Intégration de données environnementales : une approche basée sur les entrepôts de documents XML et les ontologies

Ousmane SALL* ** Moussa LO*
* Laboratoire d'Analyse Numérique et d'Informatique – UFR SAT
Université Gaston Berger de Saint-Louis
BP 234 Saint-Louis (Sénégal)
lom@ugb.sn
** Laboratoire d'Informatique du Littoral
Université du Littoral Côte d'Opale
50, rue Ferdinand Buisson - BP 719
62228, Calais Cedex (France)
sall@lil.univ-littoral.fr

Résumé. Cet article présente l'approche que nous avons adoptée pour résoudre le problème d'intégration de données dans le contexte du projet SIC-Sénégal dont l'objectif est de permettre à plusieurs organismes partenaires de partager leurs sources de données environnementales. Nous réalisons une intégration en deux phases. Une première phase d'intégration structurelle, basée sur l'utilisation d'entrepôts de documents XML, permet de créer un entrepôt pour chaque organisme participant au projet. Une deuxième phase consiste alors à effectuer l'intégration de ces entrepôts de documents XML en associant une ontologie à chaque entrepôt. Cela se fait par une construction automatique d'ontologie OWL[1] à partir des données XML de l'entrepôt et par une réutilisation de l'ontologie AGROVOC[2].

1 Introduction

L'intégration est guidée par le besoin de regrouper des données ou documents présentant une hétérogénéité à deux niveaux : structurelle et sémantique. Du point de vue structurel, les documents pouvant exister sous plusieurs formes et les données organisées suivant des structures différentes. Du point de vue sémantique, les données portent des sens et significations différentes selon leurs identificateurs.

Nos travaux se situent dans le contexte du projet SIC[3]-Sénégal initié à l'Université Gaston Berger pour offrir des solutions aux problèmes rencontrés dans la gestion et l'exploitation des données liées à la mise en valeur de la vallée du fleuve Sénégal (BDSIC, 2004). Les données sont distribuées sur plusieurs sources hétérogènes et appartiennent souvent à des organismes différents appelés *partenaires* (dans le cadre du projet). Pour résoudre le problème d'hétérogénéité structurelle, nous proposons une approche dataweb (Lô, 2002). Cela consiste à réaliser, pour chaque partenaire, un entrepôt de documents XML intégrant les sources de données dudit partenaire.

[1] http://www.w3.org/TR/owl-features/
[2] http://www.fao.org/agrovoc/
[3] Système d'Information et de Connaissances

Le problème revient ensuite à effectuer une intégration des différents entrepôts de documents XML obtenus. L'approche que nous proposons, et qui est décrite dans cet article, consiste alors à construire une ontologie OWL (Ontology Web Language) à partir de chaque entrepôt de documents XML. L'approche est basée sur la réutilisation d'ontologie, ici l'ontologie AGROVOC de la FAO (Food and Agriculture Organization) compte tenu de la nature (environnementale) des données manipulées dans le cadre du projet SIC-Sénégal.

La section 2 présente un état de l'art sur la construction d'ontologie, en particulier à partir de documents XML. Dans la section 3, nous présentons le projet SIC-Sénégal qui constitue le cadre d'application de nos travaux ; nous insistons notamment sur l'approche entrepôt XML. La section 4 décrit la méthodologie d'intégration d'entrepôts de documents XML que nous proposons et qui s'appuie sur une construction automatique d'ontologies OWL à partir des données XML.

2 Etat de l'art

Dans le cadre du projet SIC, il est nécessaire d'un point de vue sémantique, au sein de chaque partenaire d'harmoniser et de mettre en relation les différents sens dans le vocabulaire contrôlé et hiérarchisé de chaque document XML. Notre approche utilise un processus de construction automatique des différentes ontologies à partir de la structure des documents XML.

2.1 Méthodologies de construction d'ontologie

Il existe différentes méthodologies de construction d'ontologie. Mais, généralement, c'est celle d'Uschold et Gruninger qui est utilisée, elle est constituée de quatre étapes, le lecteur peut se référer à Uschold et Gruninger (1996) pour plus de détails.

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 après coup.

Nous avons aussi des approches comme Kietz et al. (2000), Navigli et al. (2003) et 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), Biébow et al. (2000)) qui repose sur l'analyse de corpus linguistiques, et celles transformant un thésaurus en ontologie par sa migration 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[1], GermaNet[2], SemCor (Fellbaum (1998)) comme base de connaissance.

