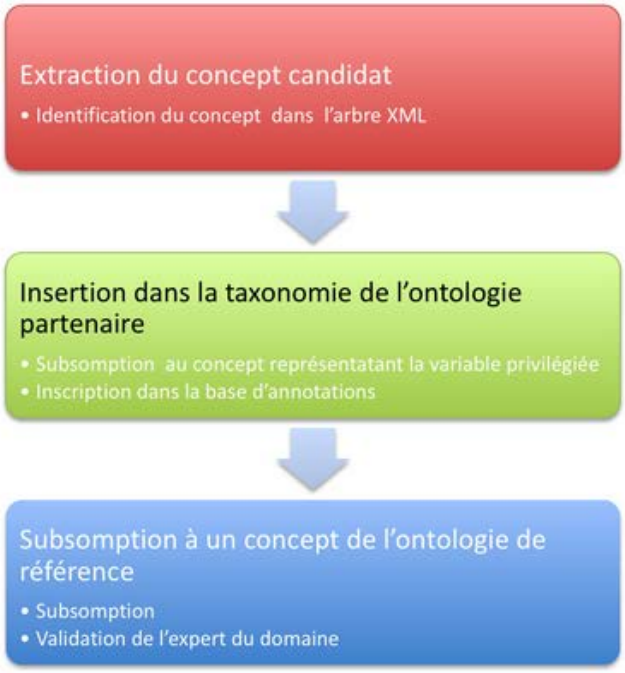


Figure 5-13 Illustration de l'extraction d'un concept à la recherche de relations

Dans la deuxième approche que nous avons adoptée, le sens d'un concept est exprimé à partir du contexte entourant le mot. Ce contexte se modélise dans les documents XML par les nœuds entourant dans un schéma le nœud concerné ainsi que leurs relations sémantiques.

Nous y introduisons en plus les termes qui le composent ainsi que la référence à son concept synonyme dans une ontologie existante. Dans les outils de traitement automatique de textes, elle est utilisée sous la forme de « fenêtre ». Ici notre fenêtre se résume aux composantes du nœud qui constituent notre micro contexte et son voisinage. En résumé, un concept candidat est un élément multidimensionnel d'un ensemble d'arrivée qui est l'ensemble des concepts candidats. Cet ensemble réalise une bijection avec la partie de l'ensemble des concepts subsumeurs dans l'ontologie de référence.

5.4.1 Extraction et subsomption des concepts candidats

Cette phase est celle de la construction du vocabulaire contrôlé de chaque *partenaire*, i.e. le niveau lexical des ontologies. Elle est basée sur l'extraction et la subsomption des concepts candidats à l'ontologie externe de référence.

Pour atteindre le niveau sémantique requis pour une ontologie, il faut rajouter à la couche sémantique des relations de nature sémantique. C'est alors dans cette phase que nous servons de la subsomption. Enfin une restructuration des ontologies est effectuée. Les deux phases précédentes sont semi-automatiques requérant l'assistance de l'expert du domaine.

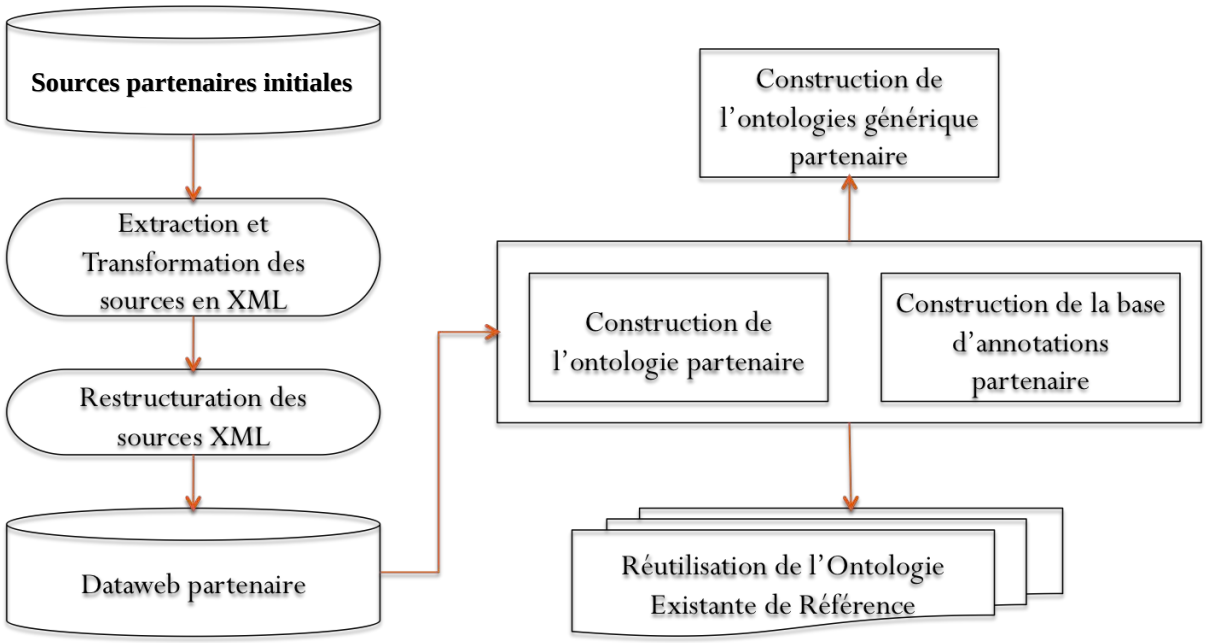


Figure 5-14 Processus de construction des ontologies

Dans les développements qui suivent, il est question de l'extraction des concepts candidats, des relations hiérarchiques et la subsomption pour la construction de l'ontologie du *partenaire*.

5.4.1.1 Extraction des concepts candidats

Nous avons adopté une stratégie de mapping direct de XML à OWL [Sall et Lo, 2007] en définissant le schéma de mapping suivant consistant à considérer :

— comme concept candidat ou classe OWL tout nœud de l'arbre d'un document XML non terminal et les nœuds terminaux ayant au moins un attribut ;

— comme attribut ou propriété OWL du concept candidat qui est leur nœud père dans l'arbre DOM, tout nœud terminal sans attribut ainsi que les attributs de nœuds. Ce seront donc les labels ou termes du concept candidat. Dans le cas de l'extraction de ces termes d'un concept, il est possible dans un document XML et dans le même niveau de l'arbre DOM, de rencontrer le même nœud avec des attributs différents. Dans ce cas, nous construisons le concept candidat comme ayant pour labels l'union des occurrences des attributs et nœuds de ses occurrences.

Cependant, après le développement du premier prototype, une problématique quasi générale s'est posée concernant une redondance sémantique entre des concepts faisant référence à des parties de documents XML de même niveau, ou d'un groupe de même niveau avec un de niveau immédiatement supérieur. Nous avons alors proposé une phase de restructuration des ontologies qui a confirmé l'étendue de cette problématique. Par conséquent une solution consistait à revenir un peu en amont pour en chercher les causes. La stratégie de mapping d'XML à OWL, elle, ne pose pas de problème. En effet l'information qu'ils rendent persistante sur la structure des documents XML n'est pas nécessaire et est de surcroît déductible en phase de construction. Si dans l'étape de construction de la taxonomie de l'ontologie à partir des documents XML, le problème ne se posait pas, la cause doit être liée à l'origine tabulaire. En cherchant à camper les données des *partenaires* sur une structure permettant de les modéliser pour faire leur relation avec les éléments XML, nous avons eu une confirmation de ces faits.

Reprenons l'exemple du tableau 5.2, déjà examiné, les tableaux de données sur lesquels nous travaillons sont de type *« individus * valeurs »* avec une certaine particularité liée au fait qu'ils véhiculent l'expression de mesures de certaines propriétés portant sur des individus. Comme nous l'avons déjà spécifié, C qui constitue le nœud venant hiérarchiquement après le nœud racine, agit ici comme une classe dans le concept objet. Il permet de déterminer la nature des autres individus qui sont les objets que l'on ne peut constater qu'en représentation sous format XML comme le montre la structure générale présentée dans le tableau 5.1. Les autres individus ne reprennent en réalité dans leur structure que le format structurel de cet individu. Il fournit la signification, la construction des autres individus grâce aux termes ou valeurs qui le décrivent. C'est la définition de ce qu'est un concept d'une ontologie.

Nous considérons ici une des définitions rapportées par N. Hernandez selon laquelle un concept se définit par Bachimont à trois niveaux [Bachimont, 2004]. Un concept est une signification. Sa place dans un système de significations permet de le comprendre, de le distinguer et de le différencier par rapport à d'autres concepts. Un concept est une construction. Comprendre un concept revient à construire l'objet dont il est le concept. Un concept est une prescription. On le comprend en exécutant l'action qu'il entreprend.

Ainsi nous distinguons d'un point de vue théorique trois ensembles de la structure du tableau 5.1 un tableau de n individus et k valeurs :

— Un singleton constitué par ce que nous nommons la variable privilégiée C, qui est la classe qui exprime la structure de l'ensemble des éléments constituant le second ensemble;

— le second ensemble la structure de l'ensemble des instances constitué par les n individus;

— le troisième ensemble est celui des k propriétés $\mathcal{P}_j$ sur lesquelles portent les mesures et qui constituent les descripteurs ou termes dudit concept.

Une alternative pour ne pas glisser vers le niveau des bases de connaissances est d'intégrer ces composants à la structure d'un concept. Soit C un concept définissant la variable privilégiée C, alors C est un triplet :

$$C := \{nom, \{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}, \{\mathcal{P}_j,\ 1 \leq j \leq k \in \mathcal{N}^*\}\} \text{ où :}$$

– *nom*, est la chaîne de caractère permettant de spécifier le nom du concept ;

– $\{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}$ est l'ensemble des instances ou objets du concept C, on dira que C subsume chaque élément de l'ensemble $\{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}$. En effet dans un tel cadre, C étant la variable privilégiée, lorsque du point de vue sémantique un observateur humain interprète ce tableau dans un premier temps on ne s'intéresse plus aux valeurs observées en tant que tels. Mais à la représentation et structure de la variable privilégiée et les individus deviennent des concepts construits à partir d'elle ;

– $\{\mathcal{P}_j,\ 1 \leq j \leq k \in \mathcal{N}^*\}$ constitue l'ensemble ordonné et structuré suivant les indices spécifiant les attributs(termes ou descripteurs) du concept C. Cet ensemble est constitué par l'union des propriétés directes des éléments ou attributs communs à l'ensemble des éléments de $\{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}$ que nous remontons à leur subsumeur. Chaque propriété $\mathcal{P}_j$ sera définie du point de vue de structurelle selon le couplet :

$$\mathcal{P}_j := \{nom, \{a_{jk},\ 1 \leq k \leq \alpha \in \mathcal{N}^*\}\} \text{ où :}$$

– *nom*, est la chaîne de caractère permettant de de spécifier le nom du concept ;

– $\{a_{jk},\ 1 \leq k \leq \alpha \in \mathcal{N}^*\}$, constitue l'ensemble ordonné et structuré suivant les indices spécifiant les composantes de la propriété $\mathcal{P}j$, dans tous les cas, nous nous occupons de bien récupérer le type de données le type $\mathcal{T}_j$ associé à a^{jk}.

Les individus $\{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}$ contribuent tout aussi à la conceptualisation du domaine des *partenaires* et expriment aussi un plus de la sémantique. Il s'agit d'une extension du concept. Nous avons décidé de les considérer comme des concepts qui n'intègrent pas dans leur structure explicite les termes de la variable privilégiée C mais les héritent. Cet individu va subsumer tous les éléments de $\{I_i,\ 1 \leq i \leq n \in \mathcal{N}^*\}$. Il est plus général que ces derniers et évidemment chaque concept subsumé hérite de ses relations sémantiques.

Ainsi, un concept issu d'un tableau de données respectant la structure d'un tableau de type *« individus × valeurs »* du point de vue structurel sera défini comme ayant un label, un

ensemble d'instances et un ensemble ordonné de propriétés qui définit sa structure (celle des ses instances).

Composant XML	Composante structurelle	Composant OWL
Nœud non terminal avec au moins un attribut ou fils	Un concept C_i	*owl:Class*
Nœud terminal sans fils ni attribut sur des données de type $\mathcal{T}_k$, i.e un nœud a^{jk} composante d'un nœud $\mathcal{P}_k$ ou un nœud $\mathcal{P}_j$ sans fils ni attribut	$\sigma_{\mathcal{A}} : \mathcal{A}_k \rightarrow C_m \times \mathcal{T}_k$ passant par une relation *« has_ a^{jk} »* ou *« has_ $\mathcal{P}_j$ »*	*owl:DatatypeProperties*

Tableau 5-3 Tableau de mapping des composants d'XML, structure d'ontologie et OWL

La prise en compte de ces dimensions d'un concept pose un nouveau problème qui est celui des instances. Il existe une divergence selon les approches sur leur intégration dans la structure de l'ontologie. On a d'ailleurs remarqué que dans les différentes approches de modélisation des ontologies, aucune n'intègre dans la structure les instances en dehors de celle de SOWA [Sowa, 1999] qui intègre une composante I constituée d'un ensemble de marqueurs individuels pour les instances de concept. Pour cause, une intégration d'un ensemble d'instances requiert l'ajout d'une relation permettant de les instancier aux concepts. Cette structure ressemble fortement à celle d'une *base de connaissances*.

Même si la frontière entre une *base de connaissances* et une ontologie est assez floue. Dans la littérature une *base de connaissances* est souvent présentée comme une ontologie ainsi que l'ensemble des instances individuelles des concepts ou classes.

En dehors du passage par XML-Schéma, cette stratégie se situe dans la même idéologie que celle présentée dans la section concernant la construction d'ontologies à partir de sources XML comme le montre le tableau 5.3.

Pour l'exemple du tableau 5.2, nous aurons donc individus classe *« Region »* ayant comme propriété le 5-uplet *« Superficies infestées (an2/2004) »* qui va apparaître comme un concept et les individus ou concepts *« Label_Saptial1 », « Podor »* et *« Saint-Louis »* vont alors constituer des instances de la classe *« Region »* qui va les susbsumer.

5.4.1.2 Subsomption des concepts candidats

La subsomption des concepts candidats est l'opération consistant à établir une relation de subsomption ou d'instanciation entre un concept candidat et un concept plus générique de l'ontologie de référence du domaine, *AOS* dans notre cas. Nous partons du principe que dans

le contexte actuel, en considérant que les connaissances abordées dans l'ontologie $\mathcal{AOS}$ et nos données abordant la même thématique agricole et environnementale, le système de mapping à $\mathcal{AOS}$ peut se faire en se basant sur le critère de l'homonymie. En se référant toujours à la structure générale présentée dans la figure 5.2, notre approche consiste à subsumer tout ce qui peut trouver un subsumeur. Ce sont donc les a_{jk}, $\mathcal{P}_j$, I_i et la variable C..

Il se présente également le cas des concepts qui ne trouvent pas de subsumeurs dans l'ontologie réutilisée $\mathcal{AOS}$. Pour ces cas de concepts, nous les subsumons automatiquement au concept généraliste « *Objet* » et laissons pour le moment dans le module de restructuration manuelle à implémenter le soin à l'expert du domaine de définir éventuellement des relations d'autres natures avec les ontologies existantes.

5.4.2 Extraction et inférence des relations sémantiques

Nous distinguons ici deux faattribut1les de relations : celle qui sont déduites de la hiérarchie des nœuds des documents XML et celles qui sont inférées grâce à l'ontologie $\mathcal{AOS}$. Nous précisons que lorsque nous abordons la notion de profondeur dans la hiérarchie d'un document XML, nous supposons de manière délibérée que la racine du document XML constitue le premier niveau numéroté à *« 1 »*.

5.4.2.1 Extraction de relations à partir du dataweb partenaire

Trois relations peuvent être extraites, déduites des documents XML. Ce sont les relations de subsomption, les relations associatives entre les concepts et les relations d'attributs.

5.4.2.1.1 Les relations de subsomption

La première faattribut1le est celles des relations de subsomption déduites des documents XML, elles sont implicites dans le document XML. Nous en avions déjà discuté dans la section précédente, l'origine tabulaire des données induit que dans un tableau de type *« individus × variables »,* la première case de la colonne des individus fait office de classe comme dans le langage objet et les autres individus font office d'objet dont la structure permet de définir celle des autres. Ainsi, chaque individu va constituer une instance de cette classe, d'où la relation de subsomption qui existe entre ce que nous nommons la variable privilégiée C et l'ensemble des individus I_i. C'est donc une relation entre deux concepts dans la structure de l'ontologie, sachant que dans la taxonomie nous aurons :

$$\forall\ i \in [1, n]\ \mathcal{H}_{part}^{C}(I_i, C)$$

Signifiant ainsi que I_i est un sous-concept ou subsume le concept C, chaque concept I_i va ainsi hériter de l'ensemble des attributs de C.

Cette relation sera convertie en description logique pour tout i sous la forme $I_i \sqsubseteq C$, et sous OWL par la structure *RDFs: subClassOf*.

L'identification des éléments de l'ensemble $\{I_i\ ,\ 1 \leq i \leq n\}$ est simple, ce sont selon la structure des documents XML, les éléments situés au troisième niveau.

5.4.2.1.2 Les relations associatives

La deuxième faattribut1le est celle des relations associatives permettant de faire la relation entre une variable privilégiée C, et une de ses propriétés $\mathcal{P}_j$ de dimension β-*aire*. Nous insistons sur cette caractéristique *n-aire* de l'attribut $\mathcal{P}_j$ qui est constitué par un ensemble d'éléments a_{jk} d'arité $\beta \in \mathcal{N}$.

Pour les intégrer dans la structure, nous définissons une relation nommée « $has\_\mathcal{P}_j$ » $\in \mathcal{R}_{part}$ dont la signature est celle des relations de type associative $\sigma_{\mathcal{R}} : \mathcal{R}_{part} \rightarrow (C_{part} \times C_{part})$.

Cette relation sera traduite sous forme d'*ObjectProperties* en OWL avec comme domaine C, et comme co-domaine $\mathcal{P}_j$.

Les propriétés $\mathcal{P}_j$ d'arité non nulle selon la structure et pour le moment de manière éprouvée sur les échantillons de données *partenaires* sont les nœuds non terminaux du quatrième niveau s'ils existent.

5.4.2.1.3 Les relations d'attribut

La troisième faattribut1le est celle des relations d'attribut permettant d'associer soit un élément ou individu non subsumeur I_i à une propriété $\mathcal{P}_j$ non *n-naire* ou une propriété $\mathcal{P}_j$ *n-aire* $n \in \mathcal{N}^*$ et une des ses composantes a^{jk}. C'est donc une relation d'attribut. Pour intégrer ce type de relation dans la structure de l'ontologie :

1. si la relation lie un I_i à une propriété $\mathcal{P}_j$ d'arité nulle sur des données de type $\mathcal{T}_j$, alors la relations associative est nommée « $has\_\mathcal{P}_j$ » $\in \sigma_{\mathcal{A}}$. Nous avons une relation $has\_\mathcal{P}_j(I_i)=v_i^j \in \mathcal{T}_j$ et sa signature est donc de la forme : $\mathcal{A}_{part} \rightarrow C_{part} \times \mathcal{T}_{part}$. Ce type de relation sera traduit dans OWL comme un *owl:DatatypeProperties* ayant comme domaine I_i et co-domaine le type $\mathcal{T}_j$.
2. si la relation lie une propriété $\mathcal{P}_j$ β-aire, i.e d'arité $\beta \in N^*$ et une des ses composantes a^{jk} parmi l'ensemble de ses propriétés d'arité β exprimant des mesures sur des données

de type $\mathcal{T}_j$, alors la relation associative est nommée $has\_a^{jk}(\mathcal{P}_j) \in \mathcal{T}_j$ ayant une signature est donc de la forme : $\mathcal{A}_{part} \rightarrow C_{part} \times \mathcal{T}_{part}$. Ce type de relation sera traduit dans OWL comme un *owl:DatatypeProperties* ayant comme domaine $\mathcal{P}_i$ et co-domaine le type $\mathcal{T}_j$.

Dans le cas où le même concept a une occurrence dans plusieurs documents, pour l'intégration semi-structurée avec des entités syntaxiquement siattribut1aires, étant dans le même domaine, nous avons utilisé le compromis consistant à définir une entité de ce genre comme un concept ayant l'union des attributs trouvés. Ainsi si un concept nommé *« label »* est extrait lorsqu'il se présente à nouveau nous fusionnons les attributs.

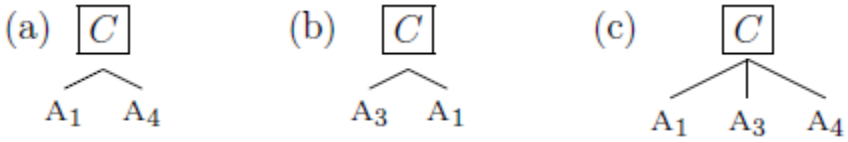


Figure 5-15 Intégration semi-structurée

La figure 5.15 montre un exemple où le concept C dans un document est extrait avec deux attributs $\mathcal{A}1$ et $\mathcal{A}4$ et dans un autre document avec les attributs $\mathcal{A}3$ et $\mathcal{A}1$, nous faisons alors l'union des attributs et ainsi ce concept aura l'attribut $\mathcal{A}1$ qui est présent dans les deux occurrences en plus des nouveaux attributs à considérer que sont $\mathcal{A}3$ et $\mathcal{A}4$.

Composants XML	Relation structurelle	Relation OWL
$\forall\ i \in [1, n]$, I_i et C	$\forall\ i \in [1, n]\ \mathcal{H}_{part}^{C}(I_i, C)$	*rdfs : subClassOf*
C, et une de ses propriétés Pj n -aire n $\in$ N^*	has_P_j(I_i, P_j)	*ObjectProperties* sous OWL avec comme domaine C, et comme co-domaine $\mathcal{P}_j$
Un individu I_i et une propriété $\mathcal{P}_j$ sur des données de type $\mathcal{T}_j$	$has\_\mathcal{P}_j(I_i)=v^{ji} \in \mathcal{T}_j$	*DatatypeProperties* ayant comme domaine I_i et codomaine le type $\mathcal{T}_j$.
Une propriété $\mathcal{P}_j$ d'arité $\beta \in \mathcal{N}^*$ et une des ses composantes a^{jk} parmi l'ensemble de ses β propriétés sur des données de type $\mathcal{T}_j$	$has\_a^{jk}(\mathcal{P}_j) \in \mathcal{T}_j$	DatatypeProperties ayant comme domaine $\mathcal{P}_i$ et co-domaine le type $\mathcal{T}_j$

Tableau 5-4 Tableau des correspondances des relations implicites d'XML à OWL

5.4.2.2 Extraction de relations sémantiques par inférence

Notre intérêt porte ici sur l'extraction automatique des relations sémantiques. Ces types de relations sont extraits en s'aidant des subsomptions effectuées sur des concepts d'*AOS*. Il faut préciser que cette phase peut être scindée en deux étapes réalisées après la construction des ontologies : une première par un raisonnement logique sur la structure

construite d'une ontologie et, la deuxième après l'implémentation sous OWL en se servant d'un moteur d'inférence.

L'utilisation d'*AOS* pour subsumer chaque concept permet d'extraire des relations sémantiques supplémentaires pour renforcer la taxonomie. En effet, si deux concepts sont subsumés par des concepts dans *AOS*, alors ils héritent de toutes les relations sémantiques et attributs de leurs subsumeurs.

Le tableau 5.4 résume la stratégie de mise en correspondance des éléments XML à OWL pour les relations de subsomption, celles associatives et celles d'attributs.

5.5 Construction des bases d'annotations

Les ontologies *partenaires* sont construites à partir des documents XML. Pour identifier la source XML originelle des concepts dans l'ontologie *partenaire*, nous utilisons des annotations. Le langage OWL que nous utilisons permet d'annoter des concepts et des propriétés selon un schéma de métadonnées prédéfini. Nous annotons les concepts de chaque ontologie *partenaire* avec un schéma d'annotation qui contient une propriété permettant de spécifier la source XML du concept à annoter (urlLocalSource).

```
<rdf :RDF(...) >
        (...)
<c_x rdf :ID="label_c_x">
        <urlLocalSource>
                file :/.../fichier1.xml/.../.../label_c_x
        </urlLocalSource>
</c_x>
<c_y rdf :ID="label_c_y">
        <urlLocalSource>
                file :/.../fichier1.xml/.../.../label_c_y
        </urlLocalSource>
</c_y>
        (...)
</rdf :RDF>
```

Figure 5-16 Extrait d'un format de base d'annotations partenaire d'un partenaire

La construction des bases d'annotations est faite en même temps que celle des ontologies *partenaires*. A chaque fois qu'un concept est extrait, nous l'annotons en spécifiant le chemin Xpath pointant vers le fragment du document XML d'où il est extrait. Ceci permet de faire le lien entre ce concept et l'ensemble de ses occurrences dans les documents XML [Sall et al., 2009].

La figure 5.16 montre l'extrait d'une *base d'annotations* d'un *partenaire*. Dans cet exemple on peut voir les deux concepts *« label_c_x »* et *« label_c_y »* et les chemins indiquant leur source d'origine.

5.6 Construction des ontologies génériques aux sources

Pour le partage des informations entre les *partenaires*, nous nous plaçons dans le contexte où les *partenaires* ne partagent que ce qu'ils ont en commun. Ainsi, chaque *partenaire* expose dans son *ontologie générique* les concepts de l'ontologie *partenaire* qu'il est susceptible d'avoir en commun avec les autres.

Considérons que l'ontologie *AOS* constitue déjà un vocabulaire commun à tous les *partenaires*, couvrant un domaine de connaissances plus large que celle abordé par nos données, et que tous les ontologies *partenaires* se réfèrent à elle pour renforcer la taxinomie de leurs concepts. On peut alors utiliser cette même ontologie et la subsomption faite sur les concepts de l'ontologie *partenaire* pour déterminer les concepts qui vont constituer l'*ontologie générique* d'un *partenaire*.

```
<rdf :RDF (...) >
        (...)
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_x">
        <rdfs :label xml :lang="FR"> label_c_x </rdfs :label>
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_a"/>
</owl :Class>
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_y">
        <rdfs :label xml :lang="FR">label_c_y</rdfs :label>
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_b"/>
</owl :Class>
<owl :Class rdf :about="Http ://.../ sall/Dataweb_Partenaire1#c_z">
        <rdfs :label xml :lang="FR">label_c_z</rdfs :label>
</owl :Class>
        (...)
</rdf :RDF>
```

Figure 5-17 Extrait d'un format d'une ontologie générique en construction

Pour ce faire, nous considérons tout concept de l'ontologie *AOS* subsumeur comme étant un concept générique, d'où sa possibilité d'être subsumé par d'autres concepts d'autres *partenaires*. Nous instancions alors dans l'*ontologie générique* ce concept ainsi que le lien de subsomption qui le lie avec les concepts subsumés. Cela permet de spécifier qu'il existe, dans l'*ontologie générique*, un concept général de l'ontologie *AOS* spécialisé par certains concepts

de l'ontologie *partenaire*. Par conséquent il sera partagé par ces derniers avec les éventuels autres concepts des autres ontologies *partenaires* qui l'auront spécialisé.

