

XML et travail collaboratif : vers un Web sémantique

Abderrazak MKADMI¹⁻²

¹Laboratoire Paragraphe, Université Paris8, France

²Institut Supérieur de Documentation, Université de Manouba, Tunisie

amkadmi@yahoo.fr

Résumé

Cet article présente l'apport de XML, avec tous les langages dérivés à savoir XML Schema, XSLT, XPath, XQuery... dans le développement Web (pour qu'il devienne compréhensible aussi bien par les machines que par les êtres humains : « Web sémantique »), dans la recherche d'informations (à travers les métadonnées), ainsi que dans les différentes activités collaboratives qui peuvent exister entre les différents utilisateurs des ressources numériques. Cette présentation se fait à travers les travaux du laboratoire Paragraphe à l'université Paris8, en l'occurrence HyWebMap qui représente un système de création et de génération des espaces d'informations personnalisés à partir du Web.

Mots clés : XML, Web sémantique, Recherche d'information, Métadonnées, Travail collaboratif, Document numérique

Abstract

This article presents the contribution of XML, with all the derived languages such as XML Schema, XSLT, XPath, XQuery... in the Web development (so that becomes comprehensible as well by the machines by the human : "semantic Web"), in information research (through the metadata), as in the various collaborative activities which can exist between the various users of the digital resources. This presentation is done through work of the Paragraphe laboratory at the Paris8 university, in fact HyWebMap which represents a system of creation and generation of spaces of information personalized starting from the Web.

Key-Words: XML, Web semantic, Information retrieval, Metadata, Collaborative work, digital document

Introduction

Suite à un ensemble de travaux au sein du laboratoire Paragraphe à l'université Paris8 et qui ont abouti dernièrement à la conception et la réalisation de certains outils et systèmes tels que ICRS (Mkadmi, 2004), HyWebMap (Bouhaï, 2002), nous allons essayer dans cet article de parler de l'importance de la technologie XML, qui apparaît comme ressource intéressante pour tout ce qui peut constituer de véritables réservoirs de documents numériques, dans le développement des systèmes de recherche d'information et qui notamment permet d'évoluer d'un Web traditionnel à un Web sémantique. Nous allons évoquer également le lien entre cette technologie et le travail collaboratif.

1- XML : langage du Web sémantique

Le « Web sémantique » (contrairement au Web actuel qui est vu comme un Web syntaxique) est une expression attribuée à Tim Berners-Lee au sein du W3C, sous laquelle se regroupe un ensemble de travaux de recherche très variés qui se situent à des niveaux de complexité très différents (Figure 1) faisant 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. Ces travaux ont tous un objectif commun : rendre le contenu des ressources Web interprétables non seulement par l'homme mais aussi par la machine (Berners-Lee, 2001).

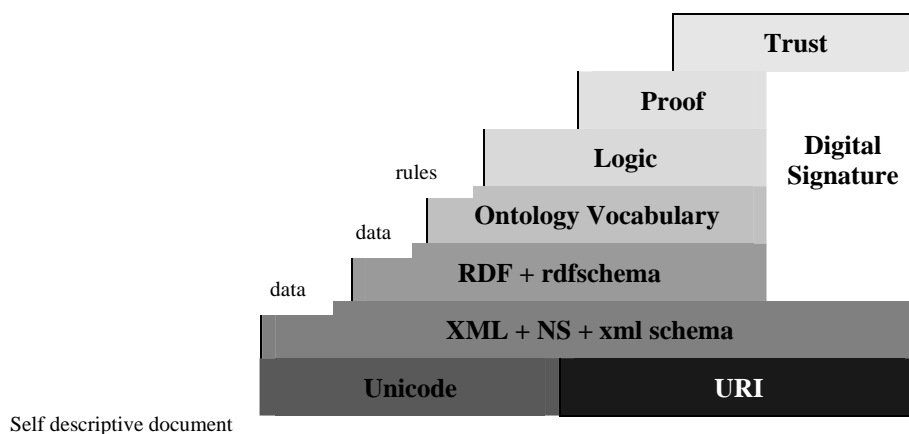


Figure 1: Les couches du Web Sémantique

1-1 RDF et RDFS

Un des points de départ de ces travaux consiste à décrire ce contenu sémantique dans des métadonnées. Pour décrire de telles métadonnées, le langage recommandé par le W3C est : RDF et le vocabulaire conceptuel sur lequel reposent ces métadonnées peut s'exprimer dans un schéma RDF (le langage RDF Schema ou RDFS).

RDF exprime le sens des balises. C'est encodé sous la forme de **triplets**, un peu comme **sujet-verbe-complément** d'une phrase très simple. Ce langage propose de définir un cadre formel de définition de métadonnées avec comme objectifs :

- rendre plus pertinent le traitement automatisé des informations contenues sur le Web, par la possibilité de fournir aux outils de traitement une information plus sémantique que les seuls mots contenus dans un document, la capacité de rendre plus "intelligente" l'information nécessaire aux moteurs de recherche et, plus généralement, nécessaire à tout outil informatique analysant de façon automatisée des pages Web ;
- fédérer les vocabulaires et syntaxes de description de métadonnées existantes dans un cadre (framework) commun :

RDF permettra de répondre précisément à une requête de type : "recherche de documents ayant Tim Berners-Lee comme auteur".