[1] http://wordnet.princeton.edu/

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

Ici nous nous intéressons uniquement à la construction automatique d'ontologies à partir de documents ou données semi-structurées par réutilisation d'ontologies. Nous n'allons pas nous étendre sur le débat consistant à défendre si oui ou non les ontologies sont réutilisables, les différents points de vue sont largement développés dans Bachimont (1996), Charlet (2002) et Furst (2004).

2.2 Construction d'ontologie à partir de documents XML

En termes d'intégration de documents XML des approches, comme celle de Klein (2002), proposent une procédure de transformation directe des données XML en données RDF en annotant les documents XML via des spécifications RDFS externes.

Il existe aussi des approches comme celles proposées par Patel-Schneider et Siméon (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 XML 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 Steve (2004). Dans Reif et al. (2005) est proposée une approche dans le contexte d'un projet nommé WESA (Web Engineering for Semantic Web Applications) pouvant être utilisée pour générer des méta-données RDF à partir de schémas de documents XML. Cette phase de construction se passe en deux étapes : d'abord, dans 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 de XML-Schema à une ontologie OWL soient générées manuellement.

Une méthodologie de mapping de XML à RDF et aussi de XML-Schema à OWL est proposée dans Matthias (2004). 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 Hannes (2004) 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 aussi, 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 mapping entre XML-Schema et OWL permet de spécifier que tout nœud terminal et attribut de nœud est mappé dans l'ontologie comme un « owl:DatatypeProperties ».

Dans Isabel (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.

3 Le projet SIC-Sénégal

3.1 Problématique

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) (voir BDSIC, 2004). 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 collections distinctes, d'un vocabulaire propre au partenaire ou partagé avec une sémantique diverse. La nécessité s'est fait 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.

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. Dans la suite de cet article, nous désignerons par *partenaire* tout organisme fournisseur de données.

3.2 Une approche entrepôt XML

Notre approche d'intégration au sein d'un partenaire est basée sur l'utilisation d'entrepôts de documents XML en nous appuyant sur des travaux précédemment menés dans le contexte, entre autres Lô (2002) et Faye et al. (2006).

Nous procédons à une phase de pré-intégration permettant une migration des données de chaque partenaire vers XML avec la mise en place d'un entrepôt de documents XML (appelé dataweb) issu des sources de données de chaque partenaire, et ensuite de construire une ontologie pour chaque partenaire à partir du dataweb.

L'approche entrepôt a été utilisée par plusieurs produits (avec des documents XML) dont les approches sont intéressantes dans le cadre du SIC ; on peut citer Xylème, Xedit, Enosys,

Tamino. On procède pour cela à une copie des données sources, ou matérialisation des données avec une migration des données vers l'entrepôt. Cette approche a plusieurs avantages, en terme de performance, avec la possibilité de procéder à des optimisations et la génération d'indexes. Les données étant stockées localement, il est possible de les organiser, annoter, personnaliser.

Un dataweb, comme dans Lô (2002), permet ici de palier à l'hétérogénéité structurelle des données partenaires en disposant d'un même format de représentation et aussi du vocabulaire contrôlé de chaque partenaire. Les données considérées sont généralement stockées dans des documents Excel et des bases de données relationnelles. Nous travaillons actuellement avec uniquement des échantillons de données provenant de tableaux comme le montre le tableau 1.

Ces tableaux ont été construits et remplis par des experts de leur domaine, donc les noms de colonnes et titre de chaque tableau véhiculent des termes du vocabulaire partenaire et aussi de manière implicite les relations entre ces dernières. XML convient particulièrement à la représentation de ces données où chaque colonne devient un nœud comme la colonne « Années » du tableau 1. Les colonnes imbriquées comme « Tonnage exporté » seront des nœuds ayant deux fils que sont ici les colonnes « Tomates séchées » et « Haricots verts ». C'est d'ailleurs l'une des raisons pour lesquelles dans le mapping de XML à OWL, chaque nœud fils et son père sont reliés par une relation de «part-Of» comme une information et ses granules. Ce tableau sera mappé automatiquement autour de son titre qui devient le nœud racine du document XML. Ceci montre que, du point de vue lexique, ces documents XML véhiculent l'idée de vocabulaire contrôlé ou d'abstraction sur le vocabulaire et d'hiérarchisation avec les relations de compositions entre les nœuds.