Signalons aussi que l'*ontologie générique* peut être considérée comme une extension de l'ontologie *AOS*.

La figure 5.16 montre l'extrait d'une *base d'annotations partenaire*. Pour les deux concepts apparaissant dans le code on peut distinguer leur source Xpath d'origine dans les *dataweb partenaires*.

La figure 5.17 expose un extrait de l'*ontologie générique* du même *partenaire*. Les deux premiers concepts subsument ceux de l'ontologie *AOS*, et par conséquent, font partie de l'*ontologie générique*. Le dernier concept ne peut pas être généralisé dans l'ontologie *AOS* et ne figure pas donc dans l'*ontologie générique*.

L'*approche dataweb sémantique* permet de réaliser l'intégration sémantique des données au sein des *partenaires*. Il en résulte un ensemble de *dataweb sémantique*s correspondant chacun à un *entrepôt de documents XML* enrichi d'une *base de connaissances*. Cette dernière est composée :

— d'une ontologie *partenaire* décrivant la sémantique des données ;
— d'une *base d'annotations* décrivant les correspondances entre l'ontologie *partenaire* et les documents XML;
— d'une *ontologie générique* contenant les concepts que l'ontologie *partenaire* est susceptible d'avoir en commun avec les autres ;

Pour réaliser l'intégration des données de tous les *partenaires*, il convient de faire la médiation de ces différents *dataweb sémantique*s.

5.7 Un système à base de hubs pour une médiation entre partenaires

Rappelons la problématique de l'intégration des données dans le contexte de nos travaux. Il s'agit d'intégrer afin de partager et d'interroger des données distribuées sur plusieurs sources hétérogènes et appartenant à différents *partenaires*. La résolution de ce problème a nécessité la réalisation d'une première phase, consistant à l'intégration structurelle des données au sein des *partenaires*. Elle conduit à la création d'un *dataweb* pour chaque *partenaire*. Après cette phase le problème se résume à l'intégration des différents *dataweb*.

Pour ce faire, nous avons proposé une approche reposant d'abord sur la construction de *dataweb sémantique*s, dans le but de résoudre les problèmes liés à l'hétérogénéité

sémantique des données de chaque *partenaire*, et à la médiation de ces différents *dataweb sémantique*s [Sall et al., 2009]. C'est de cette deuxième phase de médiation que nous étudions cette partie. Nous proposons un système à base de hubs (ou serveurs pairs) [Gandon et al., 2008] pour effectuer une médiation entre les différents *partenaires* contenant ces *dataweb sémantique*s.

5.7.1 Architecture du système

Le système proposé dans [Gandon et al., 2008] repose sur des hubs ou serveurs pairs dont l'architecture est la même partout où ils sont déployés. Un hub contient systématiquement:

— Une interface utilisateur : un serveur web proposant des applications accessibles aux utilisateurs par leur navigateur web ;

— Une interface programmatique : des services web proposant un accès distant aux

— applications du hub pour d'autres applications.

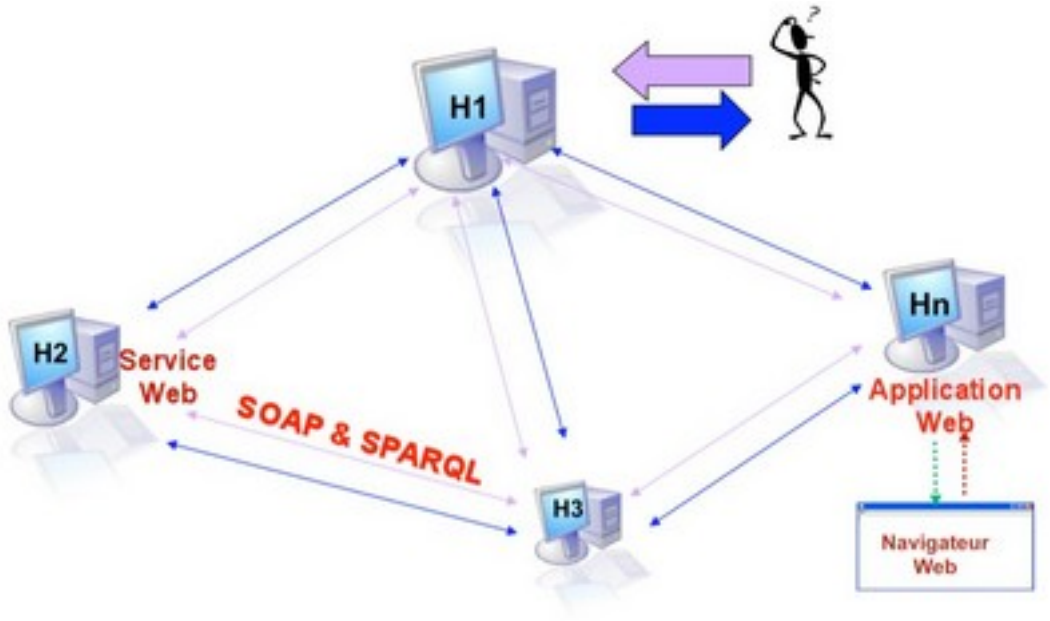


Figure 5-18 Architecture globale d'un système à base de hubs

Dans cette architecture, illustrée en figure 5.18, un utilisateur peut se connecter à n'importe quel hub pour utiliser une application web et notamment pour soumettre une requête. Pour cet utilisateur et sa demande, ce hub est alors chargé d'identifier les autres hubs susceptibles de l'aider à répondre à la requête et d'orchestrer la résolution distribuée de la requête avec les hubs identifiés. Pour ce contexte deux hypothèses fortes ont été fixées :

— Chaque hub a les mêmes ontologies que les autres, i.e. les mêmes schémas RDFS et OWL sont répliqués sur chaque hub.

— Les hubs sont en nombre restreint et leur connectivité est stable.

Un hub implante les interfaces SPARQL et ceux nécessaires à la gestion de la distribution des requêtes. Chaque hub propose aussi une application web permettant de déclarer les autres hubs et en particulier l'URL du point d'accès à leurs services web. Cette description est elle même une annotation RDF du hub. Avec cette architecture, chaque hub peut donc envoyer des requêtes à n'importe quel autre hub.

5.7.2 Adaptation du système par rapport à notre contexte

Pour effectuer la médiation entre les différents *partenaires*, nous reprenons ce système en l'adaptant à notre contexte. Chaque *partenaire* dispose alors d'un hub qui contient le *dataweb sémantique* dudit *partenaire*.

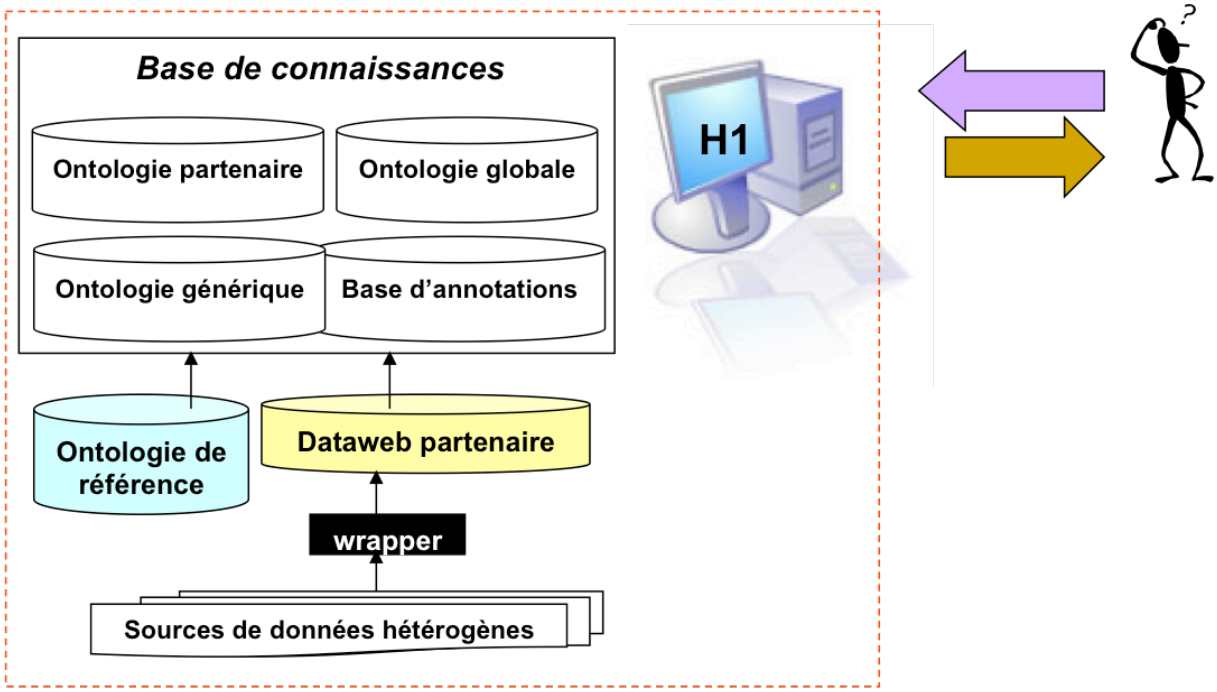


Figure 5-19 Architecture globale d'un hub

Etant donné que les travaux présentés dans [Gandon et al., 2008] concernent des sources d'information et d'annotations potentiellement distribuées, les annotations sont réparties sur des bases différentes, elles mêmes réparties sur des serveurs géographiquement distribués.

De ce fait, une requête peut impliquer des éléments d'annotation distribués entre plusieurs de ces serveurs. Cela les a conduits à adopter une approche consistant à orchestrer

une résolution collective d'une requête par plusieurs hubs. Les travaux réalisés dans [Gandon et al., 2008] et allant dans ce sens ne seront pas alors repris par notre système.

En effet, dans notre contexte une requête ne peut impliquer qu'un seul hub à la fois. Cela est dû au fait que nous ne disposons, dans chaque *dataweb sémantique*, que d'annotations spécifiques au *partenaire*.

Cependant, les *partenaires* peuvent partager les mêmes informations même si dans notre contexte ce partage ne concerne que les informations sur lesquelles les *partenaires* ont un vocabulaire commun. L'hypothèse faite dans [Gandon et al., 2008] sur les ontologies reste valable, c'est-à-dire que les hubs partagent une ontologie commune qui est utilisée pour la résolution des requêtes distribuées. Mais nous ne procédons pas à une réplication de cette ontologie sur tous les hubs. Nous la construisons plutôt de manière collaborative [Niang, 2008] grâce à l'existence de l'ontologie de référence et l'utilisation d'un moteur de recherche sémantique.

5.8 Construction de l'ontologie globale

L'*ontologie globale* constitue un vocabulaire commun à tous les *partenaires*. Ce vocabulaire existe déjà, mais il est distribué entre les différents *partenaires*. En effet, en construisant son *ontologie générique*, constituée par l'ensemble des concepts qu'il peut avoir en commun avec les autres, le *partenaire* construit une partie du vocabulaire commun. La création de ce vocabulaire permettra alors de regrouper toutes les parties des *partenaires*.

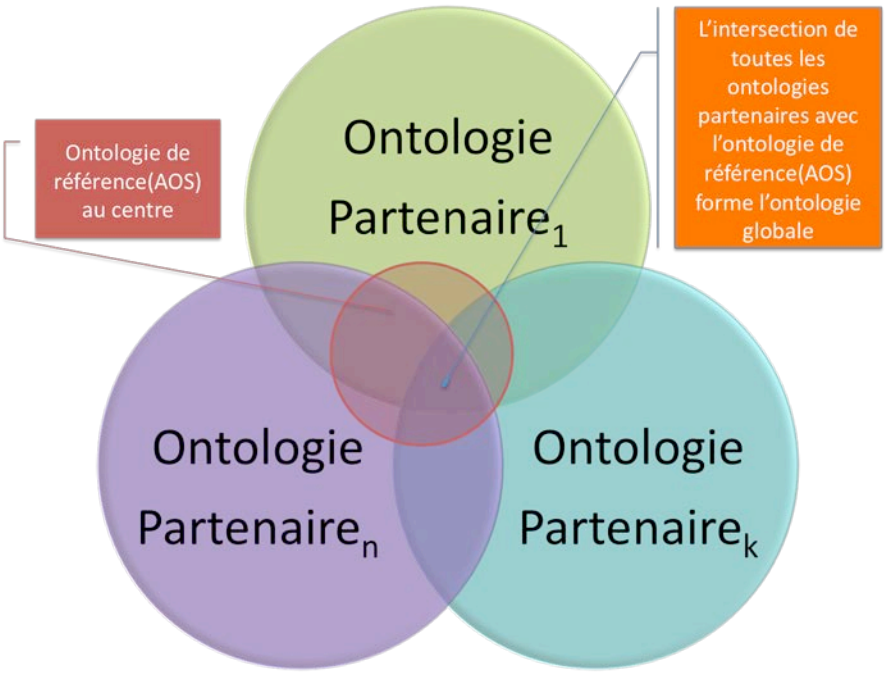


Figure 5-20 Schéma sur la construction de l'*ontologie globale*

En d'autres termes, nous allons procéder à la fusion collaborative des ontologies génériques des *partenaires* pour constituer une *ontologie globale* commune à tous ces derniers.

Cette ontologie représente le point d'entrée du système. Elle sert de support pour l'expression des requêtes et fait office d'interface pour les ontologies *partenaires*, qui permettent d'interroger les sources.

C'est un ensemble de concepts classifiés en une hiérarchie laquelle ne contient pas de propriété reliant les différents concepts. Ces propriétés existent chacune dans leur ontologie *partenaire*. Une requête peut donc contenir des concepts de l'*ontologie globale* et éventuellement des propriétés des ontologies *partenaires*. Pour disposer de l'*ontologie globale*, chaque hub devra, après avoir déclaré un autre, lui demander son *ontologie générique*, par l'intermédiaire d'un service web dédié à cette demande.

La fusion de l'*ontologie générique* de ce *partenaire* avec celles qu'il reçoit des autres va constituer l'*ontologie globale* pour ce *partenaire*. Cette fusion se fait directement par l'intermédiaire d'un moteur de recherche sémantique, que nous utilisons également pour l'interrogation des différentes bases de connaissances.

5.9 Conclusion

Dans ce chapitre, nous avons traité du processus d'intégration sémantique et structurelle. D'abord, notre présentation concerne la nature et la structure des données initiales, à savoir des données de nature environnementale structurées sous forme de tableaux de données de nature *« individus*variables »*.

L'*approche dataweb sémantique* que nous avons proposée est basée sur la construction d'un *entrepôt de documents XML* dit *dataweb* pour chaque *partenaire*. Cette approche permet l'intégration structurelle des documents, résolvant du même coup dans le processus d'intégration la problématique de la dimension organisationnelle des données avec leur aspect privé et propriétaire.

Ensuite, a été présenté le processus d'apport sémantique aux entrepôts par l'extraction semi-automatique d'ontologies à partir de ses documents XML. L'utilisation du vocabulaire utilisé pour décrire les données permet de capturer le vocabulaire utilisé dans le domaine. Ce qui est du coup permet de constituer un début de taxonomie, renforcé par l'utilisation d'une ontologie externe de référence. L'utilisation de cette ontologie de référence dont certains de

ses concepts subsument ceux des ontologies construites permettent la découverte et l'héritage de relations sémantiques, renforçant ainsi sa structure.

D'un autre côté, nous avons exposé le processus de construction des bases d'annotations qui permettent de faire le lien entre le niveau des données et celui de la connaissance sémantique qui leur est associée par l'intermédiaire des ontologies. Cela permet de faire le lien entre le niveau de données et celle des connaissances les décrivant. Permettant ainsi d'utiliser le niveau de connaissances pour régler la problématique de l'intégration des sources de données. Il suffit pour cela de faire inter-opérer les niveaux sémantiques.

L'objectif visé est la coopération des différents *dataweb sémantique*s, laquelle passe par la construction d'une *ontologie générique* pour chaque source. Cela permet à un *partenaire* de déclarer aux autres les connaissances qu'il désire partager avec eux. Ces ontologies génériques rendent possible la construction collaborative d'une *ontologie globale* par un mécanisme d'échange, utilisée avec un système à base de hubs pour la médiation entre les différents *partenaires* du système d'intégration.

Nous avons appliqué cette approche dans le contexte du projet *SIC-Sénégal*. Dans la partie qui suit, nous allons présenter le contexte applicatif, ainsi que le cahier des charges des attentes par rapport à nos travaux dans notre équipe de recherche dans le contexte des différentes approches et autres pistes de recherches ouvertes par le contexte de mise en valeur de la vallée du fleuve Sénégal.

Dans le prochain chapitre, nous présentons le domaine d'application cible dans la vallée avec les données environnementales ainsi que le prototype implémentant l'*approche dataweb sémantique*.

Troisième partie

Validation

Chapitre 6
Les données environnementales comme domaine d'application cible

Sommaire

6.1 Introduction 148

6.2 Le projet *SIC*-Sénégal 149

6.2.1 Description du projet 149

6.2.2 Participation et tâches dans l'équipe BDI*SIC* 150

6.2.2.1 Tâches de cette thèse dans ce contexte 151

6.2.2.2 Autres approches et contributions dans l'équipe 152

6.2.3 Ressources existantes 153

6.2.3.1 Relevés de données 153

6.2.3.2 Le Service d'Ontologie Agricole(SOA) 154

6.2.3.3 Jena 156

6.2.3.4 Corese(Conceptual Resource Search Engine) 157

6.2.3.5 SeWeSe (Semantic Web Server) 158

6.2.4 Evaluations 159

6.3 Construction des *dataweb partenaires* 160

6.3.1 Extraction et transformation des données sources 161

6.3.2 Restructuration des données sources 162

6.3.2.1 Restructuration ciblant les caractéristiques spatiales 163

6.3.2.2 Restructuration des caractéristiques temporelles 164

6.3.2.3 Restructuration ciblant les unités de mesure 165

6.4 Construction des bases de connaissances 167

6.4.1 Processus d'extraction des ontologies *partenaires* 167

6.4.3 Construction des ontologies génériques 173

6.4.4 Construction de l'*ontologie globale* 174

6.5 Une médiation utilisant un système à base de hubs 174

6.6 Validation des ontologies 175

6.7 Conclusion 176

6.1 Introduction

L'environnement est devenu depuis quelques décennies l'un des centres d'intérêts les plus importants de la communauté scientifique. La compréhension de l'évolution des facteurs environnementaux ainsi que la prévision de l'impact des activités de l'homme sur celui-ci sont devenues vitales à la pérennité de la survie de l'Homme sur terre.

Différentes études environnementalistes s'intéressent à observer les comportements et à détecter les moindres changements pendant que d'autres s'évertuent à en prévoir l'évolution.

Ces études, entre autres spécificités du point de vue des observations produisent ainsi une masse de données importante et difficilement exploitable de manière automatique. En plus de la durée de validité des données relativement importante, les données proviennent de sources diverses, hétérogènes et réparties : bases de données, fichiers plats, systèmes d'information géographique, etc. Du point de vue connaissances, les facteurs et attributs environnementaux étant fortement liés, l'interprétation et la compréhension des données nécessite une maîtrise de leur signification ainsi que des différentes relations existantes entre elles.

Comme introduit dans [Lo, 2002] l'approche la plus adaptée dans ce contexte est celle dite « conservation/observation ». L'aspect conservatoire exprime le fait qu'il y a acquisition et conservation de données ou de documents et que cette information n'est pas mise à jour mais archivée pour ensuite être restituée à l'utilisateur final. L'aspect observatoire exprime le fait que les usagers de tels systèmes peuvent effectuer des observations (mesures, synthèse, recherche, extraction, etc.) [Dzeakou et Derniame 1998].

Ce contexte constitue le domaine d'application de nos travaux dans le cadre du projet *SIC*-Sénégal. L'aspect conservation des données est pris en compte en constituant un système d'intégration structurel pour chaque ensemble de document provenant d'une même source. L'aspect observation des données est pris en compte en appliquant les méthodes d'intégration de données sémantiques de nos travaux à l'entrepôt ciblé utilisant ainsi l'*approche dataweb sémantique*. Mieux, dans un cadre comme la vallée du fleuve Sénégal où plusieurs organismes interviennent, notre approche permet une intégration des différentes connaissances des différents organismes. Le succès de leur combinaison justifie la notion de *partenaire* ou partenariat choisie désignant une mise en commun des hétérogénéités pour atteindre une même finalité, celle de l'intégration participative des données distribuées et propriétaires sur la vallée du fleuve Sénégal Ce chapitre présente le cadre d'application de nos

travaux dans le contexte du projet *SIC-Sénégal* dans la vallée du fleuve Sénégal. La première section présente les objectifs du projet *SIC-Sénégal* ainsi que les ressources existantes et leur évaluation.

6.2 Le projet SIC-Sénégal

Le projet *SIC-Sénégal* est une initiative, dans un contexte de mise à disposition de manière ouverte de données brutes environnementales, de leur donner un sens exploitable. Ce projet vise à mettre en relation ces données, à extraire la connaissance qu'elles véhiculent créant ou découvrant ainsi les relations existantes entre elles. Plus la masse de données est importante, plus le vocabulaire contrôlé constitué couvre le domaine et plus une donnée comme dans le cas des cartes cliquables peut renvoyer à de nouveaux détails, de nouvelles informations en relation avec elle.

Dans cette partie, nous allons présenter le projet *SIC-Sénégal*, le contexte de l'équipe *BDISIC (Bases de Données et Ingénierie des Systèmes d'Information et de la Connaissance)* ainsi que les ressources structurelles et sémantiques existantes dans le projet.

6.2.1 Description du projet

Les différentes études sur l'analyse des modifications des impacts environnementaux dans le nord de la vallée du fleuve Sénégal (le sud-ouest mauritanien) montrent à quel point la sécheresse des années 70 a marqué l'environnement dans cette partie de la vallée du fleuve Sénégal. Il apparaît dans ces études que la mise en valeur de la vallée du fleuve Sénégal, avec les grands aménagements hydro-agricoles, a fini par faire de l'homme un acteur important de la modification de l'environnement avec des conséquences multiples.

L'aménagement des rizières au détriment des zones d'inondation ainsi que celui des barrages ont eu par exemple pour conséquences la désorganisation du système de pâture au niveau de l'élevage, le déplacement des populations, le délaissement des cultures de crue en plus de l'apparition de maladies avec l'arrêt de la remontée de la salinité comme la bilharziose.

Récemment, la vallée est devenue une zone cible des oiseaux granivores avec leurs conséquences inhérentes sur les autres types de cultures granulaires, la nutrition des populations, les semences et le remboursement des dettes paysannes. Depuis de nombreuses années, des données issues de l'étude des observations sur ces diverses thématiques sont produites par les différents experts intervenant dans la vallée. La table 6-1 donne un exemple de tableau de données.

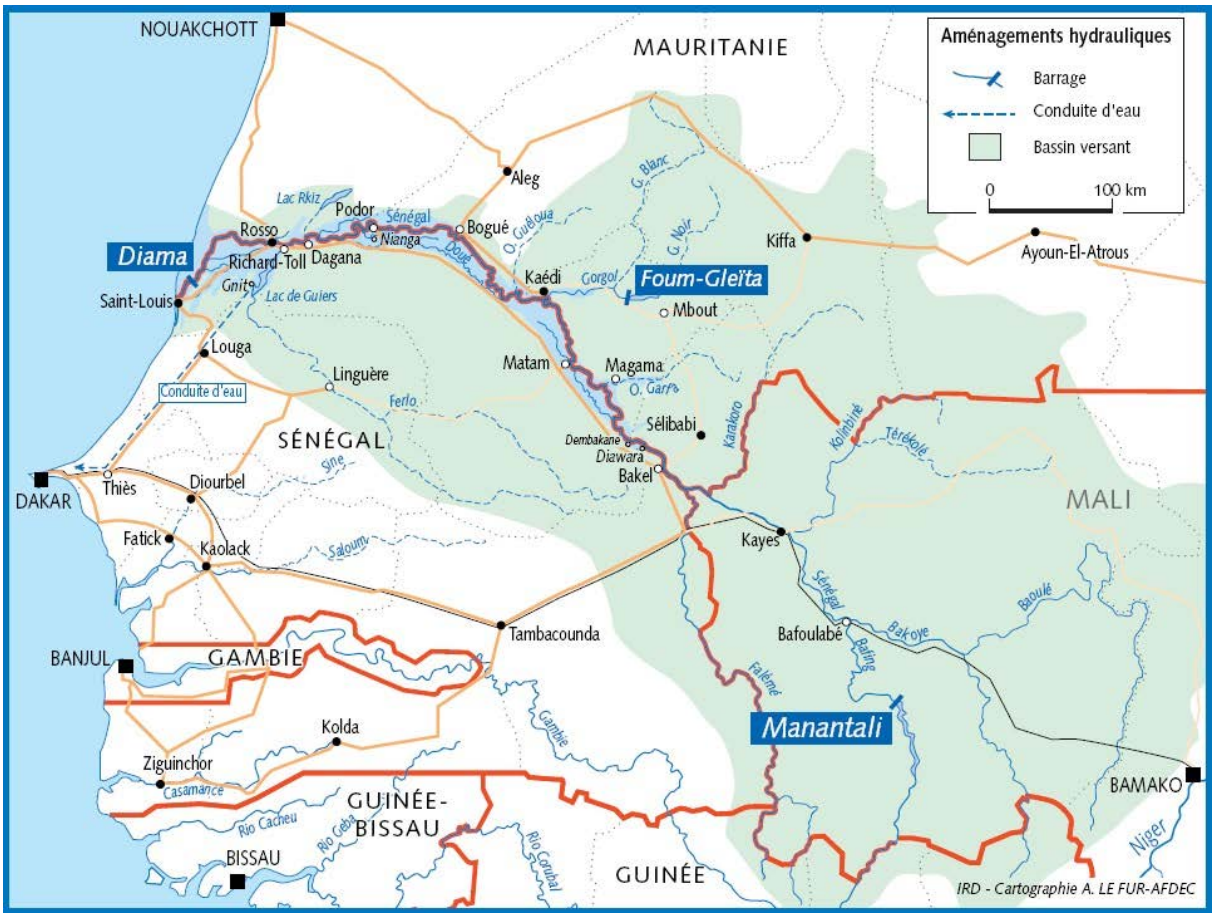


Figure 6-1 La vallée du fleuve Sénégal

Ce qui montre la diversité des thématiques et à quel point la masse de données est importante sur la vallée ainsi que l'interaction des études qui y sont menées. La combinaison de ces connaissances nécessite l'existence d'un modèle fédérateur structurel et sémantique.