Ces triplets sont :

- une ressource (document ou extrait de document) ;
- des propriétés ;
- des valeurs (chaîne descriptive ou document).

Une ressource est définie par des propriétés. L'association d'une ressource à une propriété par une valeur de propriété est une déclaration RDF. Ainsi, il est possible de décrire, avec RDF, des phrases du type :

- "<http://www.edifrance.org>" est créé par l'"association EDIFRANCE"
- "<http://www.edifrance.org>" est la ressource ;
 - "est créé" est une propriété dont la valeur est "association EDIFRANCE".

```
<rdf :RDF xmlns:rdf = http://www.w3.org/1999/02/22-rdf-syntax-ns#>
  <rdf :description about = "http://www.edifrance.org">
    <createur>EDIFRANCE</createur>
  </rdf :description>
</rdf :RDF>
```

La syntaxe utilisée par RDF étant XML, la déclaration précédente pourra s'écrire :

Pour que la syntaxe RDF soit compréhensible, la signification doit être connue par la standardisation des propriétés et la définition des Schémas RDF. Ces derniers représentent la structure permettant de définir des concepts et leur relation. RDFS ajoute à RDF la possibilité de définir des hiérarchies de classes et de propriétés dont l'applicabilité et le domaine de valeurs peuvent être contraintes à l'aide des attributs `rdfs :domain` et `rdfs :range`. Dublin core

par exemple est un schéma RDF qui définit les propriétés : Title, Creator, Subject, Description, ...

On peut considérer les métadonnées RDF sur une ressource Web comme des annotations sémantiques sur cette ressource et le schéma RDF comme une ontologie. Plusieurs propositions étendent ces langages avec des caractéristiques des langages issus du domaine de la représentation des connaissances en les articulant avec des langages proposés pour la représentation des ontologies comme OIL, DAML ainsi que différentes propositions pour le multimédia. Le défi est donc de produire un langage qui permet aux règles en provenance de n'importe quelle "représentation du savoir" d'être exportée par le Web.

Pour résumer, nous pouvons dire que XML peut être vu comme la couche de transport syntaxique, RDF peut être considéré comme un langage relationnel de base et RDFS offre des primitives de représentation de structures ou primitives ontologiques.

1-2 Les Topic Maps

Une proposition concurrente à RDF(S) pour représenter les métadonnées pour les ressources Web est celle des Topic Maps dont un des buts originaux est de gérer et de fusionner des index de livres. Standardisés par l'ISO (International Organization for Standardization) sur la base d'une DTD SGML, les Topic Maps reposent sur trois notions fondamentales : les *Topics* qui représentent n'importe quel sujet, les *associations* qui marquent les relations entre topics et les *occurrences* qui sont des ressources disposant d'une URI (Unified Resource International) qui peuvent être liées à des topics (LAU 02). Supposons que l'on dispose du fichier *actes-h2ptm'03.pdf* qui représente la version électronique des actes de la conférence *H2PTM'03* (Laboratoire Paragraphe, Université Paris8) et du document *bioSaleh.txt*, biographie de *I. Saleh*, président de la conférence. Une façon de décrire ces ressources est de relier le fichier *actes-h2ptm'03.pdf* au concept plus général *H2PTM'03*, et le document *bioSaleh.txt* au concept de *I. Saleh*. Dans le vocabulaire des Topic Maps, les concepts *H2PTM'03* et *I. Saleh* sont des *topics*, et les ressources *actes-h2ptm'03.pdf* et *bioSaleh.txt* sont des *occurrences* de ces topics. Une *association* dans ce contexte peut être exprimée entre *I. Saleh* et *H2PTM'03* par *préside*.

Il faut signaler ici que les Topic Maps de l'ISO et RDF du W3C sont compatibles, des travaux ont été menés pour représenter des Topic Maps au format RDF et vice versa. Cependant, la notion de *scope*, qui permet de définir des contextes différents dans lesquels les

éléments nommés identiquement peuvent avoir des significations différentes, est une particularité des Topic maps qu'il est difficile de retranscrire avec RDF.

Une syntaxe XML existe depuis 2001 sous le nom XML Topic Maps (XTM). De même que pour RDF(S), des langages de requêtes existent, par exemple TMQL (Topic Maps Query Language).

