

Vers des réseaux d'informations sémantiques

A. MKADMI, N. BOUHAÏ, I. SALEH

Laboratoire PARAGRAPHÉ, Université Paris8,
2, rue de la Liberté 93526 - SAINT-DENIS cedex 02,

abderrazak.mkadmi@univ-paris8.fr ; nasreddine.bouhai@univ-paris8.fr ; imad.saleh@univ-paris8.fr

Résumé :

Cet article présente le contexte général dans lequel se situe la problématique de la sémantique des documents sur le Web pour qu'ils soient compréhensibles aussi bien par les hommes que par les machines. Ceci est devenu possible avec le développement de XML comme langage de structuration et d'échange de données qui a pour objectif de séparer structure et contenu de documents, ainsi que par ses possibilités de définir des métadonnées se dotant de plus de précision et de pertinence. Une description des possibilités de créer et/ou de récupérer des espaces d'informations à partir du Web, souvent peu structurés et donc peu exploitables, et de pouvoir les rendre plus structurés et plus sémantiques, en exploitant la souplesse et l'extensibilité du langage XML sera également présentée. Cette description se fait avec le système « HyWebMap ».

Mots-clés : Documents structurés, espaces d'information, Langage XML, Schémas XML, Web sémantique, bases de données XML.

Abstract :

This article presents the general context in which the problems of the semantics of the documents on the Web, so that they are comprehensible as well by the men by the machines. This became possible with the development of XML like language of structuring and exchange of data which aims to separate structure and contained documents, like by its possibilities of defining metadata being equipped of more than precision and of relevance. A description of the possibilities of creating and/or of recovering information spaces from the Web, often little structured and thus not very exploitable, and of being able to make them more structured and more semantic, by exploiting the flexibility and the extensibility of XML language will be also presented. This description is done with the system "HyWebMap".

Keywords : Structured documents, information space, XML Language, XML Schemas, Web semantic, XML Databases.

1 Introduction

Les normes et les recommandations apparues ces dernières années dans les domaines du Web, du texte, du multimédia, de la vidéo, etc., représentent les documents à la fois par leur contenu et par leur structure logique. Elles offrent des représentations des documents beaucoup plus riches que les documents ordinaires, permettant de joindre au contenu brut des informations logiques, des rapports entre éléments d'information, des métadonnées, des liens vers d'autres sources d'information, etc. Ces normes et recommandations de description ont été développées pour répondre à de nouveaux besoins des utilisateurs et permettre des accès à l'information plus sophistiqués et plus aisés, ainsi que pour offrir des possibilités d'échange et de travail collaboratif.

On commence aujourd'hui à disposer de grands corpus textuels ou multimédias qui ont été conçus ou convertis suivant ces descriptions, par contre, il existe très peu d'outils et de moyens capables de prendre en compte ces nouvelles représentations de l'information et d'exploiter leur richesse.

Dans ce contexte et dans le cadre de l'enseignement universitaire, l'équipe du laboratoire Paragraphe a développé le système HyWebMap¹, qui peut être défini comme étant un système de navigation, de structuration et de création des espaces de navigation personnalisés². Par HyWebMap, nous répondons à trois problématiques, la première est liée aux problèmes connus des systèmes hypertextes (désorientation et surcharge cognitive), la deuxième est celle de l'écriture hypertextuelle collaborative et la troisième que nous allons la développer dans cet article a un lien avec la recherche et l'échange de l'information. L'objectif étant de structurer d'une part les documents sur le Web, souvent peu structurés et de leur donner plus de sens, tout en répondant aux besoins plus exigeants en matière de recherche, de navigation, et d'appropriation des documents sur le Web, et d'autre part d'exploiter la technologie XML pour pouvoir évoluer vers un Web où les informations seront compréhensibles aussi bien par l'être humain que par la machine.

2 HyWebMap : Description

¹ Pour plus d'informations sur HyWebMap (Hypertext Web Mapping), veuillez consulter le site : <http://h2ptm.univ-paris8.fr/hywebmap/docs/>

² Un espace de navigation personnalisé est un espace d'information dont on a modifié le parcours de navigation, renommé les liens entre les documents, annoté les documents et ajouté éventuellement des descripteurs et des mots clés.

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. 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 cinq niveaux :

- 1- un agent de veille sur le réseau virtuel, ce qui permet de faire des mises à jours permanentes des nœuds³ composant le document, tout en avisant les utilisateurs ;
- 2- un outil d'appropriation des nœuds qui permet à l'utilisateur de créer son propre parcours en renommant les nœuds et liens d'un document ;
- 3- un outil d'aide à la navigation qui offre au lecteur les possibilités de recherche des documents par date, par niveau d'arborescence, par historique, etc. ;
- 4- un outil de génération des réseaux virtuels créés ou téléchargés en format HTML, ce qui facilite l'affichage sur des navigateurs classiques ;
- 5- et enfin un outil de personnalisation des documents à travers les annotations, les commentaires et les liens personnalisés.