6.2.2 Participation et tâches dans l'équipe BDISIC

L'équipe BDISIC s'intéresse aux thématiques liées aux bases de données et aux systèmes d'informations et de connaissances. Le projet *SIC-Sénégal* offre plusieurs thématiques de recherche de l'intégration des données, l'imputation des données manquante, à leur exploitation plus efficace du point de vue datamining avec une maîtrise des connaissances décrivant les données.

Nous allons dans un premier temps présenter les tâches et contributions attendues de cette thèse dans le contexte du projet, les autres approches utilisées notamment basées sur les systèmes pair à pair ou ceux consistant à transformer toutes les données en RDF ou l'intégration d'applications.

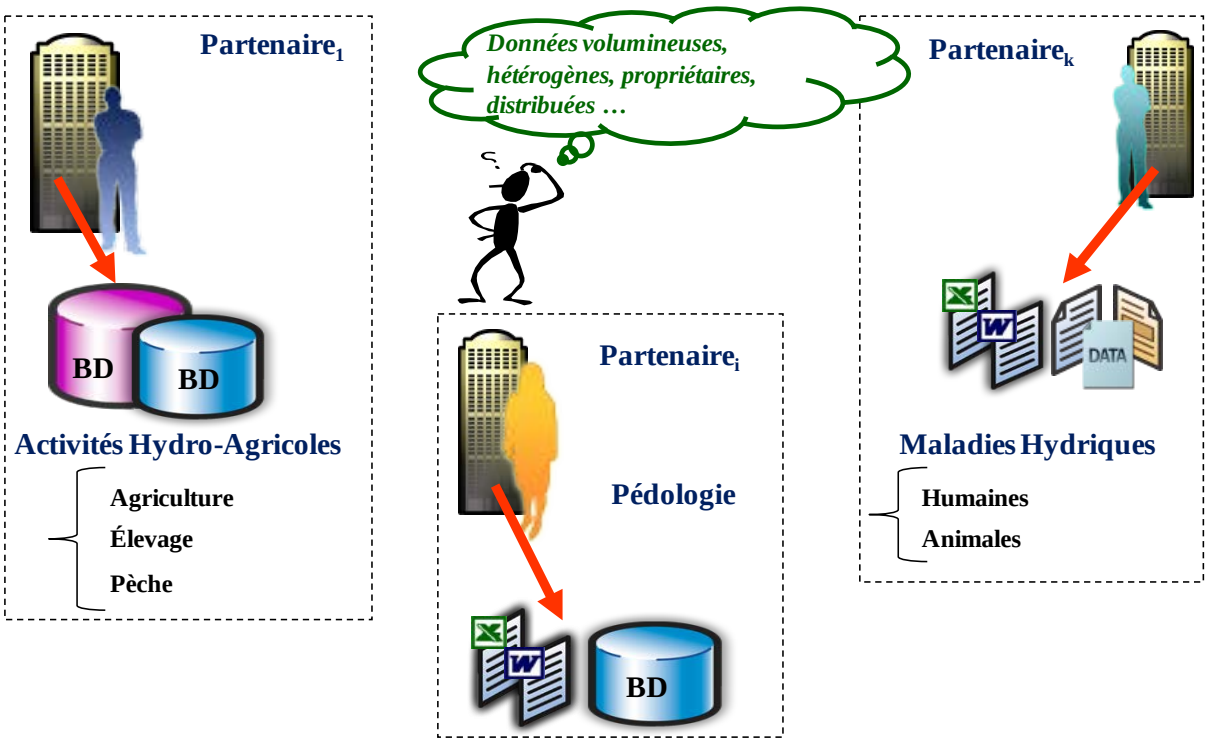


Figure 6-2 Problématique du besoin d'intégration dans la vallée du fleuve Sénégal

6.2.2.1 Tâches de cette thèse dans ce contexte

Une première approche pour l'intégration de données dans le contexte du projet a été initiée dans les travaux exposés dans [Lo, 2002] avec l'*approche dataweb*. Cependant, entre autres, cette approche présente des limites lorsqu'il s'agit d'intégrer les connaissances chez un organisme fournisseur de données et globalement dans un contexte où il faut intégrer des connaissances issues de plusieurs sources en gardant l'aspect propriétaire et privés des données. Cette limite est essentiellement liée à l'utilisation du langage XML pour la description des connaissances sur les données et l'utilisation d'un entrepôt pour l'intégration de l'ensemble des données des *partenaires*. Ainsi les résultats attendus de cette thèse sont une extension du modèle de *dataweb* proposé permettant ainsi à chaque organisme participant de disposer de son propre *dataweb* ainsi qu'une approche d'intégration des connaissances du *partenaire* d'une part et entre les *partenaires* d'autre part. Notre tâche dans ce projet consiste essentiellement à proposer un cadre le plus automatique possible permettant l'intégration structurelle et sémantique des données permettant ainsi de le préparer à une exploitation plus intelligente.

L'étude de la problématique nous a orienté vers la substitution de l'utilisation d'XML à celle des ontologies extraites semi-automatiquement des données *partenaires* représentées en XML.

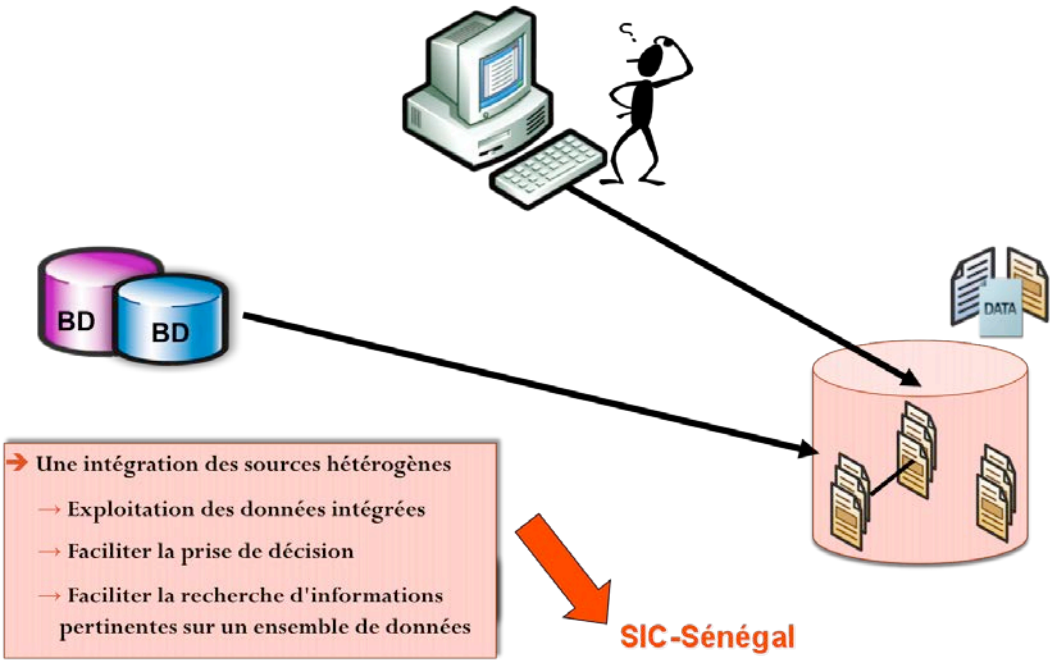


Figure 6-3 Contributions au projet SIC-Sénégal

Nous homogénéisons ainsi structurellement les données *partenaires* en étendant la notion de *dataweb* à celle dite *dataweb sémantique*, c'est-à-dire un *entrepôt de documents XML* issues de sources hétérogènes muni d'une base de connaissance à base ontologique extraite du *dataweb* et décrivant sémantiquement ses données.

6.2.2.2 Autres approches et contributions dans l'équipe

D'autres approches ont été expérimentées pour la médiation des données, notamment celle utilisée dans SenPeer [Faye et al., 2006]. Cette approche consiste à utiliser les systèmes pair-à-pair pour permettre la médiation de données décentralisée. Ainsi dans cette approche, l'architecture finale substitue aux hubs *partenaires* des nœuds du système de pair-à-pair et aux ontologies des réseaux sémantiques. Dans la démarche d'intégration nous adoptons une architecture quasi-identique sous la forme d'un système à base de hubs utilisant des ontologies, mais en plus l'organisation par domaine de connaissances des données sera plus facile, il suffit de dissocier l'*ontologie globale* servant de médiateur dans notre approche en plusieurs sous-ontologies thématiques, évitant ainsi de mettre en place un réseau de super-pair par domaine.

Comme proposée par M. Klein [Klein, 2002], une approche consisterait à transformer automatiquement toutes les données XML en RDF, cette approche est à la base de la méthodologie d'intégration des applications et de recherche d'informations et d'applications proposées. Cette approche est utilisée dans les travaux de Fatou Kamara. En plus des limites

du langage RDF ayant poussé vers l'utilisation des ontologies, nous n'avons pas adopté cette approche pour être conforme à celle web sémantique consistant à séparer le niveau des données à celle des connaissances largement discutée dans l'état de l'art. Nous avons donc dissocié les données et les connaissances en les extrayant et enrichissant grâce à une ontologie existante plus générique.

6.2.3 Ressources existantes

Dans cette section, nous étudierons la variété et la typologie des échantillons de données fournies par les organismes participant au projet d'intégration. Nous présenteront également les principales ressources logicielles utilisées comme *Corese, Jena, SeWeSe* et l'ontologie externe de référence du domaine *AOS*.

6.2.3.1 Relevés de données

Comme énoncé la région de la vallée du fleuve Sénégal est ciblée par plusieurs études et observations. Ces observations portent sur la dynamique de variables environnementales. Les données issues des observations peuvent donc varier des types de données pluviométriques, aux données sur la santé comme l'évolution du paludisme ou carrément le recueil de données biologique sur l'observation d'espèces. En résumé il peut s'agir de rapports sous format textuel, de statistiques, ou données au format multimédia. Par conséquence, dans cette application, les données peuvent provenir généralement : de bases de données, de systèmes d'information géographique, de fichiers Excel (résultats d'enquêtes), etc.

Ces sources de données appartiennent, le plus souvent, à des entités différentes : SAED, ISRA (Institut Sénégalais de Recherche Agronomique), etc. Les types de données sont nombreux et complexes : il y a des données attributaires (celles provenant de bases de données relationnelles par exemple) et des données géographiques (provenant des SIG comme la carte sur les sols dans la vallée du fleuve Sénégal). Les données sont elles-mêmes hétérogènes du point de vue structurel mais homogènes du point de vue sémantique vu qu'elles représentent les mêmes objets du monde réel. Dans ce contexte d'application, nous nous sommes intéressés aux données statistiques. Les données statistiques représentent le résultat sous forme de tableaux d'observations sur des caractéristiques d'individus dans un contexte donnés.

Ces entités sont appelées *partenaires* dans le contexte du projet. Ce sont des organismes désireux de partager leurs données de même domaine avec les autres. Ce qui nous

permet de préciser cet aspect du partage assez particulier qui ne concerne que les domaines de connaissances en commun. Ce qui du coup facilite et assure le processus extraction de la sémantique de la structure de leur données sous forme de concepts d'ontologie et leur alignement.

Régions	Superficies infestées (2002/2003)				
	Mil(ha)	Arachide(ha)	Niébé(ha)	Pastéq./beref(ha)	Riz(ha)
Dagana	3400	1800	100	400	5500
Podor	2700	600	1800	1100	2500
Saint-Louis	70	5500	3500		

Tableau 6-1 Exemple de tableau de données partenaire (Ici de la DRDR-SL)

Une particularité des données environnementales, nous l'avons déjà souligné, est de se présenter sous la forme d'une statistique sur un ensemble de caractères communs à une population ou ensemble d'individus. Chaque individu de cet ensemble constitue un cas particulier d'une abstraction plus générale qui résume leurs caractères communs. La table 6.1 illustre un exemple de tableau de données où l'on s'intéresse à l'évolution d'une infection sur plusieurs types de cultures dans trois régions (qui sont les individus) données de la vallée du fleuve Sénégal.

La structure générale des tableaux de cette nature est déjà donnée dans le chapitre précédent.

6.2.3.2 Le Service d'Ontologie Agricole (SOA)

Nous avons utilisé l'ontologie *AOS* (Agricultural Ontology Service) de la l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO) comme ontologie de référence pour enrichir les ontologies construites. *AOS* est un service d'ontologie agricole construit à partir du thésaurus AGROVOC depuis 2003 par la FAO. Ayant pour objectif de constituer une base multilingue de concepts dans le domaine de l'agriculture (le serveur de concept (SC)), il prend en considération les relations sémantiques et lexicales de manière plus finie et précise.

Cette initiative, liée au projet global de la FAO de mettre en place un service d'ontologie en agriculture (*AOS*), est censée fonctionner comme un outil de structuration et de standardisation de la terminologie en plusieurs langues avec pour objectif de l'utiliser dans différents systèmes. A partir du SC, le thésaurus traditionnel AGROVOC et d'autres formes d'organisation des connaissances (KOS) pourront être exportés. Par la même opération, des

concepts ontologiques peuvent être extraits et utilisés pour la construction d'ontologies spécifiques par domaine de connaissances.

Par ailleurs, la FAO entend instituer une base de référence mondiale commune :

— aux ressources de langages scientifiques

— aux outils pour décrire des vocabulaires, des plans de classification, des ressources terminologiques au service de la description, de l'enregistrement et du repérage de l'information relative à l'agriculture et à l'alimentation.

Exemple de définition du concept Abattoir dans *AOS :*

```
<owl:Class RDF:about="Http ://www.fao.org/aos/agrovoc #c _8">
        <rdfs:label xml:lang="AR">بصاقم</rdfs:label>
        <rdfs:label xml:lang="CS">jatky</rdfs:label>
        <rdfs:label xml:lang="DE">SCHLACHTHAUS</rdfs:label>
        <rdfs:label xml:lang="EN">Abattoirs</rdfs:label>
        <rdfs:label xml:lang="ES">Mataderos</rdfs:label>
        <rdfs:label xml:lang="FR">Abattoir</rdfs:label>
        <rdfs:label xml:lang="HU">vágóhíd</rdfs:label>
        <rdfs:label xml:lang="JA">屠場</rdfs:label>
        <rdfs:label xml:lang="PT">Matadouro</rdfs:label>
        <rdfs:label xml:lang="SK">bitúnky</rdfs:label>
        <rdfs:label xml:lang="TH">โรงฆ่าสัตว์</rdfs:label>
        <rdfs:label xml:lang="ZH">屠宰场</rdfs:label>
        <rdfs:subClassOf rdf:resource="Http://www.fao.org/aos/agrovoc #c _25201"/>
                <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty rdf:resource="Http://www.fao.org/aos/agrovoc #r _90"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="Http://www.fao.org/aos/agrovoc #c _4674"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
        </rdfs:subClassOf>
                <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty rdf:resource="Http://www.fao.org/aos/agrovoc #r _90"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="Http://www.fao.org/aos/agrovoc #c
        _28609"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
        </rdfs:subClassOf>
</owl:Class>
```

Il existe aussi d'autre norme de la FAO comme AgMES (Série d'Eléments de Métadonnées Agricoles). AgMES est la norme de métadonnées développée par la FAO pour la description et la découverte de ressources d'information agricole. AgMES fournit un ensemble d'éléments de métadonnées qui peuvent être utilisés pour décrire tous les types de ressources d'information dans les domaines de l'agriculture, de la sylviculture, de la pêche, de la sécurité alimentaire et d'autres domaines associés. (Wkipédia).

En guise de contributions, l'*AOS* offre un noyau sémantique clair constituant une ontologie réutilisable. Il convient d'appeler une ontologie pour signifier que la classification tend à se situer hors cadres du thésaurus classique et à toucher des aspects plus profonds comme la causalité, la correction/rémédiation, la succession temporelle, les effets par zone climatique, certains facteurs biologiques, etc. En effet, l'élaboration d'une taxonomie biologique qui sous-tend une capacité d'interférence à offrir pour favoriser un repérage sélectif, assisté s'avère d'une importance certaine. Il s'agit d'un registre sémantique de domaine proposé par la FAO. AGROVOC, un thésaurus multilingue existant, constituera le noyau de départ. Dans chacune des cinq langues officielles de la FAO (anglais, français, espagnol, arabe, chinois), les concepts ont des termes correspondants.

Ce noyau d'*AOS* couvrant tous les domaines ayant trait à l'agriculture, à la pêche, à l'alimentation et aux domaines connexes (l'environnement, par exemple) permet de disposer d'une structuration assez générique du domaine environnemental, cause pour laquelle nous la réutilisons dans notre travail pour subsumer les concepts que nous allons extraire des tableaux de données structurés XML afin que pour deux concepts ayant des subusmeurs, nous puissions extraire des relations sémantiques par héritage. Pour cela, nous utilisons la version en OWL. A terme *AOS* devrait être composé d'un ensemble d'ontologies de domaine. Dans tout le document nous ferons la confusion volontaire entre *AOS* et l'ontologie OWL AGROVOC.

6.2.3.3 Jena

JENA est un framework permettant de lire et de manipuler des ontologies décrites en OWL et d'y appliquer certains mécanismes d'inférences grâce au langage SPARQL.

Il offre également un environnement facilitant le développement d'applications dédiées au web sémantique. Il fournit un environnement permettant de manipuler des documents de type RDF, RDFS et OWL, et fournit en plus un moteur d'inférence à base de règles.

Cette API permet le raisonnement sur les ontologies, mais aussi le traitement des données XML.

C'est d'ailleurs son moteur d'inférence qui est utilisé dans Protégé. Dans certains outils, le validateur OWL intégré de Jena est utilisé pour la validation des ontologies OWL.

Nous l'avons d'ailleurs utilisé pour accéder à l'ontologie *AOS*, construire les nouvelles ontologies et procéder à la validation logique des ontologies construites. La figure 6.4[40] montre l'architecture de Jena avec ses quatre composants que sont le noyau du modèle RDF, le module de manipulation des ontologies, son moteur d'inférence et celui de réification.

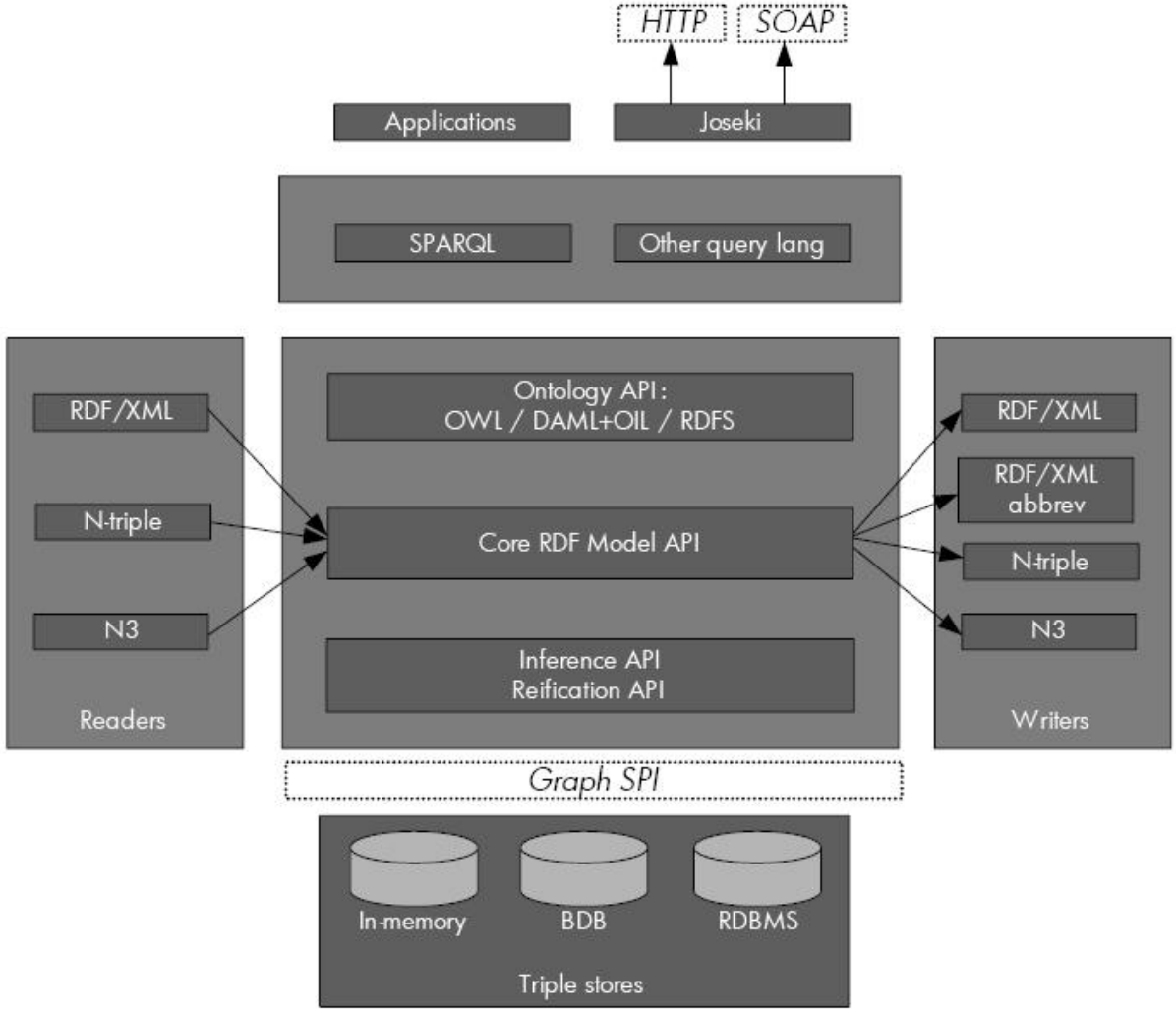


Figure 6-4 Architecture de Jena

6.2.3.4 Corese (Conceptual Resource Search Engine)

Dans ce travail, nous avons aussi utilisé Corese. C'est un moteur de recherche sémantique utilisant comme Jena le langage de requête SPARQL. Il permet de charger des ontologies et effectuer des recherches d'information sur la *base d'annotations* qui utilise des termes définis dans ces ontologies.

Comme le montre la figure 6.5, Corese charge des ontologies au format RDF(S) ainsi que des annotations décrivant des ressources sous forme d'énoncés RDF. Il construit ensuite une représentation interne de ces informations sous forme de graphes conceptuels. Puis réalise

[40] Extrait de Http://www.semanticsupport.org/About_Jena.html

des inférences grâce aux règles et répond aux requêtes de recherche d'information dans la *base d'annotations* en utilisant le langage de requêtes SPARQL.

Pour inférer des résultats, Corese fait la projection d'une requête sur des graphes conceptuels afin d'extraire un graphe conceptuel résultat. Ce graphe résultat est traduit ensuite en un langage standard (RDF(S) ou XML) permettant ainsi sa présentation à l'utilisateur ou sa réutilisation par un autre programme [Luong, 2007].

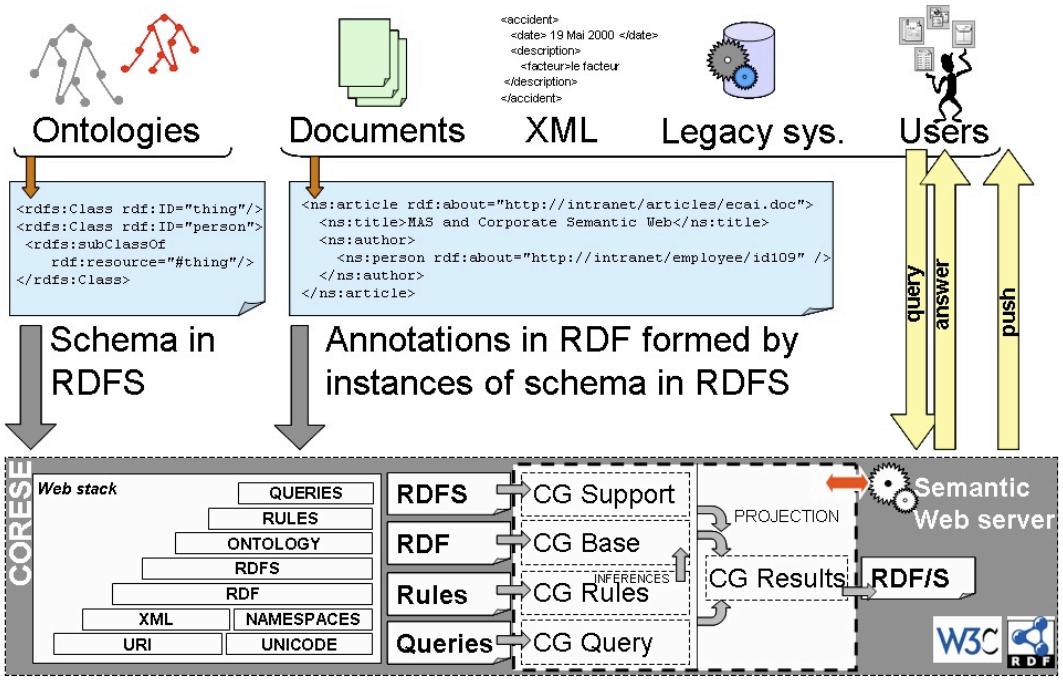


Figure 6-5 Architecture de Corese

6.2.3.5 SeWeSe (Semantic Web Server)

SeWeSe est un serveur Web sémantique permet la création dynamique d'applications Web à partir d'une *base de connaissances* représentées en RDF. Il est basé sur le moteur de recherche Corese et facilite son utilisation au sein d'une application web en proposant un jeu de tags JSP.

L'objectif de la plate-forme SeWeSe est de fournir des primitives et des composants réutilisables, configurables et extensibles afin de réduire le temps passé à développer de nouvelles applications Web sémantique et permettre à ces applications de se concentrer sur leur domaine d'application. Le principal objectif de SeWeSe est l'intégration des opérations web sémantique récurrentes (par exemple, effectuer une requête SPARQL), en technologies Web classique (par exemple, des pages JSP, des servlets) [Durville et Gandon, 2005].