1-3 XML & Document numérique

Jusqu'à présent, la recherche d'informations et l'interrogation des bases de données se sont développées de façons indépendantes. Le but d'un système de recherche d'information classique est de retrouver les documents ou plutôt les références des documents dont le contenu correspond aux termes de la requête. La structure des documents n'est prise en compte ni dans la requête pour exprimer des proximités structurelles, ni dans la réponse pour ne retourner que des parties des documents pertinentes par rapport à la requête. Par contre l'interrogation des bases de données (qui sont le résultat d'une modélisation très stricte d'un domaine d'application) a pour objectif d'extraire et de restructurer tout sous-ensemble de données utiles à la réalisation d'une tâche (Le Maître, 2004). L'usage de plus en plus de XML pour décrire des documents numériques est en train de faire évoluer cette situation en offrant la possibilité de combiner les deux approches, et ce à travers sa façon très fine de description des documents et des liens entre leurs différentes parties. C'est dans cet ordre d'idées que nous nous sommes orientés, dans notre travail, vers XML, comme langage de structuration et de balisage. Ce métalangage qui a favorisé l'expression des spécifications des standards et des normes de description, comme RDF (Resource Description Framework) (RDF, 2003), DC (Dublin Core) (Dublin, 2003), LOM (Learning Object Metadata) (LOM, 2004) ou MPEG-7 (Motion Picture Expert Group 7) (MPEG, 2003) nous a permis de créer des documents qui peuvent être traités comme une base de données intrinsèque. Ils sont auto-descriptifs, extensibles et surtout convertibles en plusieurs autres formats (HTML, XML, PDF, RTF, etc.) à l'aide des feuilles de styles, définies, elles mêmes par un langage en XML qui s'appelle XSL (eXtensible Stylsheet Language). En plus ces documents peuvent être conformes à des structures, basées elles-mêmes sur le langage XML (selon deux recommandations existantes qui sont DTD et XML Schema).

L'objectif du Web sémantique est d'augmenter donc l'efficacité des recherches d'information. Ceci en faisant évoluer les techniques d'indexation basées sur des thésaurus vers des techniques qui utilisent la représentation de connaissances et l'Intelligence Artificielle.

2- Web sémantique et travail collaboratif

Évoluer vers un Web sémantique, c'est en quelque sorte adopter des normes, des standards et des recommandations développés au sein des organismes chargés de la normalisation des technologies Web. Le W3C en est l'organisme par excellence. Toutefois, comme le Web sémantique repose sur des langages dont l'objectif est de se donner la possibilité d'enrichir le Web actuel à l'aide d'informations « sémantiques » qui facilitent la recherche et l'usage des ressources Web, l'adoption de certaines normes liées à ces langages participe à l'évidence à la facilité de description et surtout d'échange des données entre utilisateurs Web, au sens le plus large. Nous allons parler dans ce qui suit de quelques possibilités que nous avons explorées dans notre projet, de collaborations qui peuvent servir « un » Web sémantique¹.

2-1 Construction collaborative des modèles de documents

Les modèles de documents sont des modèles qui définissent la structure logique des documents, dans notre travail, nous parlons des DTD et surtout des schémas XML (Mkadmi, 2003). Ces derniers définissent les éléments utilisables dans chaque langage XML tout en exprimant les règles régissant l'assemblage de ces éléments pour construire des structures de documents valides. Cependant, ces modèles (schémas XML) ne sont intéressants que s'ils sont à la fois ouverts et partagés. Par cette conviction que, dans notre système ICRS² (Mkadmi, 2004), nous proposons un espace permettant aux différents utilisateurs de collaborer pour créer des modèles de documents et de les mettre à la disposition du grand public. Afin d'accroître la lisibilité d'un document numérique, il est nécessaire d'assister l'utilisateur dans la construction de son modèle mental. Cet espace de collaboration a été utilisé par les étudiants de maîtrise et de DESS du département hypermédia dans le cadre de leurs cours et leurs projets de fin d'études pour créer un premier modèle basé sur le langage XML Schema permettant de définir la structure logique des trois types de documents : mémoire de fin d'études et/ou rapport de stage, thèse de doctorat et article scientifique. Dans la création de ce modèle, plusieurs initiatives normatives ont été consultées à savoir la norme

¹ Nous avons bien utilisé « un » Web sémantique, parce que nous estimons que le Web sémantique n'existe pas encore, mais en l'attendant, nous allons parler plutôt des « webs sémantiques ».

² ICRS : pour Information Collaborative Retrieval System

Dublin Core et le projet BiblioML³ pour définir les éléments des métadonnées, le modèle DocBook pour définir la structure des tableaux et des figures, ainsi que la structure des paragraphes, etc. Ceci a été possible à travers les fonctionnalités de notre système qui permet d'avoir des activités collaboratives synchrones et asynchrones.

2-2 Construction collaborative de terminologie

Un des grands principes du Web sémantique est d'associer aux ressources Web des informations qui peuvent être exploitées par des agents logiciels afin de favoriser l'exploitation de ces ressources. Ces informations sont généralement ajoutées par l'auteur des ressources et parfois même par leurs gestionnaires. L'utilisateur, lui, n'est pas forcément présent dans cette opération d'indexation, malgré qu'elle soit faite pour lui. Dans ICRS, nous avons pris en compte le fait qu'une information associée à une ressource Web doit être bien structurée, descriptive et utilisable et que l'utilisateur ait un rôle primordial dans cette opération. Ceci se traduit essentiellement par les possibilités offertes par le système aux utilisateurs d'ajouter des annotations et des métadonnées aux ressources trouvées en réponse à leurs requêtes. Nous avons bien utilisé ces deux termes « annotations » et « métadonnées », malgré qu'ils n'aient pas le même sens. Ce qui nous intéresse ici c'est que ces deux termes *« prennent bien en compte cette notion d'ajout d'information à une ressource, et on pourra a priori les utiliser indifféremment pour décrire ces informations que le Web sémantique doit ajouter au Web pour le rendre plus utilisable par des machines »* (Prié, 2003).