HyWebMap dispose d'une interface graphique [figure1] offrant un ensemble d'éléments qui permettent l'accès aux différentes fonctionnalités du logiciel.

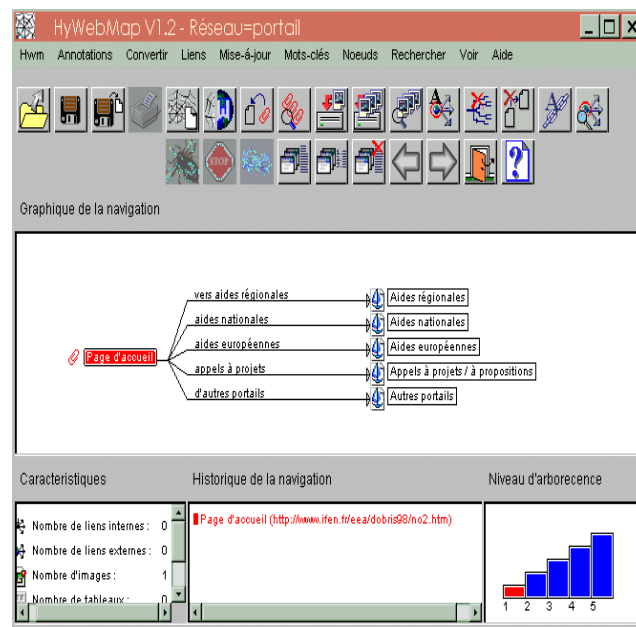


Figure 1 : Interface principale de HyWebMap [Bouhai02]

Ce système répond par l'ensemble de ses fonctionnalités à une double dimension [Pappy01] :

³ On appelle *nœud* (*node*) de l'hypertexte la représentation du document dans la mémoire logique du système.

- une dimension lecteur assurée par les agents Web en allant chercher l'information demandée par l'utilisateur et la rassembler dans l'espace informationnel de l'utilisateur ;

- une dimension auteur assurée par toutes les opérations que l'utilisateur puisse appliquer sur l'information collectée, à savoir l'ajout de nouveaux nœuds et de nouveaux liens, l'attribution des mots clés et des annotations aux nœuds d'informations, etc.

La figure 2 présente la version HTML générée à partir de la version HyWebMap créée ou téléchargée par l'utilisateur.

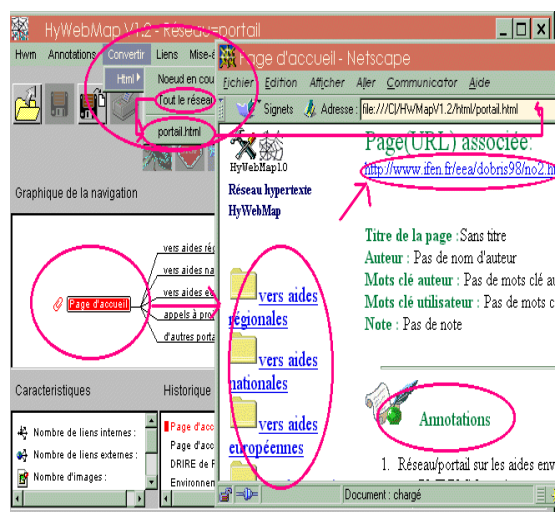


Figure 2: Interface de génération d'une version HTML à partir d'un réseau HyWebMap

3 Structuration de l'information : Documents Réticulaires Structurés (DRS)

Au-delà des notions avérées que sont les Documents Virtuels Personnalisables (DVP) et qui « peuvent être considérés comme des ensembles d'éléments (que l'on peut appeler fragments) associés à des mécanismes de filtrage, d'organisation et d'assemblage sous contraintes ... en respectant un modèle de l'utilisateur et des principes narratifs » [Garlatti99] ou les Objets Pédagogiques [Bourdoncle00], nous assistons à l'apparition d'un nouveau type de « documents-réseau », nommé « documents réticulaires structurés » [Papy01], un type de documents situé à la convergence du monde des documents structurés et des réseaux hypertextes [Clement95] dont les sites Web sont la plus récente illustration.

Des documents qui sont l'objet résultant de la construction des espaces de connaissances sur le Web [Bouhaï02]. Les « Documents Réticulaires Structurés » sont à considérer comme des entités

atomiques - au même titre qu'une page HTML - mais néanmoins constitués de multiples fragments qui, distinctement ne peuvent prétendre à une quelconque autonomie mais concourent à donner une cohérence structurelle et sémantique aux documents auxquels ils sont intégrés. Une nouvelle réalité documentaire très largement liée à l'existence de la Toile.

Il s'agit de mettre en évidence un type de document en phase avec des pratiques documentaires inédites depuis que le Web est devenu un lieu de recherche d'informations valides, pertinentes et pérennes comme le souligne [Dkaki99] « ... Il est inutile de préciser que le Web est devenu une source incontournable et incontournable d'informations à haut potentiel stratégique. Rappelons toutefois les grandes méthodes d'utilisation des informations du Web qui restent largement insuffisantes pour une exploitation optimale et sont essentiellement liées à des besoins d'accès à l'information... ».