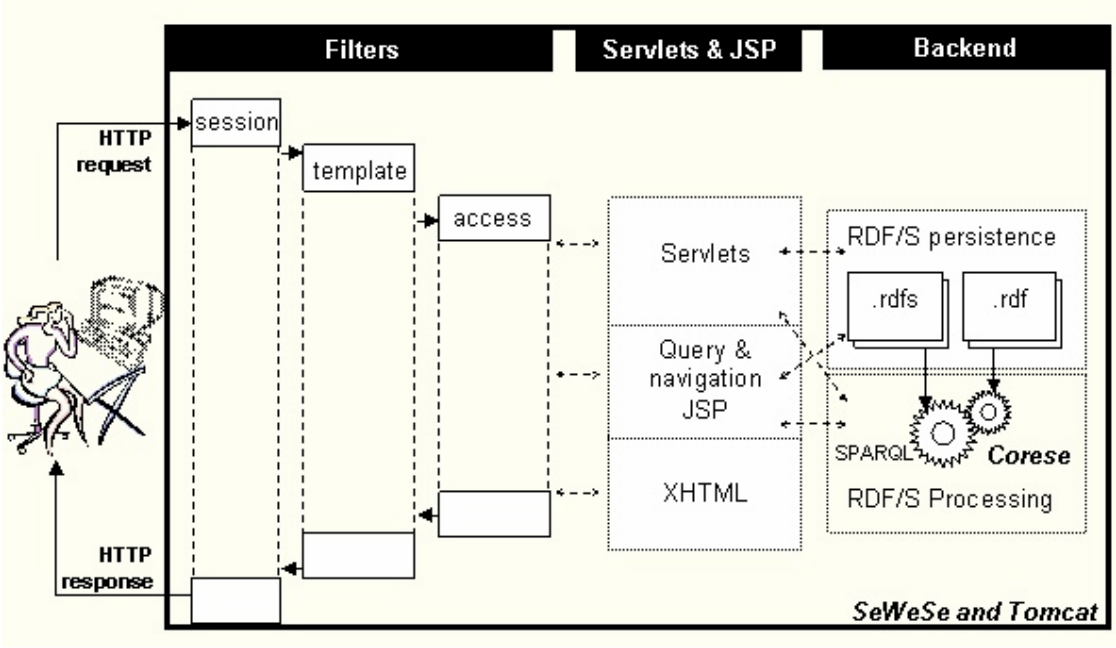


Figure 6-6 Architecture de SeWeSe

Comme le montre la figure 6.6 intègre Corese et est basée sur Tomcat avec une couche supplémentaire autour du moteur de recherche gérant les échanges entre l'application web et les ressources locales. Cette architecture implante une forme de MVC (Model-View-Controler) et se décompose en plusieurs composants réutilisables, notamment [Niang, 2007] :

— Des servlets implantant des tâches récurrentes dans les scénarios d'utilisation d'un serveur web sémantique : résolution d'une requête personnalisée, modification d'une annotation, traitement d'un formulaire, gestion des gabarits des pages, etc.

— Des feuilles de style implantant des transformations XSLT dans un langage déclaratif et donc modifiable par l'utilisateur. Elles sont utilisées pour des tâches de rendu (format d'affichage d'une réponse, génération du JSP d'un formulaire, génération d'une vue de l'ontologie, etc.) et des tâches de modification de fichiers XML (édition d'une annotation, d'une ontologie, etc.).

— Une bibliothèque de Tags JSP permettant de faire appel à Corese au sein d'un page JSP et permettant ainsi de générer des éléments à la volée comme, par exemple, un menu, une liste de requêtes contextuelles ou la présentation d'une sous-partie de l'ontologie. Cette bibliothèque se nomme SemTag et sera réutilisée dans l'implémentation du prototype.

6.2.4 Evaluations

Nous disposons de 39 tables d'échantillons de données *partenaires* comme le montre le tableau 6.2 fournit sous forme de documents XML dans les entrepôts, nous les avons

répartis pour le prototypage dans des dossiers différents selon les sept propriétaires de données identifiés comme les *partenaires*.

Ces tableaux sont issus d'un fichier de départ rassemblant toutes ces tables avec la source du tableau, c'est-à-dire le *partenaire* ayant fait la collecte de ces données. Par *partenaire* encore ici nous précisons que nous faisons référence aux organismes fournisseurs de données. La notion de *partenaire* n'est pas choisie ici au hasard mais reflète aussi la sémantique politique de ce concept sous formes d'entités se mettant ensemble pour coopérer afin d'atteindre un but commun.

Partenaires	**Sources**
Association pour le Développement de la Riziculture en Afrique de l'Ouest	3
Compagnie Sucrière Sénégalaise	1
Direction Régionale du Développement rural de Saint-Louis	17
Inspection Régionale des Services Vétérinaires	2
Sénégal, Pré-Recensement de l'Agriculture	6
Société de Conserves Alimentaires au Sénégal	2
Société d'Aménagement et d'Exploitation des terres du Delta	8

Tableau 6-2 Partenaires et part de chacun dans les échantillons de données

Les thématiques couvertes par les données des *partenaires* sont très variables, nous pouvons les répartir en quatre types de variables avec les données environnementales, sur l'élevage, l'eau avec surtout des données pluviométriques et la pêche même si dans les échantillons actuels il n'y a aucune occurrence de ces dernières.

6.3 Construction des dataweb partenaires

La construction d'un *dataweb partenaire* est effectuée en deux phases. Une première phase permet par la pré-intégration de transformer l'ensemble des données sources initiales sous format XML et une deuxième phase rend possible la restructuration des données. Le rôle de la restructuration consiste dans notre contexte à réorganiser chaque document XML issu de la transformation par une normalisation et une extraction des propriétés spatio-temporelles.

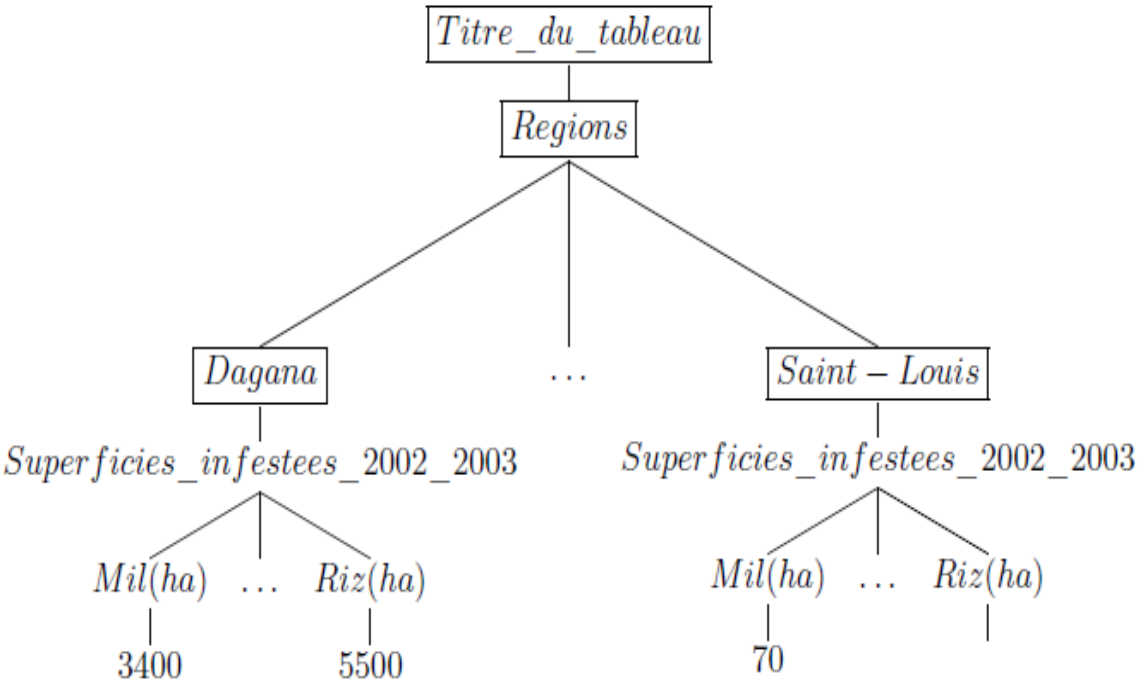


Figure 6-7 Structure XML résultant de la transformation du tableau 6.1

6.3.1 Extraction et transformation des données sources

L'extraction et la transformation des données sources visent à extraire les données de leur forme tabulaire initiale pour les transformer en XML. Les données considérées sont généralement stockées sous forme de table de données résultant de l'évolution des observations sur le terrain des caractéristiques d'une population d'individus dont les propriétés en commun sont les objets cible d'une étude.

Les formes susmentionnées ont été construites et remplies par des experts de leur domaine. Donc les noms de colonnes et titre de chaque tableau véhiculent des termes du vocabulaire *partenaire*, propre au domaine et de manière implicite les relations entre ces dernières. XML convient de façon particulière à la représentation de ces données où chaque case sera transformée en nœud du document et chaque ligne du tableau un sous-arbre à part entière.

Dans le tableau 6.1 où l'on s'intéresse aux superficies infectées, « région » et « superficies infectées » deviennent les variables privilégiées. Ainsi « Dagana », « Podor » et « Saint-Louis » deviennent des concepts construits à partir de la variable privilégiée « régions » qui qualifient leur nature. Par conséquent dans le cas du tableau général 5.1 la variable privilégiée est C, et les individus I ne constituent que des constructions ou particularités de C.

```
<?xml version="1.0" encoding="UTF-8" ?>
<Titre_du_tableau_de_donnees>
        < Regions>
                <Dagana>
                        <Superficies_infestees_2002_2003>
                                <Mil(ha)>3400</Mil(ha)>
                                ...
                                <Riz(ha)>5500</Riz(ha)>
                </Dagana>
                        ...
                <Podor>
                        <Superficies_infestees_2002_2003>
                                <Mil(ha)>2700</Mil(ha)>
                                ...
                                <Riz(ha)>2500</Riz(ha)>
                </Podor>
                        ...
                <Saint - Louis>
                        <Superficies_infestees_2002_2003>
                                <Mil(ha)>70</Mil(ha)>
                                ...
                                <Riz(ha)></Riz(ha)>
                </Saint - Louis>
                </Regions>
<Titre_du_tableau_de_donnees>
```

6.3.2 Restructuration des données sources

La restructuration est une étape ciblant les données des entrepôts *partenaires*. Elle vise comme discuté dans le chapitre précédent à reformater les données et essentiellement mettre en exergue les caractéristiques spatio-temporelles.

Du point de vue restructuration spatiale on se sert du dictionnaire des noms de localités afin d'extraire les noms et du point de vue temporelle, rechercher les occurrences temporelles.

Dans certains tableaux des fichiers de départ, on retrouve en effet certaines occurrences de caractère comme *« % »* qui ne peuvent passer le validateur XML. Nous identifions également l'occurrence des unités de mesure sous différentes forme identifiable en étudiant les labels *« f.kg_ha »,« t_ha »,« f.cfa_kg »,« f.cfa »,« kg_ha »,« t »,« ha »,« kg ».*

Pour l'identification des caractéristiques spatio-temporelles dans les documents XML nous procédons à une simple recherche sous forme de « matching » dans les noms de balise pour rechercher l'occurrence de certains mots avant d'insérer un nouveau nœud fils pour celui dans lequel s'est fait l'extraction.

Dans le contexte du *SIC*, les occurrences temporelles recherchées sont sous la forme de saisons qui s'étalent sur deux ou plusieurs années sous forme de saisons agricoles. Nous avons étudié les données pour pouvoir mettre en place une heuristique permettant d'identifier ces caractéristiques temporelles dans les noms de balises XML des données du *SIC*. Il apparaît en effet que les saisons sont représentées par les *partenaires* dans les échantillons de données que nous avons en notre possession sous la forme « debut » _« fin » soit la date début de la saison et la date de fin séparées par un « underscore ». Pour chaque document XML, ce traitement est effectué sur tous ses composants en commençant par le titre du tableau.

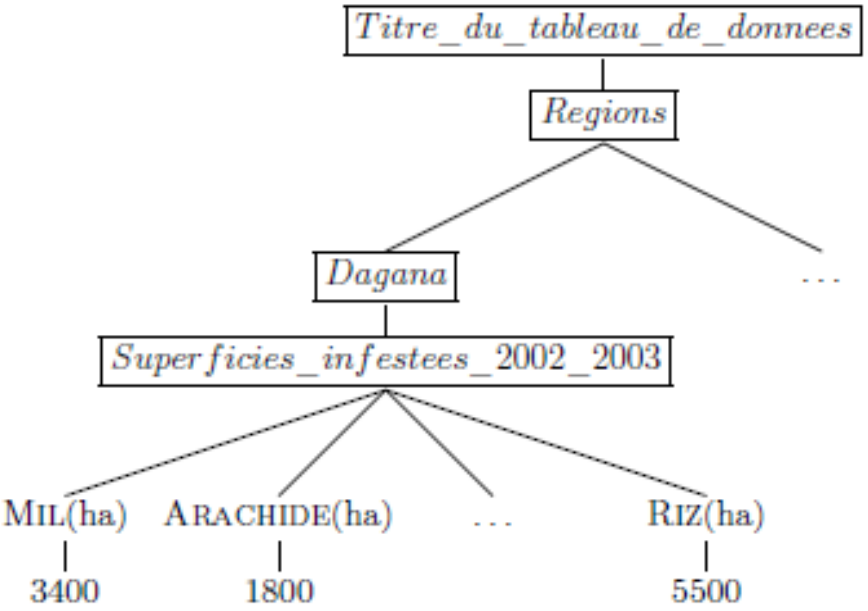


Figure 6-8 Une vue sur la structure XML résultant de l'XMLisation du tableau 6.1

Ainsi donc, chaque document issu de la pré-intégration est passé dans le module de réorganisation. En sortie nous avons un document qui répond aux normes nécessaires pour faciliter une interrogation intégrant les dimensions spatiales et temporelles.

De l'exemple du tableau 6.1 nous pouvons constater l'occurrence de plusieurs unités spatiales et mesure de surface que sont « Dagana », « Podor » et « Saint-Louis » et une unité temporelle qu'est la saison « 2003_2004 » est extraite de la propriété « Superficies infestées(2002/2003) » et « (ha) » ou « hectare » que l'on peut extraire des attributs.

6.3.2.1 Restructuration ciblant les caractéristiques spatiales

L'intégration et la combinaison des données devant prendre en compte la dimension spatiale des données, détecter la provenance géographique d'une donnée est importante. Pour la restructuration spatiale, nous procédons comme déjà spécifié dans le chapitre précédent. Il

suffit de se baser sur le dictionnaire fourni par l'expert du domaine afin d'identifier la localité comme simple entité géographique et dans le cas échéant, créer un nœud XML. Par exemple si le label correspond exactement à la caractéristique recherchée comme c'est le cas d'une des localités énumérées par l'expert dans le tableau 6.1, alors le nœud sera complètement restructuré. Il sera remplacé par un nœud ayant comme label « localite » avec un attribut « nom » ayant comme valeur le nom de la localité trouvée. Dans le cas où la caractéristique trouvée est incluse dans son label, alors un attribut « localite » avec comme valeur l'occurrence spatiale lui est ajoutée, et sera rajouté à son nœud. La figure 5.5 donne la forme générale de la restructuration que nous proposons.

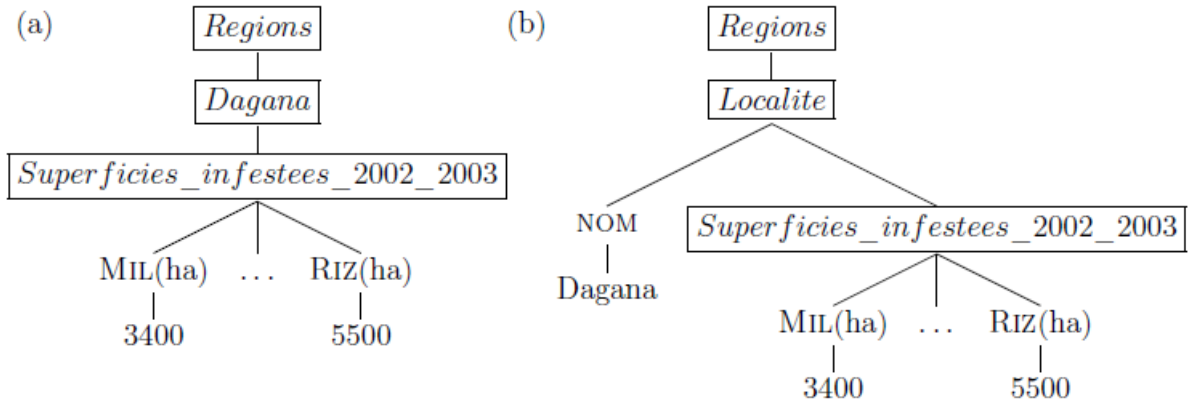


Figure 6-9 Exemple de restructuration ciblant les caractéristiques spatiales

C'est le cas du nœud « Dagana » dans la figure 6.9. Il est restructuré et devient la valeur de l'attribut nom d'un nouveau nœud « localite ». L'item (a) montre le nœud original et le (b) le nouveau nœud obtenu après restructuration spatiale.

6.3.2.2 Restructuration des caractéristiques temporelles

La restructuration temporelle peut être effectuée de manière semi-automatique comme c'est le cas dans notre contexte, il suffira d'étudier les données afin d'identifier si les occurrences temporelles respecte un ensemble de format standard. Si c'est le cas, alors la mise en place d'une heuristique permet de restructurer les données de manière automatique. Dans le cas contraire, il faudra confier la tâche à l'expert.

Dans notre contexte, une étude des données a permis de révéler qu'elles sont exprimées sous la forme de saison agricole. Comme déjà spécifié dans le chapitre précédent il suffit d'insérer un nouveau nœud nommé saison avec comme attributs les années de début et de fin de la saison à laquelle les données ont été collectées. Il convient d'attribuer au nœud

cible un nouveau fils *saison_agricole* avec un attribut *debut* et un attribut *fin* qui auront comme valeur la date de début de la saison et la date de fin.

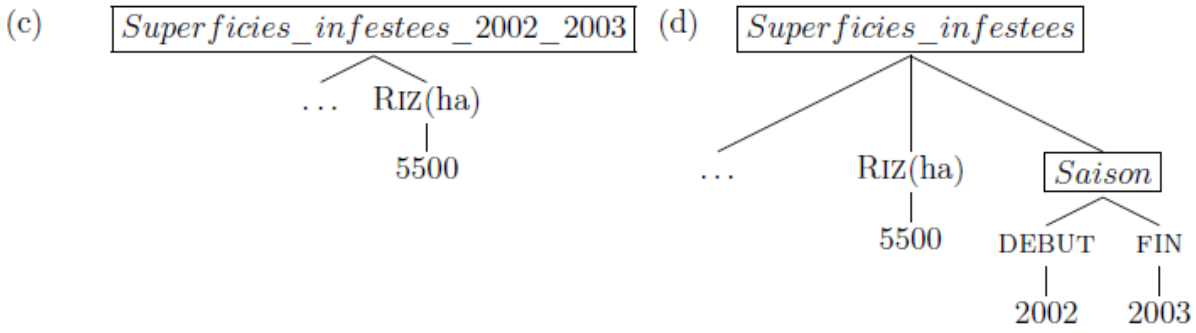


Figure 6-10 Exemple de restructuration ciblant les caractéristiques temporelles

Prenons l'exemple de Superficies_infestees_2002_2003, la période _2002_2003 sera extraite pour constituer le nouveau nœud fils et l'ancien nœud va porter comme label restructuré la chaîne de caractères *Superficies_infestees*. Dans la figure 6.10 l'item (a) montre le nœud original et le (b) le nouveau nœud obtenu après restructuration spatiale.

6.3.2.3 Restructuration ciblant les unités de mesure

Les observations environnementales de caractéristiques d'individus s'accompagne le plus souvent d'unité servant en en mesurer la grandeur. Dans la démarche de l'intégration sémantique des données, il est essentielle de les identifier et ainsi d'extraire l'unité de mesure du label du nœud cible et lui rajouter deux attributs que sont la valeur et l'unité.

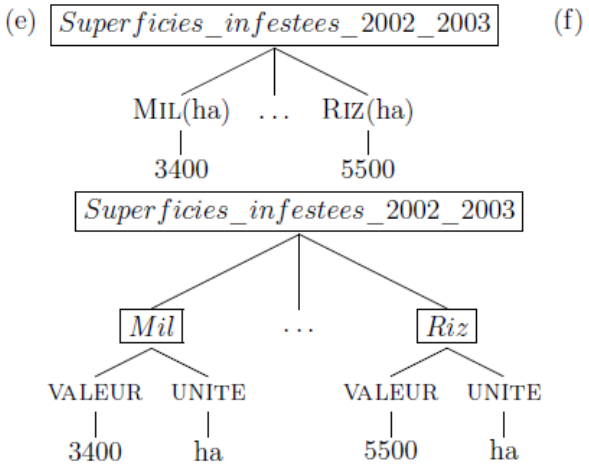


Figure 6-11 Exemple de restructuration ciblant les unités de mesure

Prenons l'exemple du nœud « Mil(ha) » et de « Riz(ha) » où l'on peut détecter l'occurrence de l'unité de mesure de surface « hectare » sous la forme « (ha) ».

Après restructuration, nous obtenons le document suivant :

```
<?xml version="1.0" encoding="UTF-8" ?>
<superficie_infestees_source_DRDRSL>
        <region>
                <localite nom="dagana">
                        <superficies_infestees>
                                <mil valeur="3400" unite_de_mesure="ha"/>
                                <arachide valeur="1800" unite_de_mesure="ha"/>
                                <niebe valeur="100" unite_de_mesure="ha" />
                                <saison_releve_superficies date_Deb="2002" date_Fin="2003"/>
                        </superficies_infestees>
                </localite>
        </region>
        <region>
                <localite nom="podor">
                        <superficies_infestees>
                                <mil valeur="2700" unite_de_mesure="ha"/>
                                <arachide valeur="600" unite_de_mesure="ha"/>
                                <niebe valeur="1800" unite_de_mesure="ha"/>
                                <saison_releve_superficies date_Deb="2002" date_Fin="2003"/>
                        </superficies_infestees>
                </localite>
        </region>
        <region>
                <localite nom="saint_louis">
                        <superficies_infestees>
                                <mil valeur="70" unite_de_mesure="ha"/>
                                <arachide valeur="5500" unite_de_mesure="ha"/>
                                <niebe valeur="3500" unite_de_mesure="ha"/>
                                <saison_releve_superficies date_Deb="2002" date_Fin="2003"/>
                        </superficies_infestees>
                </localite>
        </region>
</superficie_infestees_source_DRDRSL>
```

L'occurrence identifiée de l'unité de mesure des superficies infestées permet de restructurer les nœuds comme Mil(ha) pour lui rajouter deux attributs supplémentaires que sont l'unité de mesure en hectare et la valeur du relevé. L'identification de Dagana comme unité spatiale permet de restructurer le nœud initial.

Globalement, à partir de la structure arborescente fournie en début de cette section, la forme suivante, grâce aux informations fournies par l'expert et aux heuristiques mises en place.

Dans cette phase de restructuration de documents nous utilisons uniquement des technologies associées à XML comme DOM[41] et SAX[42]. Elles permettent d'effectuer entièrement les manipulations. Il faut aussi noter que tous les documents sont bien parsés.

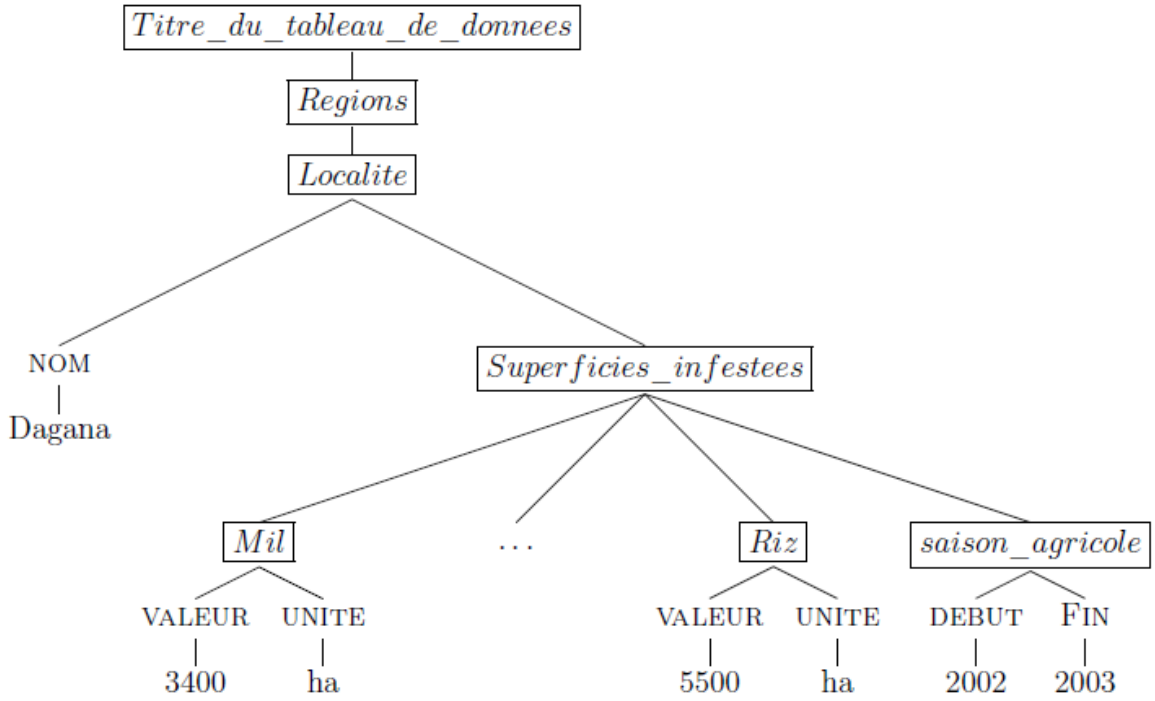


Figure 6-12 Une vue de la structure résultant de la restructuration de la figure 6.8

6.4 Construction des bases de connaissances

Dans cette section, nous allons présenter le processus de construction des bases d'annotations *partenaires*. Une *base de connaissances* est constituée par l'ontologie du *partenaire*, son *ontologie générique* et sa *base d'annotations*. Nous allons donc présenter le processus permettant de générer semi-automatiquement l'ontologie *partenaire* OWL, puis son *ontologie générique* et permettant ainsi en collaboration avec les autres organismes participants la construction de l'*ontologie globale*.

6.4.1 Processus d'extraction des ontologies partenaires

Cette phase est celle qui suit la phase de restructuration, elle permet comme exposé dans la démarche d'associer une sémantique claire à chaque entrepôt de document XML.

[41] Http://www.w3.org/DOM/
[42] Http://www.saxproject.org/

Cette approche permet d'extraire automatiquement chaque concept candidat, ses attributs et composants ainsi que les relations de cardinalité entre les concepts et leurs attributs. La réutilisation d'une ontologie plus générique couvrant le domaine comme *AOS* de la FAO dans notre contexte permet de rajouter des relations sémantiques telles que la subsomption.

Supposons disposer d'un seul tableau de données pour la DRDR, alors l'ontologie sera constituée par les seuls concepts et attributs pouvant être extraits de la structure XML issue du tableau de données 6.1. Nous avons alors huit concepts, une seule relation héritée d'*AOS* et cinq attributs.