Depuis 1994, date du fameux navigateur Mosaïc, plusieurs systèmes ont vu le jour permettant à des utilisateurs d'annoter, de lire et de partager des annotations. Quant au stockage de ces annotations, plusieurs possibilités ont été offertes : des bases de données centralisées, dans les documents eux-mêmes sous forme de balise adhoc, RDF, etc. (Denoue, 1999).

2-2-1 Recherche d'informations et métadonnées

L'utilisation d'un schéma de métadonnées dans un moteur de recherche est une des approches fondamentales, mais insuffisantes, pour évoluer vers un Web sémantique. En effet, en utilisant cette approche, le moteur de recherche propose à l'utilisateur un ensemble de descripteurs dont il peut préciser les valeurs à l'aide de mots clés. Par rapport à une recherche classique (de type google), l'utilisateur peut au moins préciser si les termes recherchés sont

³ Les formats [BiblioML](http://www.biblioml.org/) sont des représentations sous forme [XML](http://www.biblioml.org/) des formats **bibliographiques** et **autorités** préalablement définis sous [Unimarc](http://www.biblioml.org/). Ces formats sont proposés par le ministère de la Culture et de la Communication français - Mission de la recherche et de la technologie, réalisé par [AJLSM](http://www.biblioml.org/) sous la direction de Martin Sévigny. Pour plus d'informations, veuillez consulter le site : <http://www.biblioml.org/fr/>

dans le titre du document, dans son sujet, etc. et il lui est possible donc, en définissant un rôle et une sémantique aux différents termes de la requête, de cibler sa recherche et d'obtenir des résultats plus pertinents.

Le schéma de métadonnées le plus connu aujourd'hui est celui du Dublin Core. Cependant, en utilisant ce schéma, les auteurs des métadonnées, s'ils sont à peu près d'accord sur la signification de l'élément « titre » ne le sont pas forcément concernant les éléments « sujet » ou « type » d'un document. Ce qui induit des ambiguïtés sur le sens des champs. En outre, cette approche, même si elle est meilleure d'une recherche classique (qui n'utilise pas des métadonnées) n'est en fait suffisante que pour une utilisation humaine : *« même si un schéma tel que le Dublin Core définit une structure de métadonnées, l'exploitation de celle-ci par des logiciels est limitée puisque aucune sémantique – interprétable par une machine – n'est associée aux domaines de valeurs de différents champs et donc certaines ressources retrouvées pourraient ne pas convenir »* (Prié, 2003)

Nous voyons bien donc que ces schémas des métadonnées ne permettent pas de définir la sémantique opérationnelle des différents champs de description, ni celles de leurs domaines de valeurs. Ils peuvent être résumés en des standards descriptifs exprimés sous forme de DTD ou des schémas XML. Pour évoluer donc vraiment vers un Web sémantique, la solution, à notre avis, est d'utiliser des schémas des métadonnées fondées sur des ontologies.

2-2-2 Annotations sémantiques :

L'arrivée du web et des technologies associées confirme l'aspiration de V. Bush qui prônait dès 1945 la nécessité de proposer des outils interactifs pour partager l'information (Bush, 1945). Aujourd'hui de nombreux systèmes de partage d'information (Mkadmi, 2004) : cela va du Web aux outils très évolués du travail collaboratif (Lotus Notes par exemple). Toutefois, la majorité des outils visent des groupes très réduits de personnes travaillant souvent ensemble, avec des habitudes spécifiques. Par contre, le web concerne potentiellement des milliers des personnes avec des centres d'intérêts différents, des cultures et des habitudes différentes... Dans ce contexte, il est nécessaire de proposer des méthodes et des outils pour comprendre, manipuler et partager des documents.

L'annotation sémantique à partir d'ontologie semble actuellement l'approche la plus prometteuse pour partager et exploiter l'information sur le Web. Cette annotation permet d'associer des notes de lectures aux documents et de partager ainsi l'information : le lecteur devient aussi rédacteur et le système passe ainsi du « one-to-many » au « many-to-many ».

(Bringay, 2004) définit l'annotation comme étant une « *note particulière attachée à une cible. La cible peut être une collection de documents, un document, un segment de document (paragraphe, groupe de mots, mot, image ou partie d'image, etc.), une autre annotation. À une annotation correspond un contenu, matérialisé par une inscription, qui est une trace de la représentation mentale que l'annotateur se fait de la cible. Le contenu de l'annotation pourra être interprété à son tour par un autre lecteur. () l'ancre est ce qui lie l'annotation à la cible (un trait, un passage entouré, etc.)* ».

Les annotations sont multiples dans leurs formes et dans leurs fonctions, les plus courantes sont celles qui sont formalisées en langage naturel et qui sont souvent appelées « annotations libres ». Néanmoins, dans le cadre du Web, il existe un autre type particulier d'annotation appelé « annotations sémantiques » qui est utilisé dans le cadre de la recherche d'information et la classification des documents. Faisant référence à une connaissance (habituellement une ontologie) séparée du document, ces annotations sont destinées à être traitées par des machines (par opposition aux annotations libres).

Dans le cadre des annotations sémantiques, les outils permettant de choisir une ontologie, les concepts représentant les documents, ainsi que les instances des concepts sont des éditeurs d'ontologies. Les annotations sont insérées directement dans le code source du document. Ces concepts et ces instances sont soit parcourus directement par des moteurs de recherche soit utilisées pour indexer les documents.