Pour la transformation des tableaux en XML, nous avons réalisé en Java un wrapper Excel-XML (Kasset et Niang (2006)).

Statistiques d'exportation		
Tomates séchées et haricots verts		
Années	Tonnage exporté	
	Tomates séchées	Haricots verts extra fins
1997	100	150 à 200
1998	100	
1999	100	400
2000	100	400
2001	100	400
2002	100	400

TAB.1- *Exemple d'échantillons de données partenaires.*

Notre approche consiste à construire de manière automatique une ontologie pour chaque partenaire par un mapping de XML à OWL mais aussi en se servant de l'ontologie AGROVOC pour instancier les concepts candidats qui seront extraits et l'extraction de relations sémantiques. AGROVOC de la FAO est une ontologie OWL multilingue qui présente la taxonomie d'un ensemble d'environ 36.000 concepts environnementaux couvrant

plusieurs domaines comme l'agriculture, la pêche, l'élevage, la santé, l'alimentation, la géographie,…

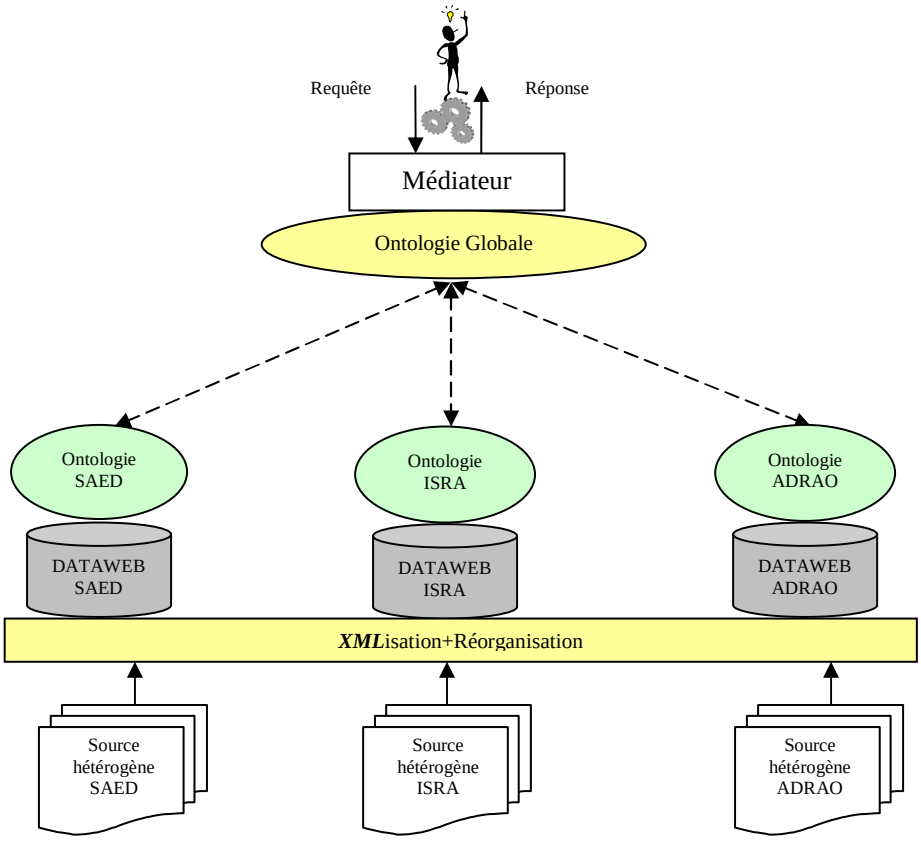

FIG. 1- Architecture générale d'intégration et de construction ascendante des ontologies

Nous utilisons donc une approche entrepôt pour l'intégration sémantique des données de chaque partenaire en associant une ontologie OWL construite à partir de la structure des documents XML. Ceci permet d'exploiter individuellement chaque dataweb de manière autonome. Pour la coopération entre les différents dataweb, nous utilisons une approche médiateur en construisant une ontologie globale issue de la fusion des différentes ontologies locales à chaque source. L'approche médiateur (Wiederhold (1995)) ou paresseuse est fondée sur la définition de mapping 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. La construction automatique se basant sur le lexique fourni par les documents XML, cela permet de les fusionner aisément, en utilisant les différents ensembles définissant la structure de chaque ontologie.