Ainsi le modèle simple d'ontologie *partenaire* O_{part} dans ce contexte sera constitué par le couplet $(S_{part}, \mathcal{L}_{part})$. La structure de l'ontologie comme décrite dans le chapitre précédent est l'octuple :

$$S_{part} := \{C_{part}, \mathcal{R}_{part}, \mathcal{A}_{part}, \mathcal{T}_{part}, C_{agreg}, \mathcal{H}_{part}{}^{C}, \sigma_{\mathcal{R}}, \sigma_{\mathcal{A}}\} \text{ où :}$$

- $C=\{c1, c2, c3, c4, c5, c6, c7, c8\}$;
- $\mathcal{R}_{part}=\{r\_261\}$;
- $\mathcal{A}_{part}=\{$'nom','unite','valeur','date_Deb','date_Fin'$\}$;
- $\mathcal{T}_{part}=\{String, Integer\}$;
- $C_{agreg}=\{agrovoc\#c\_6701, agrovoc\#c\_4236, agrovoc\#c\_8326, agrovoc\#c\_11368, agrovoc\#c\_1938, agrovoc\#c\_7536, agrovoc\#c\_6911\}$;
- $\mathcal{H}_{part}{}^{C}=\{(c1, \{agrovoc\#c\_6701, agrovoc\#c\_4236, agrovoc\#c\_8326\ \}), (c6, agrovoc\#c\_11368), (c7, agrovoc\#c\_1938), (c6, agrovoc\#c\_11368), (c1, \{agrovoc\#c\_7536, agrovoc\#c\_6911\})\}$;
- $\sigma_{\mathcal{R}}=\{r\_261(c1, c\_2), r\_261(c2, c\_3), r\_261(c3, c\_4), r\_261(c4, \{c\_5, c\_6, c\_7, c\_8\ \})\}$;
- $\sigma_{\mathcal{A}}=\{nom(c3, String), unite(\ \{c\_5, c\_6, c\_7\ \}, String), valeur(\ \{c\_5, c\_6, c\_7\ \}, Integer), date\_Deb(c8, Integer), date\_Fin(c8, Integer)\ \}$.

Soit $\mathcal{L} := \{\mathcal{L}^{C}, \mathcal{L}^{\mathcal{R}}, F, G\}$ le lexique associé tel que :

- $\mathcal{L}^{C}=\{$'superficie_infestees_source_DRDRSL','region','localite','region','mil', 'arachide','niebe','saison'$\}$;
- $\mathcal{L}^{\mathcal{R}}=\{$'est_un'$\}$;
- F('superficie_infestees_source_DRDRSL')=c1, F('region')=c2, F('localite')=c3; F('superficies_infestees')=c4, F('mil')=c5, F('arachide')=c6, F('niebe')=c7, F('saison')=c8.

Le graphe sous OWL représentant l'ontologie dérivée est la suivante :

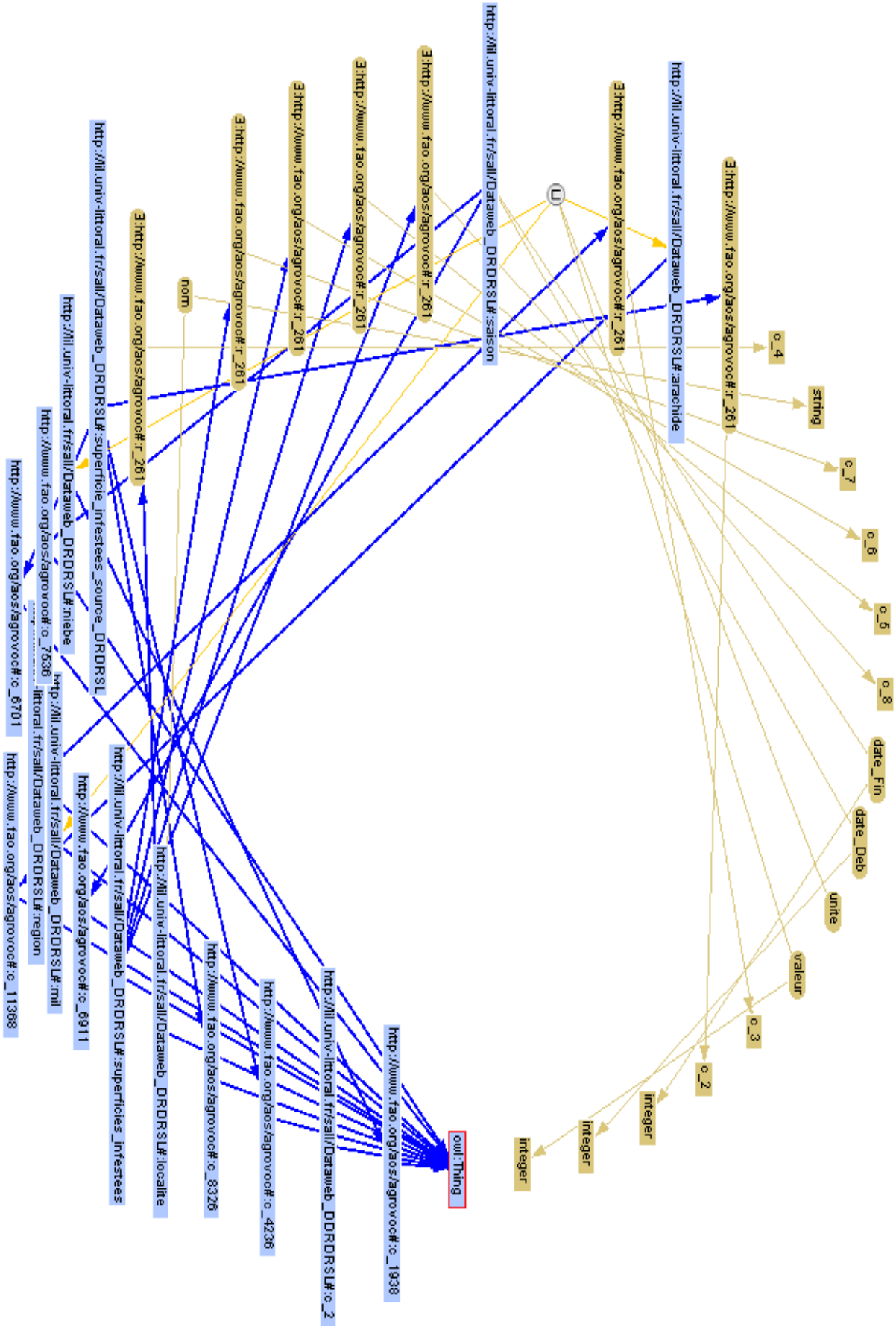


Figure 6-13 Structure graphique de l'ontologie de la DRDR-SL

Le graphe de concept résultant est le suivant :

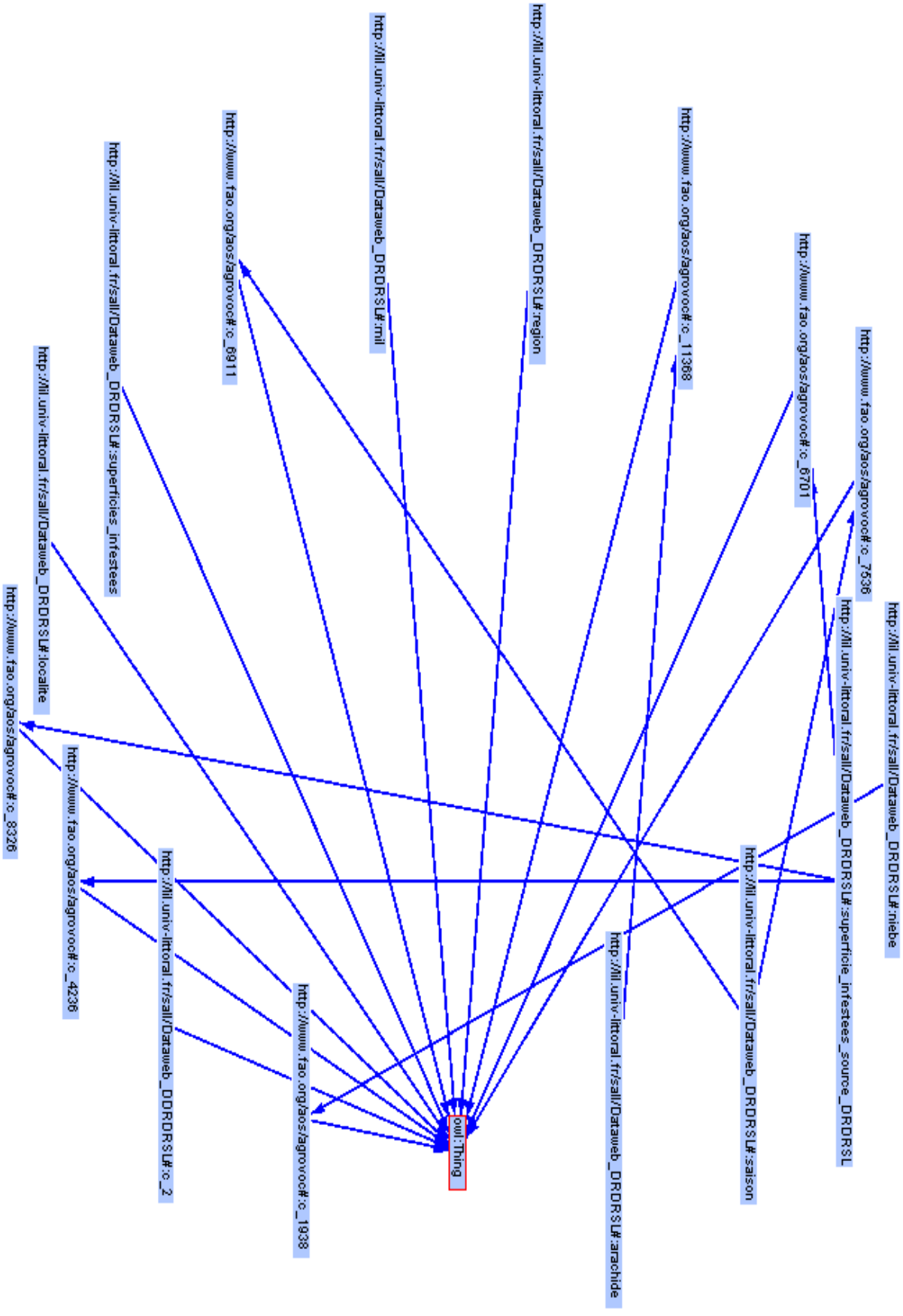


Figure 6-14 Graphe de concepts de l'ontologie de la DRDR-SL

En appliquant la technique de subsomption des concepts extraits basé sur le critère de l'homonymie au sens morphosyntaxique, nous recherchons par un système de correspondances le mot ayant la même syntaxe dans *AOS*. Dans le code OWL de l'ontologie précédemment décrite, ces relations sont exprimées via *« RDFs:subClassOf »*.

Dans le modèle à base lexical les différentes relations exprimées via celle de subsomption $\mathcal{H}_{part}{}^{C}$={(c1, {agrovoc#c_6701, agrovoc#c_4236, agrovoc#c_8326}), (c6, agrovoc#c_11368), (c7, agrovoc#c_1938), (c6, agrovoc#c_11368), (c1,{agrovoc#c_7536, agrovoc#c_6911})}. La figure 6.13 montre un exemple de subsomption où le concept candidat « Espèces Animales » subsumer au concept nommé « animal » plus générique.

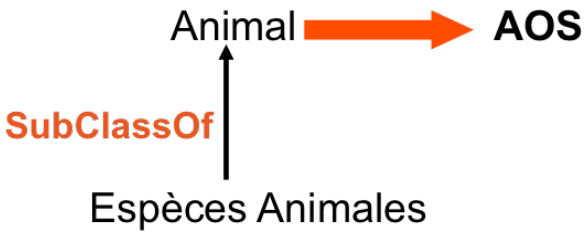


Figure 6-15 Illustration d'une subsomption à un concept générique

Cette subsomption va nous permettre après la construction des ontologies de les enrichir avec l'extraction de relations sémantiques supplémentaires entre les subsumeurs. Deux types de relations comme déjà spécifiées sont susceptibles d'être extraites de l'architecture des documents XML : celles de compositions et celles d'attributs.

La première est celle automatique de compositions via la relation r_261 de l'ontologie *AOS*. La relation de composition n'est certes pas très explicite, mais elle permet par sa flexibilité de laisser une marge à l'expert du domaine afin de mieux la préciser si nécessaire.

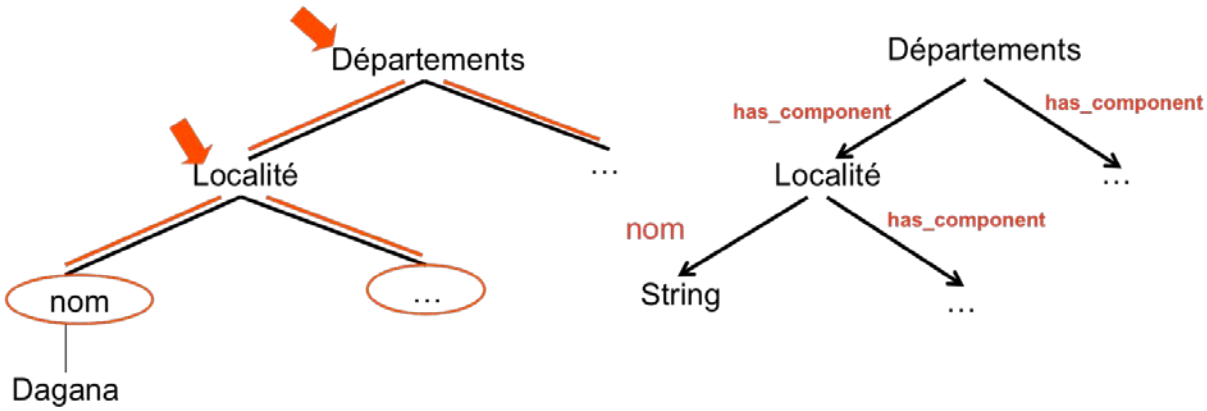


Figure 6-16 Illustration d'une extraction de relations type r_261 d'un document XML

La deuxième est constituée par les relations de type attribut permettant dans l'exemple de la figure 6.14 d'ajouter l'attribut « nom » de type « string » au concept « localité ». Elle est

représentée sous OWL grâce au nœud spécifiant la propriété sous la forme « *owl:DatatypeProperty*».

La figure 6.14 montre deux structures XML dont celle d'une partie d'un document XML où le nœud *département* a deux nœuds fils dont un nommé localité.

```
<owl :ObjectProperty rdf :about="Http ://www.fao.org/aos/agrovoc#r_261">
        <rdfs :label xml :lang="en">component</rdfs :label>
        <rdfs :comment xml :lang="en">&lt ;component&gt ;</rdfs :comment>
        <rdfs :comment xml :lang="en">Y &lt ;component&gt ; X. An object X that is a part
          of a whole Y and has also an existence independently from Y. E.g. "engine"
          &lt ;component&gt ; "engine part" ; "tree" &lt ;component&gt ; "leaf" ;
          "cell" &lt ;component&gt ; "chromosome" ; but NOT "blood" &lt ;component&gt ;
          "blood cell" (see &lt ;composed_of &gt ;) ;
        </rdfs :comment>
        <rdfs :subPropertyOf rdf :resource="Http ://www.fao.org/aos/agrovoc#r_111"/>
        <owl :inverseOf>
                <owl :ObjectProperty rdf :about="Http ://www.fao.org/aos/agrovoc#r_260"/>
        </owl :inverseOf>
</owl :ObjectProperty>
```

Figure 6-17 Définition OWL de la relation « r_261 » dans *AOS*

Cette structure permet alors de créer des relations de composition entre les niveaux XML reflétant ainsi la relation de composition. L'encadré ci-dessous montre la définition OWL de cette relation dans l'ontologie *AOS*. Une relecture d'un expert du domaine permet de valider les relations établies, même si la composition est fiable en se basant sur l'architecture des tableaux de départ.

Il est aussi possible d'exploiter la relation *r_90* de *AOS* très neutre, mais qu'un expert peut aider à rendre plus explicite. Pour chaque subsumeur issu d'*AOS*, il existe un ensemble de concepts auxquels il est lié par cette relation. C'est le cas du concept « Abattoir » dans *AOS* et celui nommé « Hygiène de la viande », mais une telle relation entre ces deux concepts peut difficilement être typée automatiquement.

Nous avons ainsi pu extraire au total (217) concepts candidats sans doublons (364 si on considère l'ensemble des concepts de toutes les ontologies, les doublons y compris) et subsumer (111 d'entre eux à des concepts d'*AOS* sans intervention de l'expert du domaine. L'intervention de l'expert du domaine permet d'améliorer considérablement ce chiffre.

6.4.2 Construction des bases d'annotations

La figure 6.15 montre l'extrait d'une *base d'annotations* d'un *partenaire*. Son utilisation a pour but de concrétiser la relation existante entre le niveau des connaissances et

celle des données. Ainsi, à chaque concept extrait est associé le chemin XPATH indiquant l'endroit d'où il a pu être extrait.

```
<rdf :RDF(...) >
        (...)
<c_4 rdf :ID="superficies_infestees">
        <urlLocalSource>
                file :/.../Superficies infestées.xml/superficie_infestees_source_DRDRSL/
                loccalite/superficies_infestees
        </urlLocalSource>
</c_4>
        (...)
<c_7 rdf :ID="niebe">
        <urlLocalSource>
                file :/.../Superficies infestées.xml/superficie_infestees_source_DRDRSL/
                loccalite/superficies_infestees/niebe
        </urlLocalSource>
</c_7>
        (...)
</rdf :RDF>
```

Figure 6-18 Extrait de base d'annotations d'un partenaire

Dans l'exemple de la figure 6.15 nous pouvons voir les deux concepts c4 et c7 de l'ontologie *partenaire* et les chemins indiquant leur source d'origine. Les bases d'annotations sont représentées en RDF facilitant ainsi leurs réutilisations.

6.4.3 Construction des ontologies génériques

La figure 6.16 montre l'extrait d'une *ontologie générique* d'un *partenaire*. Sa taxonomie n'est constituée que par des concepts ayant au moins un subsumeur dans l'ontologie *AOS*, et par conséquent, font partie de l'*ontologie générique*. Les concepts de l'ontologie *partenaire* n'ayant pas trouvé de subsumeur ne pourront pas être généralisés dans l'ontologie *AOS* et ne figurent donc pas dans l'*ontologie générique*.

L'extrait montré par la figure 6.16 est celui de l'*ontologie générique* de la SAED. Le *partenaire* a aussi la possibilité de choisir de retirer des concepts de l'*ontologie générique* qu'il ne désire pas partager afin de les rendre privés par exemple. Cela permet de gérer la problématique de l'aspect privé de certaines données. Les ontologies génériques vont servir de base à la construction de l'otologie globale comme déjà expliqué.

```
<rdf :RDF (...) >
        (...)
<owl :Class rdf :about="Http ://lil.univ littoral.fr/ sall/Dataweb_SAED#c_4">
        <rdfs :label xml :lang="FR"> superficies_infestees </rdfs :label>
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_6178"/>
</owl :Class>
<owl :Class rdf :about="Http://lil.univ-littoral.fr/ sall/Dataweb_SAED#c_12">
        <rdfs :label xml :lang="FR">tomate</rdfs :label>
        <rdfs :subClassOf rdf :resource="Http ://www.fao.org/aos/agrovoc#c_7805"/>
</owl :Class>
        (...)
</rdf :RDF>
```

Figure 6-19 Extrait d'une ontologie générique d'un partenaire

6.4.4 Construction de l'*ontologie globale*

L'*ontologie globale* est construite grâce à une fusion des ontologies génériques des *partenaires* en utilisant le moteur de recherche sémantique Corese(COnceptual REsource Search Engine)[Corby et al, 2004].

Comme nous l'avons décris dans le chapitre précédent, l'*ontologie globale* sert de point d'entrée et d'outil pour le partage des connaissances que tous les *partenaires* du projet d'intégration en commun désirent partager. Elle est donc constituée par une intersection des ontologies génériques. Du point de vue contexte applicatif, nous réutilisons les possibilités offertes par Corese et permettant sa construction automatique.

Notre approche présente l'avantage d'éviter le processus d'alignement des ontologies génériques deux à deux comme le font la plus part des approches pour construire l'*ontologie globale*. L'utilisation de l'ontologie de référence *AOS* permet de réaliser un alignement global avec le moteur Corese. En plus, Corese contient un moteur d'inférence qui permet de ne pas se préoccuper de la déduction de nouvelles connaissances à partir des ontologies génériques.

6.5 Une médiation utilisant un système à base de hubs

L'architecture du système d'intégration requiert une structuration permettant à chaque donnée de jouer à la fois le rôle d'un client comme celui d'un serveur. La solution des systèmes à base de hubs (serveurs pairs) est l'une des plus appropriées, chaque hub pouvant localement traiter les requêtes grâce à l'ontologie *partenaire* ou au besoin envoyer des requêtes aux autres hubs et recombiner les requêtes en retour via l'*ontologie globale*.

Dans cette section, nous allons étudier l'architecture conceptuelle du système à base de hubs ainsi que l'adaptation du système proposé dans [Gandon et al., 2008] au contexte du projet *SIC*-Sénégal.

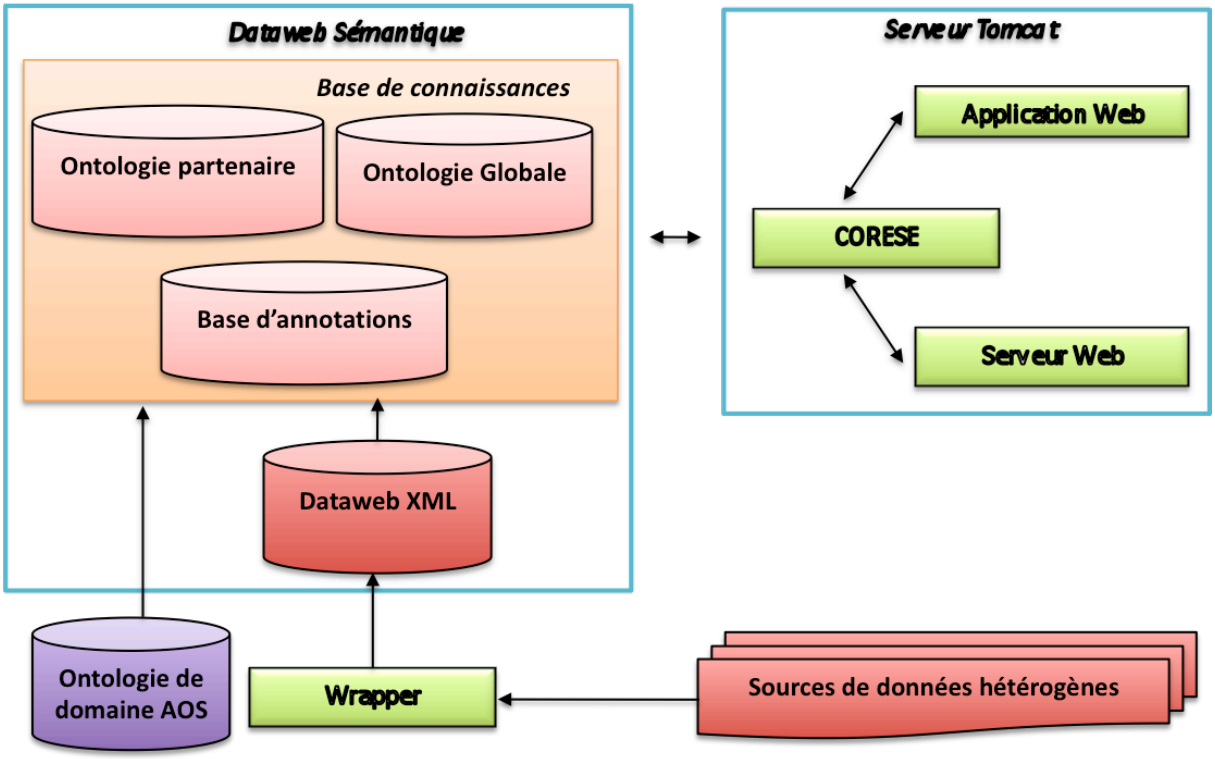


Figure 6-20 Architecture conceptuelle d'un Hub

La figure 6.17 présente l'architecture conceptuelle d'un hub. Chaque serveur peut être sollicité pour la résolution d'une requête.

L'architecture du serveur web sémantique (SWS) intégrant Corese et basée sur Tomcat est celle d'une couche supplémentaire autour du moteur de recherche gérant les échanges avec le web et avec les ressources locales sur les graphes conceptuels [Corby et al, 2004].

6.6 Validation des ontologies

Il existe deux manières de valider les ontologies : celle logique et celle par l'utilisation. Pour la validation logique, nous avons procédé deux fois la validation de chaque ontologie construite. Une première validation est effectuée avec Jena et une deuxième en utilisant Corese. Les ontologies construites passent toute cette étape de la validation avec succès.

La figure 6.18 par exemple, montre la trace du chargement avec succès de l'ontologie de l'ADRAO. Ce chargement tient lieu de validation logique, puisque Corese effectue la

vérification de la cohérence de l'ontologie en la chargeant. Si sa structure logique n'est pas cohérente, elle ne sera pas chargée.

Figure 6-21 Trace du chargement de l'ontologie de l'ADRAO sous Corese

6.7 Conclusion

Une nouvelle vision du web sémantique s'impose actuellement, celle où l'on passe de serveurs de documents avec des liens hypertextes à celle de serveurs où les liens entre documents restent mais avec une mise en place de liens entre les données contenues dans les documents. Cela nécessite des approches pour l'ajout de sémantique à ces documents et dans un cadre plus global une intégration des données et des serveurs de données.

Dans le contexte applicatif du projet *SIC*-Sénégal, nous avons présenté une démarche flexible permettant à partir d'un ensemble de documents produits par les experts d'un domaine et l'existence d'une ontologie standard décrivant le domaine de générer une *base de connaissances* conceptualisant le vocabulaire du fournisseur de données et faisant du coup le lien entre la couche de données et celle des connaissances.

Ensuite par un mécanisme de construction récursif des ontologies génériques initiales nous produisons une *ontologie globale* générique dupliquée chez tous les organismes participants. Ainsi l'utilisation d'un système à base de hubs et d'un moteur de recherche sémantique dans un hub du système d'intégration permet à un utilisateur d'interroger de manière intégrée les *dataweb sémantique*s via l'ontologie partagée.

L'application de notre approche permet, d'une part, de résoudre la problématique de l'hétérogénéité structurelle par la construction d'entrepôts de documents XML (ou *dataweb*) pour chaque *partenaire* et, d'autre part, l'association d'une couche sémantique avec une ontologie OWL à partir du vocabulaire contrôlé décrivant les données de chaque *partenaire*.

D'un autre côté, dans la perspective web sémantique, nous avons apporté notre brique dans la conceptualisation de système d'intégration ouvert sur le web, nécessitant la mise en place d'un outil pour combiner leurs données. La grande difficulté dans la philosophie web

sémantique comme nous l'avons étudié dans l'état de l'art c'est de parvenir à ajouter une couche sémantique au web ou sur des sources de données ne disposant pas d'une telle description. Il est donc nécessaire d'une part de proposer un processus de sémantisation des sources de données et des équivalences entre les niveaux sémantiques. D'autre part il convient de sécuriser les connaissances mises en ligne surtout dans le contexte d'entreprises désirant partager leurs données. Cet aspect est prise en compte par le rôle joué par les ontologies génériques qui permettront à priori de contenir l'ensemble des connaissances partagées par les *partenaires* et supposées être constituées par les concepts ayant des subsumeurs dans l'ontologie de domaine de référence.