2-2-2 Ontologies : éléments de définition

Les ontologies sont un des concepts fondamentaux du Web sémantique. Ce concept, créé au 17^{ème} siècle, a été défini comme la science de l'être. Les ontologies, une fois construites et acceptées par une communauté donnée, doivent traduire deux aspects essentiels pour permettre l'exploitation des ressources du Web par des agents logiciels : un consensus explicite et un niveau de partage.

« Les ontologies servent alors (1) pour le vocabulaire, la structuration et l'exploitation des métadonnées, (2) comme représentation pivot pour l'intégration des sources de données hétérogènes (3) pour décrire les services Web, et en général, partout où il va être nécessaire d'appuyer des modules logiciels sur des représentations sémantiques nécessitant un certain consensus » (Charlet, 2003).

Dans le domaine des ingénieries des connaissances, les ontologies sont apparues au début des années 90 pour représenter une certaine vue du monde par rapport à un domaine

particulier. Cette vue est présentée sous forme d'un ensemble de concepts. Plusieurs types d'ontologies ont été définis par les méthodes en Ingénierie des connaissances (Charlet, 2003) :

- ontologie du domaine ;
- ontologie générique : les concepts les plus abstraits du domaine ;
- ontologie d'une méthode de résolution de problèmes ;
- ontologie d'application : ontologie du domaine et ontologie de méthode ;
- ontologie de représentation : les primitives de la théorie logique.

2-2-3 Langages de définition d'ontologies

Plusieurs langages ont été définis par différents organismes pour servir de base de définition d'ontologies. Parmi ces langages, nous citons dans un ordre chronologique quelques uns qui ont marqué le domaine :

- DAML : DAML (DARPA Agent Markup Language) est un langage qui permet aux agents logiciels de comprendre et d'identifier dynamiquement la sémantique des ressources Web et de fournir une interopérabilité sémantique entre les machines;
- DAML+OIL: L'extension du langage DAML a été réalisée par l'ajout d'un autre langage OIL (Ontology Inference Layer) combinant la création d'ontologies et le marquage des informations ;
- OWL: (Ontology Web Language) est le successeur de DAML+OIL, il est utilisé pour exprimer de façon très fine les propriétés des classes définies, ainsi que les relations entre elles. Il a été fractionné en trois sous langages (qui représente chacun un langage distinct) :
- OWL lite : permet de définir une classification hiérarchique et l'expression de contraintes simples. Il ne contient d'un sous ensemble réduit des constructeurs disponibles ;
- OWL DL : contient tous les constructeurs, mais avec des contraintes particulières. Il est nommé ainsi à cause de sa correspondance avec les logiques de description. Il offre une expressivité maximale, une formalisation standard, on y trouve la notion de "domain" et de "range";
- OWL Full : très complet et sans aucune contrainte. Il se dote d'une très grande expressivité. (Baget, 2003)

Chaque langage est une extension de son précédent, l'utilisateur doit choisir celui qui correspond au mieux à ses besoins.

2-2-4 Éditeurs d'ontologies

De nombreux éditeurs existent aujourd'hui permettant de guider les utilisateurs à élaborer des ontologies suivant une méthodologie de conception plus ou moins complète. (Charlet, 2003) regroupent ces outils en deux catégories : la première fait référence aux outils les plus anciens historiquement et avec lesquels, le créateur doit se plier à un langage de représentation de connaissance donné pour définir les objets. La deuxième regroupe les outils plus récents et qui prennent en compte l'importance du niveau de connaissances. Ces outils permettent aux utilisateurs de créer des ontologies indépendamment du langage implémenté et prennent ensuite en charge la transposition de ces ontologies dans divers langages. Parmi ces outils, nous pouvons citer :

PROTÉGÉ-2000 : Il s'agit d'un outil développé par le SMI de Stanford complètement gratuit et possédant un environnement graphique permettant aux utilisateurs de créer librement leurs ontologies. De plus, cet outil, regroupant une communauté d'utilisateurs assez importante, il permet la personnalisation de ses fonctionnalités à travers des plugins qu'on peut insérer dans son interface. Dans ce outil, « *les ontologies consistent en une hiérarchie de classes qui ont des attributs (slots), qui peuvent eux-mêmes avoir certaines propriétés (facets). L'édition des listes de ces trois types d'objets se fait par l'intermédiaire de l'interface graphique, sans avoir besoin d'exprimer ce que l'on a à spécifier dans un langage formel : il suffit juste de remplir les différents formulaires correspondant à ce que l'on veut spécifier* » (Charlet, 2003).

OILED : OILED est aussi un outil gratuit développé à l'université de Manchester, mais lui il permet d'éditer des ontologies dans le langage OIL (vu ci-dessus). Il possède une interface très simple, mais qui permet de créer une hiérarchie de classes et spécifier les rôles, pareil comme avec Protégé-2000.

DOE : pour *Differential Ontology Editor*, cet outil offre une représentation graphique des concepts et des relations de l'ontologie et permet une interaction avec les hiérarchies. Son modèle de représentation est très proche de celui du langage RDFS. Il est disponible gratuitement.