4 Intégration d'entrepôts XML par construction automatique d'ontologies OWL

Afin de permettre l'intégration 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 d'ontologie OWL à partir d'un entrepôt de documents XML par réutilisation d'ontologie. Contrairement aux autres méthodologies qui opèrent un passage vers XML-Schema, nous proposons une approche permettant d'extraire automatiquement chaque concept candidat, ses attributs et composants ainsi que les relations de cardinalité entre les concepts et leurs composantes. En plus une réutilisation de l'ontologie AGROVOC nous permet de rajouter des relations sémantiques telles que la subsomption.

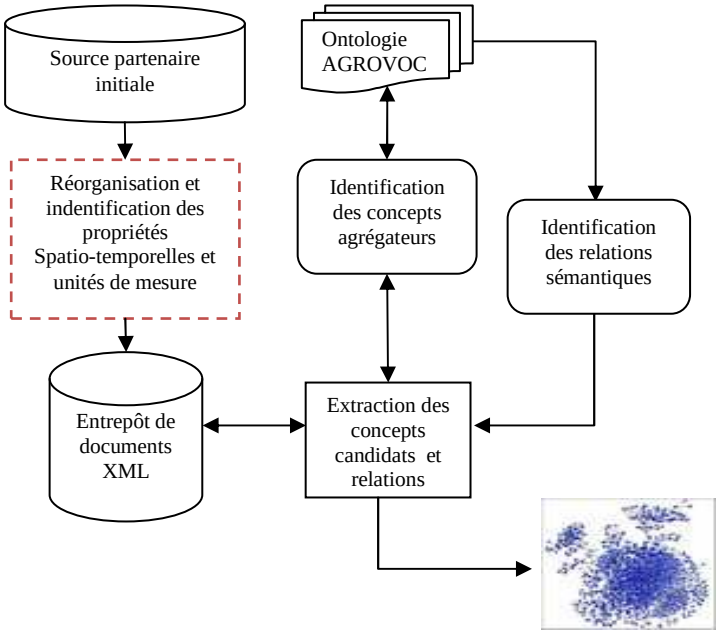

FIG. 2 - *Processus de génération d'une ontologie à partir d'un entrepôt de documents XML*

Une étude des données permet de constater qu'AGROVOC couvre quasiment l'ensemble des données manipulées dans le contexte du projet SIC-Sénégal, hormis celles pluviométriques.

XML comporte bien une couche sémantique faible, cependant elle est ambiguë (l'interprétation des données) et n'est point exploitable par la machine. De plus les relations associatives autres que hiérarchiques sont difficilement exploitables pour un utilisateur humain non expert du domaine, donc quasiment impossible à déduire par des inférences pour une machine. Cependant XML présente l'avantage pour une collection de documents représentant les connaissances du domaine d'un partenaire de constituer le vocabulaire hiérarchique et contrôlé du partenaire. Il fournit aussi de manière implicite pour l'utilisateur

humain les relations d'ordre associatives entre les termes de ce vocabulaire contrôlé et leurs propriétés. « Sémantiser » un document XML consiste donc à rendre explicite ces inférences sur l'ensemble des connaissances sur les données par une formalisation. De plus un document XML dans le contexte des données du SIC véhicule un contexte global qui constitue sa thématique (représentée par le nœud racine).

Dans notre approche nous considérons qu'un tel document 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 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 linguiste 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 agrégateur dans une ontologie. Bien sûr la bijection va d'une partie de l'ensemble des concepts de l'ontologie agrégateur de référence aux concepts candidats avec comme unique critère l'homonymie aux sens morphosyntaxiques.

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 des textes, elle est utilisée sous la forme de « fenêtre », ici notre fenêtre se résume aux composantes du nœud qui constitue 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 agrégateurs dans l'ontologie de référence.

4.1 Réorganisation des documents XML

Les données du SIC sont de nature environnementale, donc l'organisation des données après transformation des différents tableaux en XML n'est pas uniforme et nécessite une harmonisation et une mise en exergue des caractéristiques spatio-temporelles ainsi que des unités de mesures. Par réorganisation de chaque document XML, nous avons les phases d'identification des caractéristiques spatio-temporelles, ainsi que des unités de mesure et leur mise en exergue. Elles sont alors introduites comme attributs du nœud d'où elles sont extraites.