Le mécanisme d'échange des ontologies génériques *partenaires* va permettre de construire de manière collaborative l'*ontologie globale* constituée au finish par l'intersection des ontologies génériques. Cette constitution collaborative de l'*ontologie globale* se fait directement par l'utilisation des moteurs de recherche sémantiques Corese.

Dans un billet intitulé *« Semantic Web Doesn't Have to Be Difficult »*[43], Seth Ladd propose de renommer le web sémantique en « Data Web » en se basant sur la proposition de Tim Berners-Lee afin d'éviter la polysémie et les contresens possibles du terme « sémantique». Il termine son billet avec la phrase ***« Repeat after me: Data Web, Data Web, Data Web. Put my data on the web. Give it a URI. Create a Web of Data »*** pouvant bien conclure ce présente chapitre, mieux, notre approche adopte la même philosophie sous forme de « *Dataweb sémantique* ». Cette approche sous-entendant que web et données sont indissociables, que le web de données ne serait sans la sémantique d'une part des données mais aussi celles des différentes composantes du point de vue ensembliste du web de données.

Le web sémantique dans notre approche est un ensemble de *dataweb sémantique*s où les données sont interconnectées via des ontologies et la prise en compte de l'aspect organisationnelle permet à chacun de publier ses données et d'obtenir la description sémantique et la mise en commun avec l'existant. A l'image des informations dans les réseaux sociaux, il est important de faire la part entre la partie privée et celle publique, cet aspect est pris en compte par la taxonomie des connaissances décrites.

Nous distinguons une connaissance localement valable, celle générique pour le besoin d'ouverture et de coopération avec les autres et celle globale pour l'interconnexion entre sources. L'ontologie locale *partenaire* pouvant ainsi localement être utilisée pour l'accès à

[43] http://blog.semergence.com/2007/09/20/semantic-web-doesn't-have-to-be-difficult/

l'ensemble des données, celle générique servant à constituer le savoir que l'on déclare vouloir partager et celle globale que tout le monde accepte de partager.

Dans le chapitre suivant, nous allons étudier le prototype AIDE-ISH implantant l'approche d'intégration.

Chapitre 7

Le prototype AIDE-ISH

Sommaire

7.1 Introduction 179
7.2 Manager un *partenaire* 180
7.3 Générer un *dataweb sémantique partenaire* 180
7.4 Visualiser la *base d'annotations* d'un *partenaire* 181
7.5 Envoyer une requête *partenaire* ou la distribuer 181
7.6 Conclusion 182

7.1 Introduction

Nous avons développé un prototype nommé AIDE-ISH (Atelier d'Intégration de Données Environnementales Issues de Source Hétérogènes), pour implémenter notre approche. C'est un système à base de hub qui permet aux *partenaires* du projet *SIC-Sénégal* de partager leurs sources de données environnementales. Pour participer au projet, un *partenaire* doit installer un hub.

C'est le moteur de recherche sémantique Corese [Corby et al, 2004] que nous utilisons pour interroger la *base de connaissances*. Corese charge des ontologies au format RDF(S) / OWL et des annotations décrivant des ressources sous forme d'énoncés RDF. Il construit une représentation interne de ces informations sous forme de graphes conceptuels et réalise des inférences grâce aux règles et répond aux requêtes de recherche d'information dans la *base d'annotations*.

Pour faire les tests, nous avons créé 3 hubs virtuels (serveurs fonctionnant sur une même machine), correspondants respectivement aux *partenaires* SAED, SOCAS et ADRAO. Sur chaque hub nous déployons :

— Les services web nécessaires pour son fonctionnement, l'environnement d'exécution et de déploiement des services web que nous utilisons est l'outil axis. ;

— Une application web générique, fournissant les interfaces accessibles aux utilisateurs ;

— Les données initiales du *partenaire*, stockées dans des documents Excel.

L'application web générique, utilise Sewese[Durville et Gandon, 2005] et la JSTL. Elle est déployée sous Tomcat et fournit un certain nombre de fonctionnalités que nous présentons dans la suite.

7.2 Manager un partenaire

Cette opération permet à un *partenaire* de déclarer les autres, en précisant leur nom et l'URL pointe vers l'interface d'accès à leur service web (figure 7.1). Chaque service web fournit un ensemble de méthodes permettant entre autres de répondre à une requête, ou de fournir son *ontologie générique*. Ainsi, pour constituer l'*ontologie globale*, le *partenaire* doit demander aux autres à qui il a déclaré son *ontologie générique*. Il peut aussi supprimer un *partenaire* à tout moment. La suppression d'un *partenaire* entraine le retrait de son *ontologie générique*.

Figure 7-1 Interface pour manager un partenaire

7.3 Générer un *dataweb sémantique partenaire*

Cette fonctionnalité permet de générer automatiquement le *dataweb* et le *dataweb sémantique* d'un *partenaire*, à partir de ses sources de données originelles, lorsque le hub de celui-ci est installé (voir 7.2).

Figure 7-2 Interface pour générer le dataweb et le dataweb sémantique d'un partenaire

On peut toujours répéter ce processus en ayant une vue sur l'évolution des différentes phases de génération du *dataweb* et du *dataweb sémantique*.

7.4 Visualiser la *base d'annotations* d'un *partenaire*

C'est une fonctionnalité qui permet d'avoir une vue sur l'ensemble des concepts de l'ontologie *partenaire*, lesquels ont été extraits à partir du *dataweb* (7.3). Elle permet aussi de voir les chemins Xpath permettant de faire le lien entre ces concepts et l'ensemble de leurs occurrences dans les documents XML, ainsi que de préciser ceux d'entre eux qui ont étés subsumés.

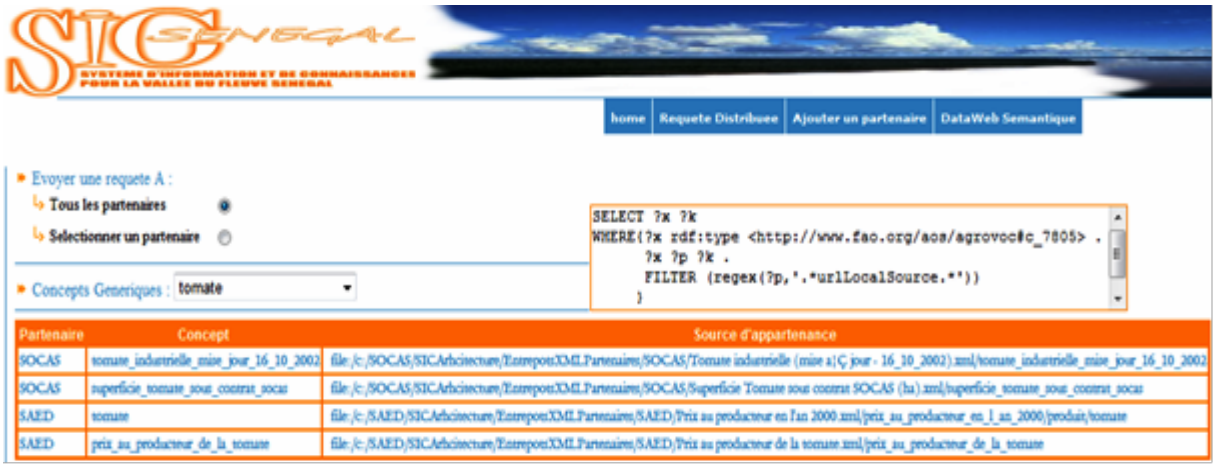


Figure 7-3 Interface pour visualiser la base d'annotations d'un partenaire

7.5 Envoyer une requête *partenaire* ou la distribuer

Dans la version actuelle du prototype, on ne peut soumettre une requête qu'en utilisant la liste des concepts disponibles dans l'*ontologie globale* .Cela permet néanmoins à un

partenaire de connaître ceux avec lesquels, parmi les autres *partenaires*, il a des informations en commun. En effet, un *partenaire* peut envoyer une requête à tous les autres en leur demandant s'ils ont des informations relatives à un concept générique qu'il choisit dans la liste des concepts de l'*ontologie globale*. Les autres répondent en lui fournissant leurs concepts concernés et les chemins Xpath indiquant leur source d'appartenance (7.4).

Le système génère une requête SPARQL qui est envoyée aux différents *partenaires* par l'invocation de leur service web. Cette requête est ensuite exécutée sur les bases de connaissances des *partenaires* recevant la requête par l'intermédiaire du moteur de recherche sémantique Corese. Les différentes réponses reçues par le *partenaire* ayant émis la requête sont fusionnées grâce à un moteur Corese temporaire prévu à cet effet.

Figure 7-4 Interface pour soumettre une requête à un ou plusieurs partenaires

7.6 Conclusion

Dans cette partie, nous avons présenté le prototype que nous avons mis en place pour implémenter notre approche. Il s'agit d'un système à base de hubs permettant aux différents *partenaires* de dialoguer à travers une application web et un ensemble de services accessibles à distance par l'intermédiaire d'un service web pour partager leurs sources de données hétérogènes.

Nous avons vu qu'avec ce prototype chaque *partenaire* peut, à partir de son application web, lancer le processus de génération automatique de son *dataweb* et de son *dataweb sémantique*. Ces opérations correspondent respectivement à la première phase, à savoir l'intégration structurelle des données initiales du *partenaire*, et à la première étape de la seconde phase, c'est-à-dire l'intégration sémantique des données, du processus générale d'intégration.

A l'issu de ces opérations, chaque *partenaire* dispose de son *dataweb* et de son *dataweb sémantique*. Il a la possibilité de visualiser sa *base d'annotations*, notamment l'ensemble des concepts de l'ontologie *partenaire* extraits du *dataweb*, le chemin de mapping permettant de faire le lien entre ces concepts et l'ensemble de leurs occurrences dans les documents XML.

Pour partager ses données, le *partenaire* doit d'abord connaître ses collaborateurs. Nous avons vu aussi que l'application web fournit une interface permettant de satisfaire à cela. En effet, à partir de cette interface, le *partenaire* peut déclarer ceux avec lesquels il veut partager ses données, en indiquant l'interface d'accès à leur service web, ce qui lui permet de communiquer avec eux et plus particulièrement de leur demander leurs ontologies génériques, requises pour la construction de l'*ontologie globale*, et de leur soumettre des requêtes.

Enfin, on a montré qu'avec la version actuelle de ce prototype on peut, à partir d'un concept générique de l'*ontologie globale*, trouver tous les *partenaires* ayant des informations relatives à ce concept. Autrement dit, un *partenaire* a les moyens d'envoyer une requête à tous les autres *partenaires* afin de connaître ceux avec lesquels il partage la même information. Il lui est également possible de localiser la provenance de ces informations dans les sources desdits *partenaires* pour d'éventuelles interrogations.

L'exemple présenté à la figure 7.4 permet de l'observer. Effectivement, cet exemple montre le *partenaire* SOCAS qui demande tous les *partenaires* ayant des informations relatives au concept tomate. Il reçoit des réponses provenant des *partenaires* SAED et ADRAO ainsi que, pour chacun d'entre eux, les sources où est localisée cette information.

Chapitre 8

Conclusion et perspectives

Sommaire

8.1 Conclusion 185

8.2 Contributions 186

8.2.1 Intégration de données par *partenaire* par une *approche dataweb sémantique* 187

8.2.1.1 Approche d'homogénéisation structurelle basée sur la notion de *dataweb* ... 187

8.2.1.2 *Approche dataweb sémantique* pour l'intégration sémantique 188

8.2.2 Une approche d'intégration des *dataweb sémantique*s 189

8.3 Travaux et perspectives de recherches 189

8.3.1 Maintenance évolutive 190

8.3.2 Langage et interface de requêtes 191

8.3.3 Extension aux autres formats et aux applications distribuées 191

8.3.4 Construction de *dataweb* thématiques 191

8.3.5 Imputation des données manquantes 192

8.1 Conclusion

L'expression web sémantique, attribuée à Tim Berners-Lee au sein du W3C, fait d'abord référence à la vision du web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés[Laublet et al., 2002]. Cependant, la diversité des formats de représentation des données et de formalisation des connaissances les décrivant lorsqu'elles sont disponibles constituent un frein à l'accessibilité et l'échange des données. La mise en place d'un système homogénéisant ou offrant un pont aux différents formats de représentations des données et des connaissances est une nécessité dans ce contexte. Ce support d'homogénéisation est ce que l'on nomme un système d'intégration.

Les travaux initiés dans cette thèse ont permis de dresser un état de l'art des approches et solutions proposées à l'intégration structurelle et sémantique des données et des sources de

données. Le projet *SIC-Sénégal* a pour objectif d'offrir une solution à l'intégration des sources hétérogènes produites dans la vallée du fleuve Sénégal en permettant ainsi une exploitation des données intégrées, facilitant ainsi la prise de décision et la recherche d'informations pertinentes sur un ensemble de données. A l'image de la démarche web sémantique, nous cherchons dans un cadre partenarial à offrir un cadre d'échange entre organismes intervenant dans la mise en valeur de la vallée un espace propice à la mise en commun intelligente des connaissances produites dans la zone malgré la diversité structurelle, sémantique et le volume des données.

Dans [Laublet et al., 2002], les auteurs affirment que l'on peut affirmer que le web sémantique doit d'abord être une infrastructure dans laquelle l'intégration des informations d'une variété de sources peut être réalisée et facilitée, c'est aussi le cas dans notre contexte où des organismes désirent partager et combiner leurs données. La mise en disposition de cette infrastructure nous a emmené à introduire un ensemble de concepts pour résoudre la problématique de l'intégration structurelle et sémantique des données : (1) un modèle de *dataweb sémantique* permettant l'intégration structurelle et sémantique des données de chaque organisme participant (2) un système d'intégration des *dataweb sémantique*s exploitant un système à base de hubs.

Ce chapitre présente une synthèse des travaux réalisés dans le cadre du projet *SIC-Sénégal* et expose les perspectives de recherche.

8.2 Contributions

L'étude d'une grande variété de solutions d'intégrations, l'existence de l'*approche dataweb* précédemment expérimentées sur les données produites dans la vallée du fleuve Sénégal nous ont permis de proposer une extension de la solution proposée dans les travaux de [Lo, 2002]. Il est ainsi apparu de cette étude que d'une part l'intégration des données passe par un cadre d'homogénéisation des formats de représentation et d'autre part, la compréhension et la combinaison automatique des données par des machines passe nécessairement par la mise à disposition d'un vocabulaire contrôlé décrivant les objets manipulés ainsi que leur hiérarchie. L'étude a également montré que la technologie la plus en vue et la plus partagée pour normaliser cette couche sémantique est résumée par la notion d'ontologie. Ainsi nous associons aux *dataweb* intégrant sémantiquement les données, des ontologies décrivant leurs données, une *base d'annotations* permet de servir de pont entre la couche sémantique et la couche des données décrites, constituant ainsi le début d'une *base de connaissances*.

Pour un ensemble d'organismes fournisseurs de données, cette approche permet le stockage et l'intégration de l'information en permettant de ne pas centraliser toutes les informations des organismes dans un même entrepôt de données mais une publication distribuée sur l'entrepôt de chaque fournisseur. L'intégration des connaissances d'un fournisseur de données dans une ontologie locale à la source suivie d'une construction d'une *ontologie globale* commune à tous permet ainsi de collecter, interroger, combiner automatiquement les données internes à un *partenaire* et entre les *entrepôts partenaires*.

8.2.1 Intégration de données par *partenaire* par une approche dataweb sémantique

L'*approche dataweb sémantique* comprend deux phases : celle permettant de résoudre la problématique de l'hétérogénéité structurelle et celle répondant localement puis globalement aux sources la dimension sémantique du besoin d'intégration.

8.2.1.1 Approche d'homogénéisation structurelle basée sur la notion de *dataweb*

En plus de la problématique des différents types d'hétérogénéité, les organismes fournisseurs ont posé deux contraintes à l'intégration des données : (1) disposer d'une autonomie d'exploitation et de gestion des données intégrées (2) préservation de l'aspect propriétaire et privées de certaines catégories de données. Ainsi, nous nous sommes basés sur l'*approche dataweb*, en transformant toutes les données *partenaires* au format XML constituant ainsi un *entrepôt de documents XML* pour chaque *partenaire*. Cette phase permet d'homogénéiser la représentation structurelle des données, favorisant ainsi leur combinaison et échange. Contrairement à l'*approche dataweb* développé dans les travaux de [Lo, 2002], nous avons choisi de représenter chaque document XML dans un entrepôt appartenant à son fournisseur, préservant ainsi l'aspect propriétaire des données.

Les données sont ainsi chargées, extraites puis transformées de leur format initial sous forme tabulaire au format XML. Une phase de restructuration permet le nettoyage des données brutes extraites et la mise en exergue des occurrences spécifiant l'origine géographique, la période de collecte des données ainsi que unités de mesures des données. Ces caractéristiques cibles sont souvent exprimées dans les tableaux de relevé de données dans les noms de colonnes décrivant les données. Ainsi cette seconde phase permet de préparer les données à leur exploitation et à l'extraction des connaissances les décrivant.

8.2.1.2 Approche dataweb sémantique pour l'intégration sémantique

Pour résoudre la problématique d'intégration sémantique des données, compte tenue de la nature environnementale des données décrites, notre approche est principalement basée sur une construction ascendante et semi-automatique des ontologies décrivant les données des sources grâce à la réutilisation d'une ontologie existante du domaine dite *AOS* (Agricultural Ontology Service). Une approche ascendante facilite l'étape d'annotation sémantique des documents puisque le vocabulaire de l'ontologie est crée à partir du vocabulaire de documents appartenant au même entrepôt. L'utilisation d'*AOS* permet de guider en partie la construction des ontologies (en particulier pour les concepts les plus génériques) et de faciliter l'alignement des différentes ontologies construites avec cette même référence. Chaque ontologie générée sera associée à une *base d'annotations* explicitant les correspondances entre les concepts de l'ontologie et les données décrites, créant ainsi une *base de connaissances*. Cette étape permet ainsi de générer une couche spécifique permettant de définir sémantiquement les données et les relations entre elles.

Dans le cadre applicatif, la mise en œuvre de cette approche passe nécessairement par la description de la sémantique encore inexistante de ces données, c'est ce que nous avons effectué avec la création d'une ontologie pour chaque *partenaire*. Cette ontologie capte le vocabulaire contrôlé du *partenaire* et permet de décrire la sémantique de ses données. Comme elle est construite à partir des documents XML, obtenus après une première phase d'intégration structurelle basées sur l'approche entrepôt de données, une mise en correspondance entre elle et ces documents XML est nécessaire afin d'identifier les sources XML originelles des concepts le composant.

C'est ainsi que nous avons alors utilisé une *base d'annotations* pour définir ces correspondances. La construction de l'ontologie *partenaire*, représentant le schéma global, à partir des documents XML, représentant les schémas locaux, et celle de la *base d'annotations*, représentant le schéma de mapping, justifient notre choix fait sur l'approche *GAV* pour réaliser l'intégration sémantique des données au sein d'un *partenaire*. Puisqu'une fois intégrées les données du *partenaire* doivent être partagées, nous avons construit, en même temps que l'ontologie *partenaire*, une *ontologie générique* contenant les concepts que le *partenaire* est susceptible d'avoir en commun avec les autres. L'ontologie *partenaire*, la *base d'annotations* et l'*ontologie générique* composent ce que l'on a appelé un *dataweb sémantique partenaire*.

En résumé, l'*approche dataweb sémantique* permet de réaliser l'intégration sémantique des données au sein d'un *partenaire* et facilite le partage de ces données entre les *partenaires*, grâce à l'*ontologie générique* qu'il intègre. Ce partage des données est mis en œuvre par l'intermédiaire d'un système à base de hubs (ou serveurs pairs), utilisé pour effectuer une médiation entre les différents *partenaires* contenant les *dataweb sémantique*s.

8.2.2 Une approche d'intégration des dataweb sémantiques

L'étape restante du processus d'intégration résultant des étapes énoncées dans la section précédente d'intégration des données d'un *partenaire* est l'intégration entre les différents *dataweb sémantique*s pour une mise en relation de leurs connaissances. Pour cet objectif, nous avons utilisé les travaux décrits dans [Gandon et al., 2008] basés sur les systèmes à base de hubs, permettant ainsi à chaque participant du système d'intégration disposant de données intégrées d'envoyer et de recevoir des requêtes. La médiation entre les différents hubs est effectuée par la construction d'une *ontologie globale* réunissant l'ensemble des connaissances que les différents participants au processus d'intégration ont en commun.

L'intégration entre les sources est donc effectuée en utilisant un système à base de hubs basé sur l'*ontologie globale* qui se trouve dupliquée chez chaque *partenaire*. Pour constituer l'*ontologie globale*, chaque *partenaire* reçoit l'*ontologie générique* des autres *partenaires* pour l'extraction des concepts en commun et subsumés dans $\mathcal{AOS}$. L'*ontologie globale* représente dans sa composante lexicale le vocabulaire partagé par tous les *partenaires* et utilisé pour la médiation entre les sources de données : les requêtes SPARQL sont effectuées sur cette *ontologie globale* via un moteur de recherche sémantique Corese dans notre prototype actuel. Elle représente le point d'entrée au niveau chaque hub. Elle sert de support pour l'expression des requêtes et fait office d'interface pour les ontologies *partenaires*, qui, permettent d'interroger les sources.

Dans cette architecture, l'utilisateur peut se connecter à n'importe quel hub par l'intermédiaire d'une application web pour soumettre des requêtes. Le hub se chargera alors de distribuer la requête aux autres hubs et de combiner les réponses. L'utilisation du moteur de recherche Corese et de l'ontologie externe de référence permet aussi d'éviter le processus onéreux de l'alignement des ontologies deux à deux.

8.3 Travaux et perspectives de recherches

Les résultats produits par cette thèse ont principalement porté sur la proposition d'une architecture d'intégration de données issues de sources hétérogènes. Les propositions ont

aussi permis de proposer un modèle formel de la démarche d'intégration ainsi que des différents composants des systèmes d'intégration *partenaire* et du système d'intégration globale. En plus de permettre aux organismes fournisseurs de données de partager autres qu'uniquement les connaissances qu'ils ont en commun, ces résultats ouvrent ainsi des perspectives de recherche pouvant faire suite à ce travail.

8.3.1 Maintenance évolutive

La problématique de l'évolution dans les *systèmes*(en considérant un système comme une source potentielle de données) est bien résumé par cette citation de Jean-Paul Baquiast: *« Il ne peut y avoir d'évolution si l'existant ne manifeste aucun changement. Dés qu'un changement se produit, il est en butte à une pression de sélection. Si le changement l'emporte sur l'existant, l'entité changée supplante l'entité précédemment existante - ceci jusqu'à renouvellement du cycle »*[44].

Les systèmes d'intégrations dans le contexte environnementale sont caractérisés par l'approche *« conservation/observation »*. Il est donc nécessaire de pouvoir contrôler les impacts de modifications des possibilités d'évolutions d'ajout ou de retrait de documents dans le système d'intégration ou carrément le retrait d'une source de données.

L'option d'utilisation des ontologies pour la médiation sémantique avec leur extraction des données *partenaires* n'est pas sans repos. Le simple retrait d'une source nécessitera des modifications des *ontologies génériques*, des *bases d'annotations* ainsi que de l'*ontologie globale*. Donc le contrôle des chemins de propagation d'impacts de modification sont d'une importance capitale.

La décomposition en composants des éléments du système d'intégration devra faciliter la prise en compte de cet aspect maintenance évolutive. Comme dans le cadre de la maintenance des ontologies, il suffira de contrôler les impacts de modification entre les composants et l'attitude à avoir dans chaque cas. Comme l'indique la citation tantôt rapportée, les changements sont cycliques, il convient de faire en sorte que l'entité nouvelle issue de la propagation des impacts de modification supplante à l'ancienne pour la pérennité du rôle qu'elle joue dans le système.

Des approches de contrôle de l'évolution d'ontologie son proposées dans [Faatz et Steinmetz, 2002], [Lin et Hovy, 2000].

[44] Http://philoscience.over-blog.com/article-17378351.html

8.3.2 Langage et interface de requêtes

L'une des difficultés rencontrées pour la validation applicative du système d'intégration est liée à la difficulté de mise en place d'un langage et d'une interface de requête utilisant les termes de l'*ontologie globale*. S'adressant à des utilisateurs ne maîtrisant pas forcément l'expression de requêtes sous SPARQL, il est nécessaire de les rendre plus naturelle.

En plus, il est nécessaire en se basant sur les résultats fournis dans [Gandon et al., 2008] de développer un algorithme de décomposition et de réécriture des requêtes utilisateurs et sur l'*ontologie globale* en requêtes locales adressées aux sources de données et de recompositions des résultats retournés. Cela fait d'ailleurs l'objet de travaux dans une thèse au sein de l'équipe.

8.3.3 Extension aux autres formats et aux applications distribuées

Nous avons travaillé dans cette thèse uniquement avec les données au format tabulaire. Cependant, elles ne sont pas la seule catégorie de données existante chez les *partenaires*. Des supports de données structurés textuellement ou sous forme multimédia par exemple existant chez les *partenaires*. Une intégration complète et efficace des données devra nécessairement les prendre en compte.

Il est aussi nécessaire de prendre en compte le cadre d'intégration des applications distribuées. Des travaux étant déjà initiés au sein du projet *SIC*-Sénégal, la problématique sera de procéder à un « binding » afin d'adapter les résultats obtenus au prototype AIDE-ISH.

8.3.4 Construction de dataweb thématiques

L'une des possibilités souhaitées et non explorées par nos travaux dans cette thèse est la mise en place de *dataweb* thématiques ou du moins la *thématisation* des connaissances selon les différents secteurs environnementaux. Elle permettra ainsi de ne par exemple interroger que les données traitant de la pêche si c'est la requête de l'utilisateur.

Divers travaux dans le domaine de la classification traitent de cet aspect. Il est aussi possible de s'inspirer de l'option que nous avons fait avec la problématique propriétaire des connaissances, la *thématisation* peut se traiter en partitionnant en autant de parties que de thématique les ontologies *partenaires* et celle globale. Une autre alternative est tout simplement d'introduire une annotation thématique supplémentaire des connaissances. Il est

bien possible de se limiter aux connaissances pour traiter cette problématique, les données constituant uniquement des sortes d'instances de celles-ci.