ONTOEDIT : Cet outil présente les mêmes fonctionnalités que les autres outils, mais il n'est pas disponible gratuitement dans sa version complète. Une version de démonstration est disponible sur le site d'Ontoprise, la société qui le développe en collaboration avec l'AIFB Karlsruhe.

2-2-5 construction des ontologies à partir de modèles XML

Dans notre travail nous nous intéresserons en particulier à la construction d'ontologies à partir de schémas XML existants.

Les travaux menés dans ce sens sont nombreux, en effet c'est une perspective du Web sémantique qui est en pleine émergence. On s'appuiera dans ce qui suit sur le projet PICSEL (Giraldo, 2002) un système médiateur permettant l'intégration de données ayant le format de documents XML.

Les DTDs et schémas d'un document XML (Figure 2) sont exploités pour la construction semi-automatique de l'ontologie, l'approche proposée repose sur trois étapes :

1. une ontologie initiale construite manuellement en reprenant les concepts du domaine en question, on aboutit à une hiérarchie de concepts. Les DTDs sont ensuite utilisées pour définir et enrichir les classes-propriétés-relations de l'ontologie. L'idée de base est de considérer les éléments décomposables dans les DTDs comme des classes, l'élément est sélectionné quand il apparaît au moins dans une DTD, étant donné que plusieurs DTDs peuvent être représentatives du domaine. Les propriétés sont repérées par le fait qu'il s'agit de termes associés à des éléments qui ne sont jamais décomposés dans aucune des DTDs.
2. La seconde étape consiste à organiser les éléments acquis dans la première étape ;
3. la dernière étape représente les connaissances préalablement collectées et organisées dans un langage formel. De ce fait l'ontologie pourra être exploitée par le moteur de requêtes de PICSEL.

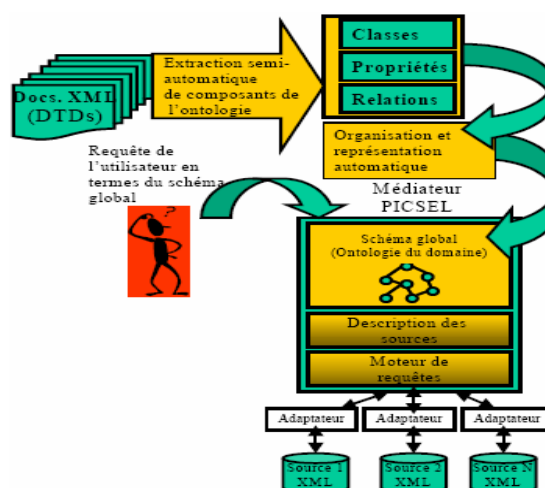


Figure 2: Construction semi-automatique d'ontologies (Giraldo, 2002)

L'intérêt majeur de l'approche réside d'une part dans l'utilisation des DTDs et des schémas XML qui représentent le langage actuel du Web et d'autre part des ontologies qui permettent une meilleure structuration et indexation des données. Cette approche est de plus en plus utilisée dans un souci de Web sémantique.

2-3 HyWebMap : vers une construction collaborative de terminologie

HyWebMap est un système de navigation et de structuration des documents sur le Web créé dans le cadre de l'élaboration et de la conception des systèmes hypermédia à vocation didactique au laboratoire Paragraphe à l'université Paris8 (Bouhaï, 2002). Il est utilisé entre autres dans le domaine de l'éducation, la veille technologique et l'enseignement à distance et constitué, suite à une étude des besoins des utilisateurs, de plusieurs niveaux (Mkadmi, 2003a). HyWebMap dispose d'une interface graphique offrant un ensemble d'éléments qui permettent l'accès aux différentes fonctionnalités du logiciel. Le réseau d'information est représenté par des nœuds. Le contenu de ces nœuds peut être un texte, une image, un graphique, une vidéo... Cet outil permet entre autres de calculer pour chaque nœud visité le nombre de liens internes, liens externes, le nombre d'images, de graphiques, de tableaux, ainsi que l'URL (s'il existe) de chaque nœud. HyWebMap permet également, de générer le contenu d'un nœud dans différents formats à savoir le HTML et le XML.

Ce système se dote aujourd'hui de plusieurs fonctionnalités que nous pouvons présenter comme suit :

- Une interface permettant à un groupe d'utilisateurs de créer des réseaux d'information à partir d'un schéma XML (figure 3) représenté sous format graphique (figure 4), à partir duquel, l'utilisateur peut créer son réseau d'information en cliquant directement sur le nœud et en remplaçant le nom de l'élément apparent par un contenu qui représentera après le nom de l'élément (XML) et après à l'aide d'une autre fonctionnalité, il peut générer tout le réseau en format XML. Les noms des nœuds représentent les noms des éléments XML et les contenus représentent les contenus des éléments.