Ce traitement est important vu la nature spatio-temporelle des données du SIC, étant donné que pour chaque partenaire les données ont été recueillies dans un espace et une

échelle de temps bien définis. Donc leur prise en compte par une phase que nous nommerons *intégration spatio-temporelle* est importante. Les caractéristiques spatiales sont identifiées grâce à la consultation des occurrences des concepts géographiques sous AGROVOC ainsi qu'un dictionnaire des localités du Sénégal. Pour les unités de mesures aussi, elles sont extraites grâce à un dictionnaire. Une étude du vocabulaire nous a aussi permis d'identifier les diminutifs et acronymes dans le vocabulaire contrôlé de chaque partenaire, ce qui nous permet d'harmoniser en les remplaçant par le nom complet. Une étude des occurrences temporelles dans les noms de nœud nous a permis d'avoir l'heuristique permettant de fixer que dans l'ensemble des données partenaires, une période de temps est en réalité une saison qui chevauche sur plusieurs années successives selon le format «date début_date fin ». C'est l'exemple de la période « 1990_2001 » qui peut aussi être écrit dans le nœud sous le format «1999_01 » ; dans ce cas nous harmonisons et représentons la date de fin de la période de relevé sous la forme « 2001 ».

Des cas particuliers se sont aussi posés comme dans ceux où le nom d'un nœud est constitué seulement du nom de période. Dans ce cas nous redéfinissons le nœud en l'appelant « période_relevé » avec deux fils « date_deb » et « date_fin ». Un travail similaire est effectué pour les unités de mesures que nous extrayons grâce à un dictionnaire, il s'agit d'unités de mesure de poids comme le kilogramme (kg), d'espace comme l'hectare (ha) et de temps. Nous pouvons illustrer cette réorganisation en utilisant la partie de collection de données sur les production et rendement de riz représentés par le nœud suivant:

```
<Riz>
      <_Sup_____ha_> 28 371</_Sup_____ha_>
       <Prod__T__> 144 875 </Prod__T__>
      <_Rendt__T_ha_> 5 </_Rendt__T_ha_>
</Riz>
```

Après une phase de réorganisation nous aurons remplacé dans le document ce nœud par le nœud suivant :

```
<riz>
    <superficie valeur="28 371" unite_de_mesure="ha" />
    <production valeur="144 875" unite_de_mesure="t" />
    <rendement valeur="5" unite_de_mesure="t_ha" />
</riz>
```

Après cette phase de réorganisation, nous procédons au mapping pour la construction des différents ensembles constituant la structure de l'ontologie, puis un mapping des éléments de la structure de l'ontologie à OWL. Ce qui permet d'associer à chaque dataweb partenaire, à partir de son contenu et par réutilisation, une ontologie.

4.2 Extraction des concepts

Après la phase de réorganisation, nous procédons à l'extraction des concepts candidats, relations hiérarchiques et agrégation pour la construction de l'ontologie du partenaire. Dans la phase de pré-intégration, pour la représentation en XML, les colonnes ne contenant pas de sous-colonne vont apparaître comme des nœuds terminaux si elles ne sont pas des sous-

colonnes et comme attributs dans le cas contraire dans l'arbre ou des attributs. Aussi les colonnes des tableaux ayant des sous-colonnes seront mappées en XML comme des nœuds non terminaux et le titre de chaque tableau devient le nœud père du document XML. Dans le lexique des éléments DOM, nous avons à traiter du cas des nœuds terminaux et non terminaux ainsi que des attributs de nœuds. Nous avons défini 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 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.

Dans cette phase, nous n'avons pas usé d'outils particuliers, le langage Java que nous utilisons dispose de parseurs permettant l'extraction automatique des noms de nœuds XML.

4.3 Extraction des relations

Notre intérêt porte ici sur l'extraction automatique des relations sémantiques. Nous avons les deux types de relation au sens des propriétés OWL qui permettent de relier ici un concept candidat à un autre <owl:ObjectProperties> et la relation reliant un concept candidat à un attribut ou <owl:DatatypeProperties>. Ces relations définissent la composition au sens de la signification ou composante d'une information qui là est représentée par le concept candidat.