8.3.5 Imputation des données manquantes

Un problème fréquemment posé par l'exploitation des données d'origine environnementale est celle des données manquantes. Le système que nous visons à mettre en place devrait aider à prendre des décisions en fonction des valeurs trouvées. Il est donc important de prendre en compte cet aspect en mettant en place des procédures d'imputation pour les données manquantes. L'imputation est le processus visant à combler les valeurs manquantes par des valeurs prédites ou simulées.

Cependant, la nature environnementale de ces données apporte son lot de complexité. En effet, elles sont géo-référencées. Ces données ont été recueillies dans des espaces bien définies et selon une période de temps définies. Donc quelque soit la procédure d'imputation mise en place, elle se devra de prendre en compte l'aspect restrictif sur l'espace ainsi que le temps. En plus, nous l'avons expérimenté en traçant le graphique de chacune des ces lignes de tableaux lorsque c'est possible, mais aucun ne ressemble à un des graphiques de loi d'estimation statistique donnée.

Dans les échantillons de données fournies par les *partenaires* dans le contexte du projet *SIC*-Sénégal, il existe de rares tableaux ne contenant pas de champs non renseignés. Il s'avère que sur les 39 tableaux de données des échantillons *partenaires*, 13 présentent des données manquantes en leur sein soit 33% des tableaux de données, ce que l'on ne peut pas ignorer comme c'est le cas pour certains traitements en statistiques. Si cette problématique n'est pas résolue au moment du passage à l'échelle, les individus non renseignés risquent d'être si nombreux que l'analyse de certaines caractéristiques de certains individus en question perd tout intérêt. Nous construisons des entrepôts, donc l'aspect analyse et combinaison relève d'une importance capitale.

Prenons l'exemple du tableau 6.1 vue en page162. Il apparaît que dans la région de Saint-Louis, les superficies infestées pendant la saison agricole 2002/2003 pour le riz et les pastèques ne sont pas renseignées. Une idée basique des méthodologies d'imputation de données manquantes consiste à se rabattre sur les données concernant les mêmes espèces de culture dans le même tableau. Cependant, le problème qui est posé ici est que ces données pour la même espèce concernent une région géographique différente.

Nous manipulons des données de nature agricole et environnementale en général, ce qui induit que selon les zones géographiques, les conditions n'étant pas les mêmes, il est

hasardeux de se baser sur les relevés d'une autre unité spatiale, peut être que des conditions particulières font que la prolifération des infections est plus accentuée d'une zone à l'autre.

D'autant plus que nous pouvons constater que l'écart des relevés entre la zone de «Saint-Louis» et «Dagana» pour tous les types de culture est considérable. L'approche que nous proposons d'utiliser ici consiste à profiter de la nouvelle connaissance sur la sémantique des données offertes par les ontologies. Autrement dit utiliser une méthodologie d'imputation des données manquantes, mais en faisant, dans le cas où c'est possible, le lien de cette donnée manquante avec d'autres individus sur le même espace-temps et avec ou partageant une sémantique. C'est d'ailleurs ce qui a en partie motivé la mise en exergue des caractéristiques spatio-temporelles lors de la phase de construction des dataweb.

Bibliographie

[Abiteboul et Duschka, 1998] Abiteboul, S. et Duschka, O.M., (1998). *Complexity of answering queries using materialized views*. In Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (Seattle, Washington, United States, June 01 - 04, 1998). PODS '98. ACM, New York, NY, pp. 254-263. DOI= http ://doi.acm.org/10.1145/275487.275516.

[Abiteboul et al., 2003] Abiteboul, S., Bonifati, A., Cobéna, G., Manolescu, I., et Milo, T. (2003). *Dynamic XML documents with distribution and replication*. In Proceedings of the 2003 ACM SIGMOD international Conference on Management of Data (San Diego, California, June 09 - 12, 2003). SIGMOD '03. ACM, New York, NY, pp. 527-538. DOI= http ://doi.acm.org/10.1145/872757.872821.

[Abiteboul et al., 2004] Abiteboul, S., Benjelloun, O., Cautis, B., Manolescu, I., Milo, T., et Preda, N. (2004). *Lazy query evaluation for Active XML*. In Proceedings of the 2004 ACM SIGMOD international Conference on Management of Data (Paris, France, June 13 - 18, 2004). SIGMOD '04. ACM, New York, NY, pp.227-238. DOI= http ://doi.acm.org/10.1145/1007568.1007596.

[Amann et al., 2002(a)] Amann, B., Beeri, C., Fundulaki, I., et Scholl, M. (2002). *Ontology-Based Integration of XML Web Resources*. In Proceedings of the First international Semantic Web Conference on the Semantic Web (June 09 - 12, 2002). I. Horrocks and J. A. Hendler, Eds. Lecture Notes In Computer Science, vol. 2342. Springer-Verlag, London, pp. 117-131.

[Amann et al., 2002(b)] Amann, B., Beeri, C., Fundulaki, I., et Scholl, M. (2002). *Querying XML Sources Using an Ontology-Based Mediator*. In on the Move To Meaningful internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated international Conferences Doa, CoopIS and ODBASE 2002 (October 30 -November 01, 2002). R. Meersman and Z. Tari, Eds. Lecture Notes In Computer Science, vol. 2519. Springer-Verlag, London, pp. 429-448.

[Arens et al., 1996] Arens, Y., Hsu, C. et Knoblock, C. A., (1998). *Query Processing in the SIMS Information Mediator*. In Austin Tate, editor, Advanced Planning Technology, pp. 61-69. AAAI Press, Menlo Park, California.

[Aussenac-Gilles et al., 2000(a)] Aussenac-Gilles N., Biébow B., et Szulman S. (2000).

Modélisation du domaine par une méthode fondée sur l'analyse de corpus. Dans : 9e Conférence Francophone d'Ingénierie des Connaissances IC 2000, Toulouse (F), 10/05/00-12/05/00, Université Paul Sabatier, Toulouse (F), pp. 93-104, mai 2000.

[Aussenac-Gilles et al., 2000(b)] Aussenac-Gilles N., Biébow B., et Szulman S. (2000). *Revisiting Ontology Design: a method based on corpus analysis*. Dans: 12th International Conference on Knowledge Engineering and Knowledge Management, Juans-Les-Pins (F),03/10/00-06/10/00, Springer Verlag, Heidelberg (G), pp. 172-188, octobre 2000.

[Bachimont, 1996] Bachimont, B. (1996). *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser*, Thèse d'épistémologie, Ecole Polytechnique, Paris, 1996.

[Bachimont, 2004] Bachimont, B. (2004). *Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*, Mémoire d'Habilitation à Diriger des Recherches, Université de Technologie de Compiègne, France.

[Battle, 2004] Battle, S. (2004). *Round-tripping between XML and RDF*. In International Semantic Web Conference ISWC, Hiroshima, Japan, November 2004. Springer, 2004.

[Bdisic, 2004] BDI*SIC*. (2004). *Projet SIC-Sénegal. Compte rendu du Workshop des 10 et 11 juin 2004*, Unité de Formation et de Recherche en Sciences Appliquées et de Technologie, Université Gaston Berger de Saint-Louis du Sénégal.

[Bellatreche et al., 2004] Bellatreche, L., Pierra, G., Xuan, Dung. et Hondjack, D. (2004). *Intégration de sources de données autonomes par articulation a priori d'ontologies*. In Proc. du 23ème congrès Inforsid, may 2004, pp. 283-298.

[Beneventano et al., 2000] Beneventano, D., Bergamaschi, S., Castano, S., Corni A., Guidetti, R., Malvezzi, G., Melchiori, M., et Vincini, M. (2000) *Information integration : the MOMIS project demonstration*. In Proc. of the 26th Int. Conf. On Very Large Data Bases VLDB 2000.

[Berners-Lee et Lassila, 2001] Berners-Lee, T., Hendler, J., et Lassila O. (2001). *The Semantic web*. In Scientific American, May 2001, pp. 35-43.

[Bohring et Auer, 2005] Bohring, H., et Auer, S. (2005). *Mapping XML to OWL Ontologies*. In Jantke, K. P., F´'uhnrich, K.-P., and Wittig, W. S., editors, Leipziger Informatik-Tage, volume 72 of LNI, pp. 147-156. GI.

[Borst 1997] Borst, W.N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Doctoral Thesis of the University of Tweenty. In Enschede Publisher, The Netherlands.

[Bouquet et al., 2003] Bouquet, P., Magnini, B., Serafini, L. et Zanobini, S. (2003). *A SAT-Based Algorithm for Context Matching*. In : 4th International and Interdisciplinary Conference, CONTEXT 2003. June 23-25, 2003, Stanford, CA, USA. Springer, 2003, Lecture Notes in Computer Science, vol. 2680, pp. 66-79.

[Buccella et Cechich, 2003] Buccella, A. et Cechich, A. (2003). *An ontology approach to data integration*. In JCS&T, Vol. 3, No. 2, October, pp. 62-68.

[Buche et al., 2005] Buche, P., Dibie-Barthélemy, J., Doussot, D., Haemmerlé, O., Hignette, G., et Thomopoulos, R.(2005). *MIEL++ : un entrepôt intégrant des données floues exprimées en graphes conceptuels, bases de données relationnelles et XML*. AFIA 2005.

[Calì et al., 2005] Calì, A., Lembo, D., et Rosati, R.(2005). *A comprehensive semantic framework for data integration systems*, Journal of Applied Logic, volume 3, number 2, pp. 308-328, Elsevier Science Publishers North-Holland, Amsterdam, 2005. ISSN 1570-8683.

[Calvier et Reynaud, 2008] Calvier, F. et Reynaud, C. (2008). *Ontology Matching Supported by Query Answering in a P2P System*. In Proceedings of the OTM 2008 Confederated international Conferences, Coopis, Doa, Gada, Is, and ODBASE 2008. Part II on on the Move To Meaningful internet Systems (Monterrey, Mexico, November 09 - 17, 2008). R. Meersman and Z. Tari, Eds. Lecture Notes In Computer Science, vol. 5332. Springer-Verlag, Berlin, Heidelberg, pp.1559–1567. DOI= http ://dx.doi.org/10.1007/978-3-540-88873-4_44.

[Carroll et al., 2003] Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., et Wilkinson, K.(2004). *Jena : implementing the semantic web recommendations*. In Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers and Amp; Posters (New York, NY, USA, May 19- 21, 2004). WWW Alt. '04. ACM, New York, NY, pp. 74-83. DOI= http ://doi.acm.org/10.1145/1013367.1013381

[Charlet, 2002] Charlet J. (2002). *L'ingénierie des connaissances, développements, résultats et perspectives pour la gestion des connaissances médicales*. Mémoire d'Habilitation à Diriger des Recherches, Université Pierre et Marie Curie, Paris, 2002.

[Chong et al., 2002] Chong, Q., Mullins, J., et Rajasekharan, R. (2002). *An ontology-based metadata management system for heterogeneous distributed databases*. CS590LWinter 2002, Project Proposal January 2002 - Heuristic Crosswalk Worktool (HCW) Group, http ://r.web.umkc.edu/rr015/590L.htm

[Clark et. al , 2000] Clark, P., Thompson, J., Holmback, H., et Duncan, L. (2000). *Exploiting*

a Thesaurus-Based Semantic Net for Knowledge-Based Search. In Proceedings of the Seventeenth National Conference on Artificial intelligence and Twelfth Conference on innovative Applications of Artificial intelligence (July 30 - August 03, 2000). AAAI Press / The MIT Press, pp. 988-995.

[Corby et al, 2004] Corby O., Dieng-Kuntz R. et Faron-Zucker C. (2004). *Querying the Semantic Web with the CORESE engine*. In R. Lopez de Mantaras and L. Saitta eds,Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-2004), Valencia,Spain, IOS Press, pp. 705-709.

[Cruz et al., 2004] Cruz, I. F., Xiao, H., et Hsu, F. (2004*). An Ontology-Based Framework for XML Semantic Integration*. In Proceedings of the international Database Engineering and Applications Symposium (July 07 - 09, 2004). IDEAS. IEEE Computer Society, Washington, DC, pp. 217-226. DOI=http ://dx.doi.org/10.1109/IDEAS.2004.10.

[Cui et al., 2001] Cui, Z., Jones, D., et O'Brien, P. (2001). *Issues in ontology-based in formation integration*. In Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, 4-5 August, Seattle, USA, pp.141-146.

[Dang-Ngoc et al., 2004b] Dang-Ngoc, T.-T., Jamard, C., et Travers, N. (2005*). XLive : An XML Light Integration Virtual Engine*. In Proc. of Bases de Données Avancées (BDA)'2005. Saint-Malo, France.

[Delobel et al., 2003] Delobel, C., Reynaud, C., Rousset, M-C., Sirot, J-P et Vodis*lav*, D. (2003), *Semantic Integration in Xyleme : a Uniform Tree-Based Approach*, Data and Knowledge Engineering Review, 44(2), pp. 267-298.

[Durville et Gandon, 2005] Durville, P. et Gandon, F.(2005). *Sewese : Semantic web serve*r. http ://wwwsop. inria.fr/edelweiss/wiki/wakka.php ?wiki=Sewese, 2005.

[Dieng, 1996] Dieng Kuntz, R.(1996). *Ontologies for Knowledge Management*, Interoperability Research School, April 13, 2006.

[Ducourneau et. al, 1998] Ducournau, R., Euzenat, J., Masini. et G., Napoli, A.(1998). *Langages et modèles à Objets Etat des recherches et perspectives*. INRIA, Collection Didactique, 1998, ISSN 0299-0733 ; ISBN 2-7261-1131

[Dzeakou et. al, 1998] Dzeakou, P., Morand, P., et Mullon C. (1998) *Méthodes et architectures des Systèmes d'Information sur l'environnement*, Actes du 4ième congrès Conférence Africaine de Recherche en Informatique (CARI), Dakar, Sénégal, INRIA ed., pp. 509-520, octobre, 1998.

[Dzeakou et Derniame 1998] Dzeakou, P., Derniame, J.C.(1998). *Conception des systèmes*

d'information des observatoires environnementaux : Une architecture de médiation. Rapport de recherche, Projet SIMES, LORIA.

[Faatz et Steinmetz, 2002] Faatz, A., et Steinmetz, R. (2002). *Ontology enrichment with texts from the WWW*. In Proceedings of the 2nd Semantic Web Mining Workshop at ECMLI/PKDD.

[Faye et al., 2006] Faye, D.C., Nachouki, G., et Valduriez, P. (2006). *SenPeer : Un système Pair-à-Pair de médiation de données*. In Volume 4 - pp. 24-52 - In ARIMA 2006.

[Fayyad et al., 1996] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., et et Uthurusamy., R.(1996). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence.

[Fayyad et al., 2001] Fayyad, U., Grinstein, G. G., et Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery*. 1st. Morgan Kaufmann Publishers Inc. 2001.

[Fellbaum, 1998] Fellbaum, C. (1998). *Wordnet : An Electronic Lexical Database*. The MIT Press, Cambridge, Massachussets, Etats-Unis, 1998.

[Florescu et. al., 1998] Florescu, D., Levy, A. et Mendelzon, A. *Database Techniques for the World-Wide Web : A Survey*. SIGMOD Record 27(3),1998.

[Frawley et al., 1991] Frawley W.J., Piatetsky-Shapiro., G., Matheus., C., (1991). *Knowledge Discovery in Databases : An Overview*. AAAI/MIT Press. Isbn 0-262-62080-4.

[Fundulaki et al., 2002] Fundulaki, I., Amann, B., Beeri, C., Scholl, M., et Vercoustre, A. (2002). *STYX : Connecting the XML Web to the World of Semantics*. In Proceedings of the 8th international Conference on Extending Database Technology : Advances in Database Technology (March 25 - 27, 2002). C. S. Jensen, K. G. Jeffery, J. Pokorný, S. Saltenis, E. Bertino, K. Böhm, and M. Jarke, Eds. Extending Database Technology, vol. 2287. Springer-Verlag, London, pp.759-761.

[Furst, 2004] Furst F. (2004). *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation*, Thèse de doctorat, Université de Nantes, France.

[Gandon et al., 2008] Gandon, F., Lo, M. et Niang, C. (2008). *Un modèle d'index pour la résolution distribuée de requêtes sur un nombre restreint de bases d'annotations RDF*. In Yannick Prié, ed., Actes d'IC'2008, Institut National Polytechnique de Lorraine, pp. 25-35.

[Gardarin et Dang-Ngoc, 2004] Gardarin, G., et Dang-Ngoc, T.-T. (2004), *Mediating the semantic web*. In Georges Hébrail ; Ludovic Lebart and Jean-Marc Petit, ed., EGC'2004 , Cépaduès-Éditions, pp. 1-14. Clermont-Ferrand, France.

[Gerat, 2007] Gerat, L. (2007). *Xedix, un système de gestion de masses de données*. In La Recherche no 408, pp. 15-17, 2007.

[Goh, 1997] Goh, C. H. (1997). *Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems*. Doctoral Thesis. UMI Order Number : AAI0597943., Massachusetts Institute of Technology.

[Grüber, 1993] Grüber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowl. Acquis. 5, 2 (Jun. 1993), pp. 199-220. DOI=http ://dx.doi.org/10.1006/knac.1993.1008

[Halevy, 2001] Halevy, A. (2001). *Answering queries using views : A survey*. The VLDB Journal 10, 4 (Dec. 2001), pp. 270-294. DOI=http ://dx.doi.org/10.1007/s007780100054

[Han et Kamber, 2001] Han, J., et Kamber, M.(2001). Data Mining : Concepts and Techniques. Morgan Kaufmann, 2001.

[Hazaël-Massieux, 2005] Hazaël-Massieux, D. (2005). *Bridging XHTML, XML and RDF with GRDDL*. In XTech 2005 : XML, the Web and beyond., Amsterdam, The Netherlands. IDE Alliance. Proceedings available at http://www.idealliance.org/proceedings/xtech05/.

[Hazaël-Massieux et Connolly, 2005] Hazaël-Massieux, D. et Connolly, D. ,2005. *Gleaning Resource Descriptions from Dialects of Languages(GRDDL)*. http ://www.w3.org/TeamSubmission/grddl/. Seen July 2006.

[Hernandez et. al, 2006] Hernandez, N., Chrisment, N., Hubert, G., et Mothe, J.(2006). *Mise à jour d'une ontologie de domaine à partir de l'analyse de nouveaux documents du domaine pour l'indexation de documents*. In Information – Interaction - Intelligence I3, Cépaduès Editions, Numéro spécial Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance, 2006.

[Hocine et. Lo, 2000] Hocine, A. et Lo, M. (2000). *Modeling and Information Retrieval on XML-Based Dataweb*. In Proceedings of the First international Conference on Advances in information Systems (October 25 - 27, 2000). T. M. Yakhno, Ed. Lecture Notes In Computer Science, vol. 1909. Springer-Verlag, London, pp. 398-408.

[Hull, 1997] Hull, R. (1997*). Managing semantic heterogeneity in databases : a theoretical prospective*. In Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (Tucson, Arizona, United States, May 11 - 15, 1997). PODS '97. ACM, New York, NY, pp. 51-61. DOI= http ://doi.acm.org/10.1145/263661.263668

[Hull et. Zhou, 1996] Hull, R. Zhou G. *A framework for supporting data integration using materialized and virtual approaches*. Proceedings of ACM SIGMOD Conference on Management of Data, pp. 481-492, Montreal, Canada, 1996.

[Isabel et al., 2004] Cruz, I. F., Xiao, H., et Hsu, F. (2004). *An Ontology-Based Framework for XML Semantic Integration*. In Proceedings of the international Database Engineering and Applications Symposium (July 07 - 09, 2004). IDEAS. IEEE Computer Society, Washington, DC, pp. 217-226.
DOI=http ://dx.doi.org/10.1109/IDEAS.2004.10

[Kasset et Niang, 2006] Kasset, C. A. et Niang, K. (2006). Développement d'un wrapper Excel-XML, Mémoire de maîtrise informatique, Université Gaston Berger de Saint-Louis Sénégal, juillet 2006.

[Kayser 1997] Kayser, D.(1997). *La représentation des connaissances*. In Hermes, ISBN 2-86601-647-5, 1997.

[Kietz et al., 2000] Kietz, J. U., Maedche, A., et Volz, R. (2000). *A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet*. In Aussenac, B.Biébow, and S. Szulman (eds.), EKAW'00 Workshop on Ontologies and Texts, vol. 51, Juan-Les-Pins, France. CEUR Workshop Proceedings.

[Kiryakov et al., 2003] Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A. et Goranov, M. (2003*), Semantic Annotation, Indexing, and Retrieval*. In International Semantic Web Conference , pp. 484-499 .

[Klein, 2002] Klein, M. C. (2002). *Interpreting XML Documents via an RDF-Schema Ontology*. In Proceedings of the 13th international Workshop on Database and Expert Systems Applications (September 02 - 06, 2002). DEXA. IEEE Computer Society, Washington, DC, pp. 889-894.

[Lakshmanan et Sadri ,2003] Lakshmanan, L. V., et Sadri, F., (2003*). Interoperability on XML Dat*a. In Proceedings of ISWC 2003 : international semantic web conference No2, Sanibel Island FL , ETATS-UNIS (20/10/2003) 2003, vol. 2870, pp.146-163.

[Lassila et Swick, 1999] Lassila, O., Swick, R. R.(1999). *Ressource description framework rdf model and syntax specification*. W3c recommendation 22. février 1999.
http ://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999.

[Laublet et al., 2002] Laublet, P., Reynaud, C., et Charlet, J. (2002). *Sur quelques aspects du Web sémantique*. In Actes des 2 assises du GdR I : Information – Interaction - Intelligence, pp. 217-231.

[Laurent, 2004] Laurent, F. (2004). 2004, l'année RSS.

http://xmlfr.org/documentations/articles/040331-0001

[Lechevallier, 2005] Lechevallier, Y. (2005). *Le tableau de données, une structure unique, des réalités multiples*, 21 March 2005, RDC'2005 ENST Paris.

[Lenzerini, 2002] Lenzerini, M. (2002). *Data integration : a theoretical perspective*. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems' , ACM, Madison, Wisconsin , pp. 233-246 .

[Lima et al., 2003] Lima, J.G.S., Medeiros, C.M.B., Assad, E.D.(2003). *Integration of Heterogeneous Pluviometric Data For Crop Forecast*, 11-2003, GEOINFO 2003 - V Simpósio Brasileiro de Geoinformá tica, Vol. 1, pp.10-20, Campos do Jordão, SP, Brasil, 2003.

[Lin et Hovy, 2000] Lin, C. et Hovy, E. (2000). *The automated acquisition of topic signatures for text summarization*. In Proceedings of the 18th Conference on Computational Linguistics - Volume 1 (Saarbrücken, Germany, July 31 -August 04, 2000). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 495-501.
DOI=http ://dx.doi.org/10.3115/990820.990892

[Lo, 2002] Lo, M.(2002). *Dataweb basés sur XML : modélisation et recherche d'informations pertinentes*. Thèse de Doctorat de l'Université de Pau et des Pays de l'Adour, Décembre 2002.

[Lo et Hocine, 2000] Lo, M., Hocine, A.(2000). *Modélisation de dataweb : une approche basée sur l'intégration de la sémantique des données et XML*. Actes du congrès 6e Conférence Africaine de Recherche en Informatique (CARI'2000), Antanarivo, Madagascar, INRIA ed., pp. 31-38, 16-19 octobre, 2000.

[Lo et Hocine, 2005] Lo, M. et Hocine A. (2005*). ISYWEB : an XML-based Architecture for Web Information Systems*. In Proceedings of SITIS'05 Conference, pp. 116-121, Yaoundé (Cameroun).

[Luong, 2007] Luong, P-H. (2007). *Gestion de l'évolution d'un web sémantique d'entreprise*. Thèse de Doctorat de l'Ecole des Mines de Paris, Décembre 2007.

[Maedche et Staab, 2001] Maedche, A. et Staab, S. 2001. *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems 16, 2 (Mar. 2001), pp. 72-79.
DOI=http ://dx.doi.org/10.1109/5254.920602

[Maedche et. al, 2002] Maedche, A., Pekar, V., et Staab, S. (2002). *Ontology learning part one - on discovering taxonomic relations from the web*, In Web Intelligence, Z.Ning et al Eds., Spinger, 2002.

[Matthias et al., 2004] Ferdinand, M., Zirpins, C., et Trastour, D. (2004). *Lifting XML-Schema to OWL*. In Nora Koch, Piero Fraternali, and Martin Wirsing, editors, Web Engineering - 4th International Conference, ICWE 2004, Munich, Germany, July 26-30, 2004, Proceedings, pages 354-358. Springer Heidelberg, 2004.

[McGuinness et Harmelen, 2004] McGuinness, D.L., et Harmelen, F.V.(2004). OWL *Web Ontology Language Overview*, W3C Recommendation http ://www.w3.org/TR/owl-features/, 10 February 2004.

[Miles et al., 2003] Miles A. J., Rogers N., Beckett D., (2003). *Migrating thesauri to the semantic web, guidelines and case studies for generating RDF encodings of existing thesauri*, SWAD Europe Thesaurus Activity, Delivrable 8.8, 2003.

[Miller et al., 1993] Miller, G. A., Leacock, C., Tengi, R., et Bunker, R. T. (1993). *A semantic concordance*. In Proceedings of the Workshop on Human Language Technology (Princeton, New Jersey, March 21 - 24, 1993). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, pp. 303-308. DOI= http ://dx.doi.org/10.3115/1075671.1075742

[Milo et al., 2003] Milo, T., Abiteboul, S., Amann, B., Benjelloun, O., Ngoc, F. D. (2003). *Exchanging intensional xml data. In Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 289-300, San Diego, California, USA, 2003.

[Mizoguchi, 1998] Mizoguchi R.(1998). *A Step Towards Ontological Engineering*. Paper presented at the 12th National Conference on AI of JSAI, pp.24-31, June, 1998.

[Navigli et al., 2003] Navigli, R., Velardi, P., et Gangemi, A. (2003). *Ontology Learning and Its Application to Automated Terminology Translation*. IEEE Intelligent Systems, vol. 18, n.1 (Jan. 2003), pp. 22-31. DOI=http ://dx.doi.org/10.1109/MIS.2003.1179190.

[Niang, 2008] Niang., C.A.T.(2008). *Atelier d'Intégration de Données Environnementales issues de Sources Hétérogènes*. Stage de Recherche, Laboratoire d'Analyse Numérique et d'Informatique, Université Gaston Berger de Saint-Louis, 2008.

[Niang, 2007] Niang., C.A.T.(2007). *Serveurs distribués pour une mémoire d'ingénierie, intégration des sources et résolution distribuée de requêtes*. Rapport de stage à INRIA Sophia Antipolis, France.

[Papakonstantinou et al., 2003] Papakonstantinou, Y., Borkar, V., Orgiyan, M., Stathatos, K., Suta, L., Vassalos, V. et Velikhov, P. (2003). XML queries and algebra in the Enosys integration platform. Data Knowledge Engineering, Vol. 44, No.3, pp. 299-322.