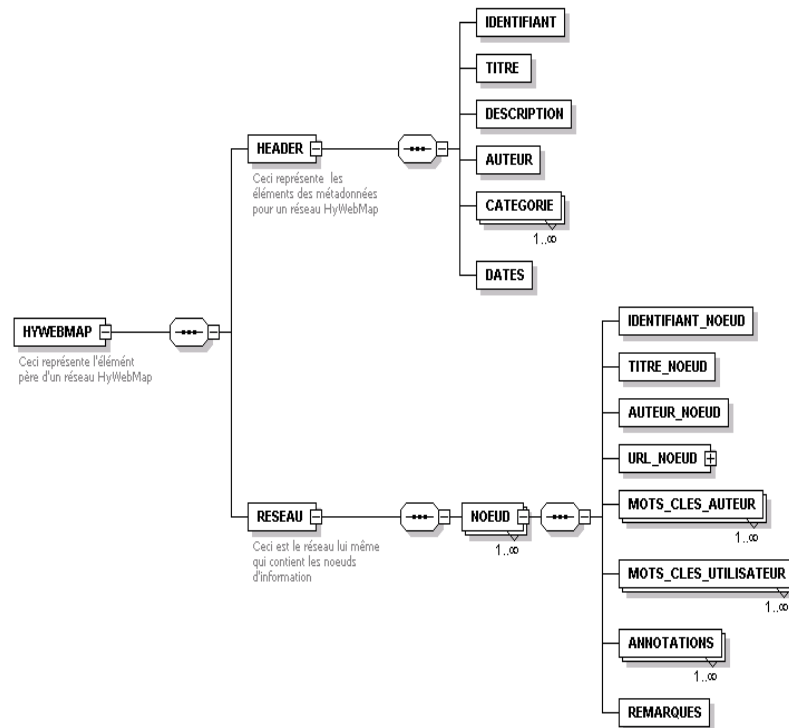


Figure 3 : Schéma XML des réseaux HyWebMap

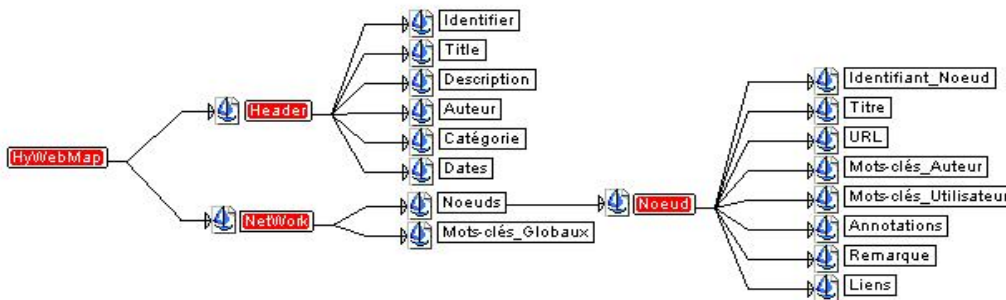


Figure 4: Schéma graphique d'un réseau HyWebMap respectant le modèle XML prédéfini

Ces deux schémas qui montrent bien la compatibilité entre l'architecture HyWebMap et XML peuvent être proposés également à un groupe d'utilisateurs de travailler en collaboration pour créer :

- des modèles de documents et/ou des réseaux d'informations personnalisés ;
- des documents et/ou des réseaux d'informations ;
- une ontologie basée sur les modèles des documents.

De plus, en utilisant XML, les documents et les réseaux d'informations (HyWebMap) sont devenus :

- plus compréhensibles : parce que toutes les balises sont devenues auto-descriptives et lisibles aussi bien par l'utilisateur que par les applications qui reçoivent ces réseaux ;

- plus échangeables entre utilisateurs : par le fait de la séparation qu'offre XML entre le contenu et la mise en forme, ce qui nous permet de stocker une seule fois le contenu et avoir plusieurs types de sorties selon les besoins de chaque utilisateur en manipulant que les feuilles de style (XSLT).

XML nous permet donc de faciliter la création, le traitement et l'accès à des réseaux d'information tout en gardant l'utilisation de l'outil HyWebMap qui simplifie, à son tour, l'utilisation de la technologie XML grâce à ses interfaces plus simples et son modèle logique de données.

Dans un souci toujours de recherche et d'échange d'informations, et vu la nature des réseaux HyWebMap créés et/ou téléchargés qui contiennent à la fois des informations hétérogènes et complexes, nous avons choisi de les stocker dans une base de donnée XML native (NXD : Native XML Database), plutôt que dans une base de données relationnelle et ce pour les raisons suivantes :

- une base de données XML native (NXD) permet de stocker les données XML sous leur forme structurée, ce qui permet ainsi d'effectuer des requêtes plus rapides et plus efficaces puisqu'on ne passe pas par une correspondance entre les tableaux et l'arborescence des données XML.
- la hiérarchie des données est ainsi préservée tout en augmentant les performances d'accès et de recherche.

Dans notre travail, le choix a été fixé sur *XINDICE*⁴, comme SGBD XML native. En utilisant cette base, nous récupérons les réseaux créés ou téléchargés du Web, nous les convertissons à un format XML, puis nous les stockons en intégralité dans la base. Quant à la recherche d'information au sein de cette base de données, nous utilisons actuellement XPath⁵ comme langage de requête et de navigation dans des documents XML.

⁴ Xindice est un logiciel libre qui fait partie du projet ApacheXML et de la DBXML Initiative qui permet de gérer une base de données composées de documents XML. Il utilise pour effectuer des requêtes le langage de requête XPath.

⁵ Si un document XML contient en même temps des données et des informations permettant d'identifier la structure et le sens de ces données, il est alors utile de pouvoir s'appuyer sur cette information pour désigner une partie d'un document XML.