Nous avons les autres types de relations dites associatives entre les concepts et entre les concepts et attributs et celles dites de subsomption ou taxonomiques que nous généralisons ici au sens de l'agrégation. Ces dernières ne peuvent être extraites automatiquement à partir de la hiérarchie des documents du dataweb. Par exemple, dans le tableau 1, il n'est pas évident pour un non expert de déduire une relation entre « haricots verts extra fins » et « tomates séchées ». Pour les extraire nous allons au préalable après extraction des concepts candidats de chaque document XML les agréger ou accrocher à un concept de l'ontologie AGROVOC. Ceci permet de spécifier que le concept accroché dans AGROVOC subsume notre concept candidat qui va hériter de l'ensemble de ses relations sémantiques. Cette agrégation est réalisée pour le moment par simple matching, permettant de chercher par exemple pour un concept candidat « Riz », le concept qui porte le même nom dans AGROVOC. Il existe des cas particuliers comme « tomates séchées » ou « arachide_de_saison_chaude ». Dans ce cas nous cherchons à trouver et éliminer les mots vides et ensuite rechercher en partant de la droite de l'expression « arachide saison chaude », les groupes de mots les plus longs que nous pouvons trouver avec une occurrence dans AGROVOC. Si au moins une composante est trouvée ici comme le singleton « arachide » et le binôme « saison chaude ». Alors nous définissons une relation de restriction par la relation http://www.***fao***.org/aos/***agrovoc***#r_90 d'AGROVOC ou « Related Term » avec les concepts composant le nom du concept candidat. Si aucun concept agrégateur n'est trouvé dans AGROVOC, alors nous définissons une nouvelle classe sans relation de subsomption. Cette construction nous permet par la suite de découvrir d'autres types de relations. Nous savons par exemple que la relation de subsomption est transitive, ce qui permet de découvrir

des relations éventuelles de subsomption entre les concepts candidats en recherchant dans AGROVOC une relation de subsomption directe ou par transitivité entre leurs agrégateurs. De même, AGROVOC étant une ontologie multilingue, elle permet pour chaque concept candidat d'avoir son équivalence dans onze langues parmi lesquelles l'anglais ; cela permettant, grâce à WORDNET, de découvrir les relations supplémentaires.

Une URL de mapping est aussi générée avec le nœud ou la partie de l'arbre DOM d'où ce concept est extrait avec l'URL du document XML suivi d'un «#» suivi du nom du nœud d'où est extrait ce concept.

4.4 Nettoyage

Après la phase d'extraction des relations sémantiques, nous procédons à un nettoyage consistant à chercher dans l'architecture globale les éventuels concepts à fusionner ainsi que les classes reprises à plus de deux fois avec les mêmes noms mais avec des dimensions différentes. Pour les concepts candidats qui ont leur synonyme direct dans AGROVOC, qui est multilingue, nous nous servons de son équivalence en anglais pour chercher dans l'ontologie que nous construisons ses éventuels synonymes, homonymes et contraires, cela grâce à WORDNET**.**

5 Conclusion

Dans le contexte du SIC-Sénégal, la problématique posée consiste à offrir une plate-forme permettant à des partenaires de partager et d'interroger des sources de données hétérogènes. Il existe 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. De plus, une des contraintes posées par tous les partenaires consiste à disposer d'une exploitation individuelle et autonome de leurs données.

Nous avons présenté dans cet article une approche d'intégration des données environnementales appliquée au contexte du SIC-Sénégal. L'utilisation d'une approche entrepôt de documents XML ou dataweb nous permet de résoudre la problématique de l'hétérogénéité structurelle et l'aspect propriétaire des données au sein de chaque partenaire. Cela permet d'obtenir, pour chaque partenaire, un entrepôt de documents XML intégrant ses sources de données. Afin d'intégrer les différents entrepôts, nous associons une ontologie OWL à chaque dataweb partenaire. Les ontologies sont construites automatiquement par une extraction des concepts candidats et leurs termes ou labels à partir de la structure des documents XML. Ce qui permet d'accrocher, en utilisant des relations de mapping la structure de chaque document XML à des concepts de l'ontologie. Une agrégation des concepts candidats au niveau de l'ontologie AGROVOC nous permet, en plus des relations hiérarchiques déductibles des documents XML, d'inférer des relations sémantiques telles que la subsomption.

Un prototype de l'approche est en cours d'implémentation dans un environnement Java. Par ailleurs, nous envisageons de travailler sur des volumes de données plus importants.

Remerciements : Nous adressons nos sincères remerciements au personnel de la Direction Régionale du Développement Rural de Saint-Louis (en particulier à Mme Diop et M. Sarr) qui nous ont fourni la plupart des échantillons de données sur lesquels nous avons travaillé.