[Pierra et al., 2005] Pierra G., Hondjack D., Jean S., Xuan D.N.(2005). *Ingénierie dirigée par*

les modèles en EXPRESS : un exemple d'application, IDM'05 Premières Journées sur l'Ingénierie Dirigée par les Modèles Paris, 30 juin, 1 juillet 2005.

[Psyché et al., 2003] Psyché V., Mendes O., et Bourdeau J., (2003). *Apport de l'ingénierie ontologique aux environnements de formation à distance*. STICEF, 10 Numéro spécial : Technologies et Formation à distance.

[Reif et. al, 2005] Reif, G., Gall, H., et Jazayeri, M. (2005*). WEESA : Web engineering for semantic Web applications*. In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM, New York, NY, pp. 722-729. DOI= http ://doi.acm.org/10.1145/1060745.1060849

[Rocha, 2003] Rocha, H.A. (2003). *Metadados para workflows cientõficos no apoio ao planejamento ambiental*. Masters thesis, Instituto de Computacão, UNICAMP, Campinas, Brazil, 2003.

[Rodrigues et al., 2006] Rodrigues, T., Rosa, P., et Cardoso, J. (2006*). Mapping XML to Exiting OWL ontologie*s, International Conference WWW/Internet 2006, (Eds) Isaías, Pedro and Nunes, Miguel Baptista and Martínez, Inmaculada J., pp.72-77, ISBN :972-8924-19-4, 2006.

[Sall et Lo, 2007] Sall, O., et Lo, M.(2007). *Intégration de données environnementales : une approche basée sur les entrepts de documents XML et les ontologies*, Revue des Nouvelles Technologies de l'Information EDA'2007, pp. 147-160, Juin, 2007.

[Sall et al., 2009] Sall, O., Lo, M., Gandon, F., Niang, C., et Diop, I. (2009). *Using XML data integration and ontology reuse to share agricultural data*, Int. J. Metadata, Semantics and Ontologies, Vol. 4, Nos. 1/2, pp.93-105.

[Schöning et Wäch, 2000] Schöning, H. et Wäsch, J. (2000). *Tamino - An Internet Database System*. In Proceedings of the 7th international Conference on Extending Database Technology : Advances in Database Technology (March 27 - 31, 2000). C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust, Eds. Extending Database Technology, vol. 1777. Springer-Verlag, London, pp. 383-387.

[Stuckenschmidt et al., 2000] Stuckenschmidt, H., Wache, H., V¨ögele, T., et Visser, U.(2000). *Enabling technologies for interoperability*. Workshop on the 14th International Symposium of Computer Science for Environmental Protection, pp. 35-46, Bonn, Germany, 2000. TZI, University of Bremen.

[Simperl et Tempich, 2006] Simperl., E.P.B. et Tempich., C. (2006). *Ontology engineering: a reality check*. In R. Meersman and Z. Tari and others, The 5th International Conference on Ontologies, DataBases, and Applications of Semantics ODBASE2006,

volume 4275 of Lecture Notes in Computer Science LNCS, pp.836-854. Springer, Montpellier, France, November 2006. [Sowa, 1999] Sowa J.(1999). Conceptual Graph Standard. Http ://www.bestweb.net/_sowa/cgdpans.htm, 1999.

[Spivack, 2007] Spivack., N. (2007). *The Semantic Web, Collective Intelligence and Hyperdata*.
Http ://novaspivack.typepad.com/nova_spivacks_weblog/2007/09/hyperdata.html.

[Ullman, 1997] J. D. Ullman.(1997). *Information integration using logical views*. In Proc. Of the 6th Int. Conf. on Database Theory ICDT'97, volume 1186 of Lecture Notes in Computer Science, pp. 19-40. Springer, 1997.

[Uschold et Gruninger, 1996] Uschold, M. et Gruninger, M. (1996). *Ontologies : Principles, Methods and Applications*. Knowledge Engineering Review 11,2.

[Visscher, 2005] Visscher, S.(2005). *RDF Extraction from XML (XR)*, October 2005.
http ://w3future.com/xr/.

[Wache et.al, 1999] Wache, H., Scholz, T., Stieghahn, H., et König-Ries, B. (1999). *An Integration Method for the Specification of Rule-Oriented Mediators*. In Proceedings of the 1999 international Symposium on Database Applications in Non-Traditional Environments (November 28 - 30, 1999). DANTE. IEEE Computer Society, Washington, DC, 109.

[Wache et al., 2001] Wache, H., T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, et S. Hübner (2001). *Ontology-based integration of information – a survey of existing approaches*. In H. Stuckenschmidt (Ed.), IJCAI-01Workshop : Ontologies and Information Sharing, pp. 108-117.

[Widom, 1995] Widom, J. (1995). *Research problems in data warehousing*. In Proceedings of the Fourth international Conference on information and Knowledge Management (Baltimore, Maryland, United States, November 29 - December 02, 1995). N. Pissinou, A. Silberschatz, E. K. Park, and K. Makki, Eds. CIKM '95. ACM, New York, NY, pp. 25-30. DOI= http ://doi.acm.org/10.1145/221270.221319

[Wiederhold, 1995] Wiederhold, G. 1995. *Mediation in information systems*. ACM Comput. Surv. 27, 2 (Jun. 1995), pp. 265-267.
DOI= http ://doi.acm.org/10.1145/210376.210390

[Wielinga et. al ,2001] Wielinga, B. J., Schreiber, A. T., Wielemaker, J., et Sandberg, J. A. (2001). *From thesaurus to ontology*. In Proceedings of the 1st international Conference on Knowledge Capture (Victoria, British Columbia, Canada, October 22 - 23, 2001). K-CAP '01. ACM, New York, NY, pp. 194-201.

DOI=http ://doi.acm.org/10.1145/500737.500767.

Annexe A

A.1 Sources Tabulaires initiales

Année	Collecte (T)	Fourchette prix paddy (F.cfa/kg)	Moyenne	Fourchette riz blanc (F.cfa/kg)	Moyenne
1994/95	31000	97,6-115	100	62-190	175
1995/96	23804	102-125	115	165-210	195
1996/97	17090	95-105	100	160-200	175
1997/98	22538	90-105,5	100	150-190	185
1998/99	24625	90-105	100	160-190	175
1999/00	21118	95-105	100	165-180	170

Tableau A-1 Evolution des fourchettes de prix des riziers

Année	Prix kg de tomate F.cfa (bord champ)
1995	34
2003	45

Tableau A-2 Prix au producteur de la tomate

Produit	Prix au producteur (F.cfa/kg)
Riz paddy	105
Tomate	39
Oignon	125
Maïs	100
Sorgho	100
Coton	170
Arachide	150
Patate	200
Gombo	350

Tableau A-3 Prix au producteur en l'an 2000

A.2 Dataweb de la SAED

Code XML résultant de la transformation XML du tableau A.1

```
<?xml version="1.0" encoding="UTF-8" ?>
<evolution_des_fourchettes_de_prix_des_riziers>
<saison debut="1994" fin="1995">
<collecte valeur="31000" unite="t" />
<fourchette_prix_paddy valeur="97,6-115" unite="F.cfa/kg" />
<moyenne> 100</moyenne>
```

<fourchette_riz_blanc valeur="162-190" unite="F.cfa/kg" />
<moyenne> 175</moyenne>
</saison>
<saison debut="1995" fin="1996">
<collecte valeur="23804" unite="t" />
<fourchette_prix_paddy valeur="102-125" unite="F.cfa/kg" />
<moyenne> 115</moyenne>
<fourchette_riz_blanc valeur="165-210" unite="F.cfa/kg" />
<moyenne> 195</moyenne>
</saison>
<saison debut="1996" fin="1997">
<collecte valeur="17090" unite="t" />
<fourchette_prix_paddy valeur="95-105" unite="F.cfa/kg" />
<moyenne> 100</moyenne>
<fourchette_riz_blanc valeur="160-200" unite="F.cfa/kg" />
<moyenne> 175</moyenne>
</saison>
<saison debut="1997" fin="1998">
<collecte valeur="22538" unite="t" />
<fourchette_prix_paddy valeur="90-105,5" unite="F.cfa/kg" />
<moyenne> 100</moyenne>
<fourchette_riz_blanc valeur="150-190" unite="F.cfa/kg" />
<moyenne> 185</moyenne>
</saison>
<saison debut="1998" fin="1999">
<collecte valeur="24625" unite="t" />
<fourchette_prix_paddy valeur="90-105" unite="F.cfa/kg" />
<moyenne> 100</moyenne>
<fourchette_riz_blanc valeur="160-190" unite="F.cfa/kg" />
<moyenne> 175</moyenne>
</saison>
<saison debut="1999" fin="2000">
<collecte valeur="21118" unite="t" />
<fourchette_prix_paddy valeur="95-105" unite="F.cfa/kg" />
<moyenne> 100</moyenne>
<fourchette_riz_blanc valeur="165-180" unite="F.cfa/kg" />
<moyenne> 170</moyenne>

```
</saison>
</evolution_des_fourchettes_de_prix_des_riziers>
```

Code XML résultant de la transformation XML du tableau A.2

```
<?xml version="1.0" encoding="UTF-8" ?>
<prix_au_producteur_de_la_tomate>
<saison debut="1995" fin="1995">
<prix_de_tomate_bord_champ unite="F.cfa/kg" valeur="34"/>
</saison>
<saison debut="2003" fin="2003">
<prix_de_tomate_bord_champ unite="F.cfa/kg" valeur="45"/>
</saison>
</prix_au_producteur_de_la_tomate>
```

Code XML résultant de la transformation XML du tableau A.3

```
<?xml version="1.0" encoding="UTF-8" ?>
<prix_au_producteur>
<saison debut="2000" fin="2000">
<produit>
<riz_paddy>
<prix_au_producteur valeur="105" unite="F.cfa/kg" />
</riz_paddy>
</produit>
<produit>
<tomate>
<prix_au_producteur valeur="39" unite="F.cfa/kg" />
</tomate>
</produit>
<produit>
<oignon>
<prix_au_producteur valeur="125" unite="F.cfa/kg" />
</oignon>
</produit>
<produit>
<mais>
<prix_au_producteur valeur="100" unite="F.cfa/kg" />
</mais>
```

```
</produit>
<produit>
<sorgho>
<prix_au_producteur valeur="100" unite="F.cfa/kg" />
</sorgho>
</produit>
<produit>
<coton>
<prix_au_producteur valeur="170" unite="F.cfa/kg" />
</coton>
</produit>
<produit>
<arachide>
<prix_au_producteur valeur="150" unite="F.cfa/kg" />
</arachide>
</produit>
<produit>
<patate>
<prix_au_producteur valeur="200" unite="F.cfa/kg" />
</patate>
</produit>
<produit>
<gombo>
<prix_au_producteur valeur="350" unite="F.cfa/kg" />
</gombo>
</produit>
</prix_au_producteur>
```

A.3 Ontologie Partenaire de la SAED

```
<?xml version="1.0" encoding="UTF-8"?>
 <!DOCTYPE rdf:RDF [
                    <!ENTITY rdf     'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
                    <!ENTITY rdfs    'http://www.w3.org/2000/01/rdf-schema#'>
                    <!ENTITY owl     'http://www.w3.org/2002/07/owl#'>
                    <!ENTITY xsd     'http://www.w3.org/2001/XMLSchema#'>
                  ]>
```

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:xsd ="http://www.w3.org/2001/XMLSchema#"
         xml:base ="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:owl = "http://www.w3.org/2002/07/owl#"
        xmlns:j0 = "http://www.w3.org/2003/05/"
        xmlns:ag = "http://www.fao.org/aos/SAED#">
       <rdf:Description rdf:about="http://lil.univ-littoral.fr/~sall/dataweb/schemas/subtree/">
              <dc:title>Fichier contenant les meta-donnees en OWL</dc:title>
              <dc:date>Fri Sep 12 10:12:31 GMT 2008</dc:date>
              <dc:creator>SALL Ousmane, LANI/LIL</dc:creator>
              <dc:description> Class declarations of Subject Tree Topic
(16)</dc:description>
              <dc:identifier rdf:resource="http://lil.univ-
littoral.fr/~sall/dataweb/schemas/subtree/"/>
              <dc:publisher>LANI/LIL</dc:publisher>
              <dc:format>RDF/XML</dc:format>
              <dc:language>fr</dc:language>
       </rdf:Description>
              <owl:ObjectProperty rdf:about="lil.univ-littoral.fr/~sall/dataweb#r_1">
              <rdfs:label xml:lang="fr">est une partie de</rdfs:label>
              <rdfs:comment xml:lang="fr">isUnPartOf</rdfs:comment>
              <rdfs:comment xml:lang="fr">X &lt;scope note reference&gt; Y. The scope
notes for the term X contains information on the term Y. E.g.: "foods"
&lt;scope_note_reference&gt; "feeds";</rdfs:comment>
                <owl:inverseOf>
              <owl:ObjectProperty rdf:about="http://www.fao.org/aos/agrovoc#r_2"/>
                </owl:inverseOf>
              </owl:ObjectProperty>
              <owl:ObjectProperty rdf:about="lil.univ-littoral.fr/~sall/dataweb#r_2">
              <rdfs:label xml:lang="fr">est compose de</rdfs:label>
              <rdfs:comment xml:lang="fr">isComposedBy</rdfs:comment>
```

<rdfs:comment xml:lang="fr">X <scope note reference> Y. The scope notes for the term X contains information on the term Y. E.g.: "foods" <scope_note_reference> "feeds";</rdfs:comment>
<owl:inverseOf>
<owl:ObjectProperty rdf:about="http://www.fao.org/aos/agrovoc#r_1"/>
</owl:inverseOf>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="urlLocalSource">
<rdfs:comment xml:lang="fr">chemin Xpath de l'élément XML à partir duquel a été extrait le concept</rdfs:comment>
<rdfs:domain rdf:resource="&owl;Thing"/>
<rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="isAccrocheTo">
<rdfs:comment xml:lang="fr">Indique le nombre de concept(AOS) agregateur si un concept(Local) est agrégé </rdfs:comment>
<rdfs:domain rdf:resource="&owl;Thing"/>
<rdfs:range rdf:resource="&xsd;integer"/>
</owl:DatatypeProperty>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_1">
<rdfs:label xml:lang="FR">evolution_des_fourchettes_de_prix_des_riziers</rdfs:label>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
<rdfs:subClassOf>
<owl:Restriction>
<owl:onProperty rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
<owl:someValuesFrom>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_2"/>
</owl:someValuesFrom>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

```
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_2">
<rdfs:label xml:lang="FR">ann_e</rdfs:label>
 <rdfs:subClassOf>
      <owl:Restriction>
            <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
            <owl:someValuesFrom>
                  <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_3"/>
            </owl:someValuesFrom>
      </owl:Restriction>
 </rdfs:subClassOf>
 <rdfs:subClassOf>
      <owl:Restriction>
            <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
            <owl:someValuesFrom>
                  <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_5"/>
            </owl:someValuesFrom>
      </owl:Restriction>
 </rdfs:subClassOf>
 <rdfs:subClassOf>
      <owl:Restriction>
            <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
            <owl:someValuesFrom>
                  <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_4"/>
            </owl:someValuesFrom>
      </owl:Restriction>
 </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_3">
```

<rdfs:label xml:lang="FR">collecte</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_4">
<rdfs:label xml:lang="FR">fourchette_prix_paddy_f_cfa</rdfs:label>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_5">
<rdfs:label xml:lang="FR">fourchette_riz_blanc_f_cfa</rdfs:label>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6599"/>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_6">
<rdfs:label xml:lang="FR">prix_au_producteur_de_la_tomate</rdfs:label>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7805"/>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
<rdfs:subClassOf>
<owl:Restriction>
<owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
<owl:someValuesFrom>
<owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_7"/>
</owl:someValuesFrom>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_7">
<rdfs:label xml:lang="FR">ann_e</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_8">
<rdfs:label xml:lang="FR">prix_au_producteur_en_l_an_2000</rdfs:label>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
<rdfs:subClassOf>
<owl:Restriction>

```
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_9"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_9">
        <rdfs:label xml:lang="FR">produit</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6211"/>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_13"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_18"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
```

```
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_17"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_14"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_16"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
```

```
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_15"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_12"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_10"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_19"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
```

```
        </rdfs:subClassOf>
      </owl:Class>
      <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_10">
      <rdfs:label xml:lang="FR">riz_paddy</rdfs:label>
      <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6599"/>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                <owl:someValuesFrom>
                    <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                </owl:someValuesFrom>
            </owl:Restriction>
        </rdfs:subClassOf>
      </owl:Class>
      <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_11">
      <rdfs:label xml:lang="FR">prix_au_producteur_f_cfa</rdfs:label>
      <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
      </owl:Class>
      <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_12">
      <rdfs:label xml:lang="FR">tomate</rdfs:label>
      <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7805"/>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                <owl:someValuesFrom>
                    <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                </owl:someValuesFrom>
            </owl:Restriction>
        </rdfs:subClassOf>
      </owl:Class>
```

```
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_13">
        <rdfs:label xml:lang="FR">oignon</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_12934"/>
          <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
          </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_14">
        <rdfs:label xml:lang="FR">ma_s</rdfs:label>
          <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
          </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_15">
        <rdfs:label xml:lang="FR">sorgho</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7249"/>
          <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
```

```
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_16">
        <rdfs:label xml:lang="FR">coton</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_1926"/>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_17">
        <rdfs:label xml:lang="FR">arachide</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_11368"/>
            <rdfs:subClassOf>
                <owl:Restriction>
                    <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                    <owl:someValuesFrom>
                        <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                    </owl:someValuesFrom>
                </owl:Restriction>
            </rdfs:subClassOf>
```

```
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_18">
        <rdfs:label xml:lang="FR">patate</rdfs:label>
          <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
          </rdfs:subClassOf>
        </owl:Class>
        <owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_19">
        <rdfs:label xml:lang="FR">gombo</rdfs:label>
        <rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_12920"/>
          <rdfs:subClassOf>
                <owl:Restriction>
                        <owl:onProperty
rdf:resource="http://www.fao.org/aos/agrovoc#r_261"/>
                        <owl:someValuesFrom>
                                <owl:Class rdf:about="http://lil.univ-
littoral.fr/~sall/Dataweb_SAED#c_11"/>
                        </owl:someValuesFrom>
                </owl:Restriction>
          </rdfs:subClassOf>
        </owl:Class>
        <owl:DatatypeProperty rdf:ID="nom">
        <rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_2"/>
        <rdfs:range rdf:resource="&xsd;date"/>
        </owl:DatatypeProperty>
        <owl:DatatypeProperty rdf:ID="moyenne">
        <rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_2"/>
```

```
<rdfs:range rdf:resource="&xsd;Literal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="unite_de_mesure">
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_3"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_4"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_5"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_11"/>
<rdfs:range rdf:resource="&xsd;Literal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="valeur">
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_3"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_4"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_5"/>
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_11"/>
<rdfs:range rdf:resource="&xsd;Literal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="nom">
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_7"/>
<rdfs:range rdf:resource="&xsd;Literal"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="prix_kg_de_tomate_f_cfa_bord_champ">
<rdfs:domain rdf:resource="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_7"/>
<rdfs:range rdf:resource="&xsd;Literal"/>
</owl:DatatypeProperty>
</rdf:RDF>
```

A.4 Ontologie générique de la SAED

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:xsd ="http://www.w3.org/2001/XMLSchema#"
         xml:base ="http://www.fao.org/aos/agrovoc#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:owl = "http://www.w3.org/2002/07/owl#"
        xmlns:j0 = "http://www.w3.org/2003/05/"
```

```
 xmlns:ag = "http://www.fao.org/aos/agrovoc#">

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6178">
<rdfs:label xml:lang="FR">prix</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_1">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6178">
<rdfs:label xml:lang="FR">prix</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_4">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6599">
<rdfs:label xml:lang="FR">riz</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_5">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6599"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_7805">
<rdfs:label xml:lang="FR">tomate</rdfs:label>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6178">
<rdfs:label xml:lang="FR">prix</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_6">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7805"/>
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>
```

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6178">
<rdfs:label xml:lang="FR">prix</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_8">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6211">
<rdfs:label xml:lang="FR">produit</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_9">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6211"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6599">
<rdfs:label xml:lang="FR">riz</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_10">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6599"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_6178">
<rdfs:label xml:lang="FR">prix</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_11">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_6178"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_7805">
<rdfs:label xml:lang="FR">tomate</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_12">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7805"/>

```
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_12934">
<rdfs:label xml:lang="FR">oignon</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_13">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_12934"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_7249">
<rdfs:label xml:lang="FR">sorgho</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_15">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_7249"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_1926">
<rdfs:label xml:lang="FR">coton</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_16">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_1926"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_11368">
<rdfs:label xml:lang="FR">arachide</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_17">
<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_11368"/>
</owl:Class>

<owl:Class rdf:about="http://www.fao.org/aos/agrovoc#c_12920">
<rdfs:label xml:lang="FR">gombo</rdfs:label>
</owl:Class>
<owl:Class rdf:about="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#c_19">
```

<rdfs:subClassOf rdf:resource="http://www.fao.org/aos/agrovoc#c_12920"/>
</owl:Class>
</rdf:RDF>

A.5 Base d'annotations de la SAED

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF [
<!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
<!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
<!ENTITY xsd 'http://www.w3.org/2001/XMLSchema#'>
]>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd ="http://www.w3.org/2001/XMLSchema#"
xml:base ="http://lil.univ-littoral.fr/~sall/Dataweb_SAED-instances"
xmlns ="http://lil.univ-littoral.fr/~sall/Dataweb_SAED#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:ag = "http://www.fao.org/aos/SAED#">
<c_1 rdf:ID="evolution_des_fourchettes_de_prix_des_riziers">
<urlLocalSource>

file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Evolution des fourchettes de prix des riziers .xml/evolution_des_fourchettes_de_prix_des_riziers
</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_1>
<c_2 rdf:ID="ann_e">
<urlLocalSource>

file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Evolution des fourchettes de prix des riziers .xml/evolution_des_fourchettes_de_prix_des_riziers/ann_e
</urlLocalSource>
</c_2>
<c_3 rdf:ID="collecte">

<urlLocalSource>

file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Evolution des fourchettes de prix des riziers .xml/evolution_des_fourchettes_de_prix_des_riziers/ann_e/collecte

</urlLocalSource>
</c_3>
<c_4 rdf:ID="fourchette_prix_paddy_f_cfa">
<urlLocalSource>

file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Evolution des fourchettes de prix des riziers .xml/evolution_des_fourchettes_de_prix_des_riziers/ann_e/fourchette_prix_paddy_f_cfa

</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_4>
<c_5 rdf:ID="fourchette_riz_blanc_f_cfa">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Evolution des fourchettes de prix des riziers .xml/evolution_des_fourchettes_de_prix_des_riziers/ann_e/fourchette_riz_blanc_f_cfa

</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_5>
<c_6 rdf:ID="prix_au_producteur_de_la_tomate">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur de la tomate.xml/prix_au_producteur_de_la_tomate

</urlLocalSource>
<isAccrocheTo>2</isAccrocheTo>
</c_6>
<c_7 rdf:ID="ann_e">
<urlLocalSource>

file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur de la tomate.xml/prix_au_producteur_de_la_tomate/ann_e
</urlLocalSource>
</c_7>
<c_8 rdf:ID="prix_au_producteur_en_l_an_2000">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000
</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_8>
<c_9 rdf:ID="produit">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit
</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_9>
<c_10 rdf:ID="riz_paddy">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/riz_paddy
</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_10>
<c_11 rdf:ID="prix_au_producteur_f_cfa">
<urlLocalSource>
file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/coton/prix_au_producteur_f_cfa
</urlLocalSource>
<isAccrocheTo>1</isAccrocheTo>
</c_11>
<c_12 rdf:ID="tomate">

```
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/tomate
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_12>
        <c_13 rdf:ID="oignon">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/oignon
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_13>
        <c_14 rdf:ID="ma_s">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/ma_s
            </urlLocalSource>
        </c_14>
        <c_15 rdf:ID="sorgho">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/sorgho
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_15>
        <c_16 rdf:ID="coton">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/coton
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_16>
        <c_17 rdf:ID="arachide">
```

```
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/arachide
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_17>
        <c_18 rdf:ID="patate">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/patate
            </urlLocalSource>
        </c_18>
        <c_19 rdf:ID="gombo">
            <urlLocalSource>
                file:/c:/SAED/SICArhcitecture/EntrepotsXMLPartenaires/SAED/Prix
au producteur en l'an 2000.xml/prix_au_producteur_en_l_an_2000/produit/gombo
            </urlLocalSource>
            <isAccrocheTo>1</isAccrocheTo>
        </c_19>
</rdf:RDF>
```

Résumé

Des données de nature environnementale sur la vallée du fleuve Sénégal ont collectées depuis de nombreuses années à partir des activités des différents experts y intervenant. Ces données de nature spatio-temporelle, présentent certaines particularités sémantiques et structurelles selon les partenaires. Des moyens de collecte et de stockage divers sont utilisés, induisant ainsi pour les données une dimension structurelle de l'hétérogénéité à laquelle est apparentée une dimension sémantique liée à leur description.

Afin de résoudre la problématique de l'hétérogénéité structurelle, nous avons proposé l'introduction d'une phase préalable de pré-intégration par une représentation de l'ensemble des données partenaires sous XML, constituant ainsi un *entrepôt de documents XML* dit ici *Dataweb*. Nous utilisons ensuite le vocabulaire contrôlé décrivant les données de chaque partenaire et la réutilisation d'une ontologie du domaine comme un support pour la construction d'une base de connaissances. Cette base est alors générée partir du dataweb partenaire constituant ainsi un dataweb sémantique. Ce dataweb sémantique permet ainsi l'intégration sémantique et structurelle des données de chaque partenaire.

Pour l'intégration entre les différents dataweb sémantiques nous exploitons les travaux développés sur les systèmes à base de hubs en définissant une ontologie avec les concepts que chaque partenaire désire partager et à partir desquels nous constituons une ontologie globale.

Mots-clés: Modèle de système d'intégration de donnes, modèle de dataweb sémantique, systèmes à base de hubs.

Abstract

Environmental data on Senegal River Valley have been collected for many years from activities of the various experts involved. These spatio-temporal data, present certain semantic and structural features as partners. Means of collection and storage are different uses, thus inducing data for a structural dimension of a heterogeneite which is affiliated dimension related to their semantic description with a vocabulary, but in control of each partner.

To solve the problem of structural heterogeneity, we propose the introduction of a preliminary phase of pre-integration with a representation of all partners with XML data, creating a warehouse of XML documents said Dataweb here. We then use the controlled vocabulary describing the data of each partner and reuse of a domain ontology as a support for the construction of a knowledge-based ontology is generated from the constituent partner Dataweb a semantic Dataweb. This allows Dataweb semantic integration and semantic structural data of each partner.

For integration between different semantic Dataweb we operate on the work developed based systems hubs by defining an ontology with the concepts that each partner to share from which we are a global ontology.

Keywords: Data integration systems, semantic Dataweb model, hubs based systems.