C'est utile lorsque l'on réalise des applications de présentation, par exemple, pour faire une table des matières où l'on ne veut sélectionner que des titres. C'est également utile lorsqu'on veut réaliser des hyperliens sur des documents que l'on ne peut pas modifier pour leur ajouter des ancres et que l'on souhaite pourtant désigner (par exemple, le deuxième alinéa du troisième chapitre d'un document).

Du coup, il est nécessaire d'avoir un langage de désignation d'objets dans un document ; c'est l'objectif de [XPath](#). Du point de vue du W3C, l'objectif de XPath est aussi de devenir un standard de base, réutilisable dans des recommandations de plus haut niveau.

Formatant les contenus dans une forme unique, l'utilisation de « Xindice » se révélerait particulièrement intéressante pour traiter de grandes quantités d'informations non structurées et d'avoir des temps de réponse nettement améliorés

Conclusion

Il est nécessaire de signaler que le Web sémantique ne se réalisera pas sans qu'il y ait un consensus et un partage de la part des personnes et des institutions d'une même sémantique d'un domaine (Charlet, 2003a). À partir de cette réalité, nous avons essayé de présenter dans cet article quelques éléments permettant aux utilisateurs de se collaborer pour créer des ressources qui ont un sens aussi bien pour les humains que pour les machines. Cette collaboration consiste en premier lieu à créer des modèles XML à travers le répertoire de schémas XML et de les mettre à la disposition du grand public. Ensuite, à travers un espace de discussion, de créer et d'enrichir une structure d'ontologie générée à partir des schémas XML prédéfinis. Ceci représente, à notre avis, une base pour créer des Webs sémantiques en attendant l'arrivée du Web sémantique qui nécessite beaucoup plus de maturité des différents outils et des différentes recommandations liés à ce domaine.

Références bibliographiques

- Baget, J.F. ; Canaud, E. ; Euzenat, J. et Saïd-Hacid, M. (2003). Les langages du Web sémantique. (Charlet, 2003), pp. : 9-24.
- Berners-Lee, T., Hendler J., Lassila O. (2001) The semantic Web. *Scientific American*, may 2001, p35-43
- Bouhaï, N. (2002) *Lire, Réécrire et partager le savoir sur le Web : problèmes et solutions*. Thèse de Doctorat en Sciences de l'Information et de la Communication. Université Paris8. 2002.
- Bringay, S.; Barry, C.; Charlet, J. (2004). Les documents et les annotations du dossier patient hospitalier. *Information-Interaction-Intelligence*. Vol. 4, n°1, 2004, pp. 191-211
- Bush, V. (1945). As we may think. *Atlantic Monthly*.1945.
- Charlet, Jean ; Laublet, Philippe & Reynaud, Chantal. *Web sémantique : rapport final de : Action spécifique* 32 CNRS/STIC. (2003). <http://rtp-doc.enssib.fr/basedoc/rapports/ASWSRapportActiviteV2-2003.pdf>, V2, 13 Octobre 2003, consulté le 15 mars 2004.
- Charlet, J. ; Bachimont, B. ; Troncy, R. (2003a). Ontologies pour le Web sémantique. (Charlet, 2003), pp. 43-63
- Denoue, L & Vignollet, L. (1999). Yawas : un outil d'annotation pour les navigateurs du Web, *IHM'99*, Montpellier, France, 22-26 novembre 1999.
- Dublin Core (2003). <http://dublincore.org/>
- Giraldo.G et Reynaud.Ch. (2002). "Construction semi-automatique d'ontologies à partir de DTDs relatifs à un même domaine". –in : B. Bachimont, Ed., *Actes des 6^{ème} Journées Ingénierie des connaissances*, p. 53-61, Rouen, France, 2002.

- Laublet, P. ; Reynaud, C.; Charlet, J. (2003). Quelques aspects du Web sémantique. – in : *Actes des deuxièmes assises nationales du GdR I3*, 2003
- Le Maître, J. ; Muriasco, E. ; Bruno, E. (2004). Recherche d'informations dans des documents XML. – in : *Méthodes avancées pour les systèmes de recherche d'informations*. – sous la direction de M.Ihadjadene. Paris : Lavoisier, 2004. – ISBN : 2-7462-0846-6
- LOM (2004). <http://ltsc.ieee.org/wg12/index.htm>
- Mkadmi, A. (2004). *Recherche collaborative d'informations : repenser l'architecture des SRIs à l'ère numérique*. – Thèse de doctorat de l'université Paris8, 2004.
- Mkadmi, A. ; Bouhaï, N. et Langlois, M. (2003). Partager des modèles XML : Quel intérêt ?". *BBF*, N°5, Septembre, 2003
- Mkadmi, A. ; Bouhaï, N. et Saleh, I. (2003a) Vers des réseaux sémantiques d'informations. *Actes de la Conférence CoPSTIC 2003*, Rabat, 11, 12 et 13 décembre 2003, pp 204-209.
- MPEG (2003). <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>
- Prié, Y. & Garlatti, S. (2003). Métadonnées et annotations dans le Web sémantique. (*Charlet, 2003*)
- RDF (2003). <http://www.w3.org/RDF>