Références

Aussenac-Gilles, N., B. Biébow et S. Szulman (2000), *Modélisation du domaine par une méthode fondée sur l'analyse de corpus*, In Actes de la Conférence en Ingénierie des Connaissances (IC'2000), pp 93-103.

Bachimont, B. (1996). *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser*, Thèse d'épistémologie, Ecole Polytechnique, Paris.

BDISIC. Projet SIC-WEB Sénégal (2004). *« Compte rendu du Workshop des 10 et 11 juin 2004, Université Gaston Berger de Saint-Louis du Sénégal »*.

Biébow, B., N. Aussenac-Gilles and S. Szulman (2000). *Revisiting ontology design : a method based on corpus analysis*, In Proceedings of the 12th European Knowledge Acquisition Workshop (EKAW'00), R Dieng, O. Corby (Eds.), pp 172-188.

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.

Clark, P., J. Thompson, H. Holmbeck and L. Duncan (2000). *Exploiting a Thesaurus based Semantic Net for Knowledge-based Search*, Proc. Of IAAI-2000.

Faye, D-C., G. Nachouki et P. Valduriez (2006), *SenPeer : Un système Pair-à-Pair de médiation de données*, In Volume 4 – pages 24 à 52 – A R I M A 2006.

Fellbaum, C. (1998). Semantic network of English verbs. In Fellbaum, (ed). *WordNet: An Electronic Lexical Database*, MIT Press.

Furst, F. (2004). *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation*, Thèse de doctorat, Université de Nantes.

Hannes, B. et S. Auer (2005). *Mapping XML to OWL Ontologies*. Leipziger Informatik-Tage 2005: 147-156.

Isabel F. Cruz, Huiyong Xiao and Feihong Hsu (2004). *An Ontology-based Framework for XML Semantic Integration*. IDEAS 2004: 217-226.

Kasset, C. A et K. Niang (2006). *Développement d'un wrapper Excel-XML*, Mémoire de maîtrise informatique, Université Gaston Berger de Saint-Louis (Sénégal), juillet 2006.

Kietz J., R. Volz and A. Maedche. (2000). *A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet*, Proceedings of EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France.

Klein, Michel C. A. (2002): *Interpreting XML Documents via an RDF Schema Ontology*. DEXA Workshops 2002: 889-894.

Lakshmanan L. V. and F. Sadri(2003) *Interoperability on XML Data*. In Proceedings of the 2nd International Semantic Web Conference (ICSW'03).

Lô Moussa (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.

Maedche, A. and S. Staab.(2001). *Ontology Learning for the Semantic Web*. IEEE Intelligent Systems, vol.16, no. 2.

Matthias F., Christian Z. and D. Trastour. (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.

Miles A. J., Rogers N. and 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.

Navigli, R., Velardi P. and Gangemi A.(2003). *Ontology Learning and its application to automated terminology translation*. IEEE Intelligent Systems, vol. 18, n.1, January February 2003.

Reif, G., Harald G. and Mehdi J.(2005). *WEESA - Web Engineering for Semantic Web Applications*. In Proceedings of the 14th International World Wide Web Conference, pages 722.729, Chiba, Japan, May 2005.

Steve Battle (2004). *Round-tripping between XML and RDF*. In International Semantic Web conference (ISWC), Hiroshima, Japan, November 2004. Springer-Verlag.

Uschold, M. and Gruninger, M. (1996). *Ontologies: Principles, Methods and Applications*. Knowledge Engineering Review 11(2).

Wiederhold, G.(1995). Mediation in Information Systems. *ACM Computing Surveys*. 27(2):265-267, June 1995.

Wielinga B. J., Th. Schreiber A., Wielemaker J. and Sandberg J. A. C. (2001). *From Thesaurus to Ontology*. International Conference on Knowledge Capture, Victoria, Canada.

Summary

We present in this paper the approach we adopted to solve the data integration problem within the context of SIC-Senegal project. In this project, many partners want to share their environmental data sources. We realize an integration in two steps. First, we perform a structural integration by using XML documents repositories; an XML warehouse is created for each partner. Secondly, we integrate those XML warehouses by associating an ontology to each of them. That is done by an automatic construction of OWL ontology from XML data of the warehouse and by a re-use of ontology AGROVOC.