Dans la réalité, ce constat est indéniablement étayé par la multiplication des systèmes de filtrage [Ben Hamadou01] tendant à séparer le bon grain de l'ivraie et des moteurs de recherches inédits comme NeuroWeb [Lelu99] qui puisent leur matière informationnelle directement à partir du Web. De là découlent des pratiques spécifiques inhérentes à la lecture de l'information électronique [Morizio99] et à sa « préhension cognitive » par les usagers [Hudon01, Dufour01]. Par ailleurs, on constate que les pratiques de recherche d'informations sur le Web ne sont pas toujours tournées vers la recherche de « LA » réponse [Feret01] mais plus fréquemment vers la constitution de références documentaires convergentes que les dispositifs de veille ou les portails ont très largement mis en évidence.

3 Des « documents réticulaires structurés » à des « espaces d'informations sémantiques »

Pour rendre ces « documents réticulaires structurés » plus sémantiques, les évolutions que nous avons mises en oeuvre couvrent notamment la structure, la portabilité, l'échange et l'accès aux documents. Bref, elles tendent d'évoluer d'un Web syntaxique où seul un être humain peut réellement utiliser l'information disponible vers un Web sémantique [Berners-Lee01] où l'information a une signification explicite permettant de donner aux machines un rôle de traitement, de description et de structuration beaucoup plus actif qui dépasse largement la fourniture des outils de localisation, de transfert, de mise en forme et de présentation.

Cette évolution, s'appuie sur trois concepts :

1. une représentation plus riche, plus structurée, plus rigoureuse de l'information elle-même, condition première pour que des programmes puissent agir sur le contenu,
2. un modèle de données permettant de valider tous les réseaux d'information créés ou téléchargés ;
3. des métadonnées, c'est-à-dire une description externe bien formalisée de l'information principale.

3-1 Un modèle de documents en XML ?

Il ne s'agit pas d'une obligation, mais dans la mesure où XML est utilisé comme un langage d'échange de données, disposer d'un modèle de documents paraît très essentiel pour que les données échangées soient comprises. Pour créer ce modèle, nous avons le choix entre deux recommandations qui sont disponibles aujourd'hui : Les DTD et les schémas XML. La première sur les DTD est historique et est devenue reconnue dans tous les domaines applicatifs. La deuxième est récente et vient palier aux déficiences de la première se rapportant notamment au typage de données, au langage utilisé et au support des espaces de noms [Mkadmi03].

Nous avons donc choisi les schémas XML pour construire notre modèle pour les raisons que nous venons de citer et qui sont :

- Les schémas XML utilisent le langage XML, ce qui facilite la manipulation des données avec les mêmes outils de manipulation de documents XML ;
- Les schémas XML contiennent un très grand nombre de types de données intégrées comme les booléens, les entiers, les intervalles de temps, etc. De plus, il est possible de créer de nouveaux types par ajout de contraintes sur un type existant ;
- Les schémas XML supportent les espaces de nom, ce qui permet d'éviter la confusion entre éléments provenant de diverses sources ;
- Etc.

Le modèle de documents [figure3] que nous avons créé sert donc à définir la cohérence de l'ensemble de documents, lesquels peuvent être utilisés par n'importe quelle application informatique en ne se définissant que par rapport au modèle sous-tendu. Il s'agit donc d'une grammaire permettant de vérifier la conformité du document XML. Cette grammaire contient les noms des balises que nous devons ou pouvons utiliser dans un document XML, et leurs imbrications possibles. Elle contient aussi les types de données contenus dans ces éléments. Ceci

permet évidemment de gagner beaucoup de temps, d'argent et de fiabilité dans les travaux coopératifs.

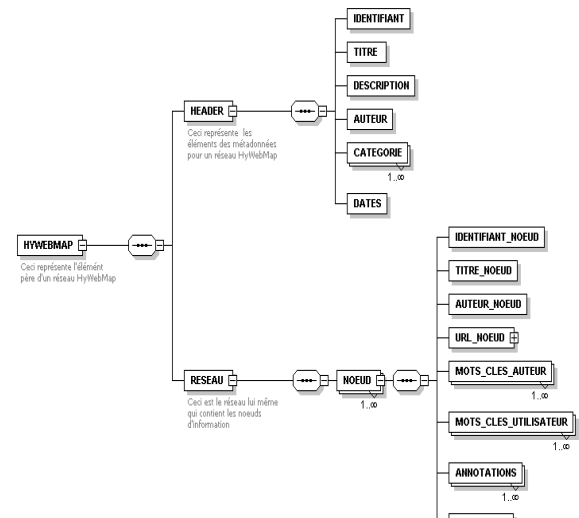


Figure 3 : Schéma XML des réseaux HyWebMap

Ce schéma XML permet de valider tous les réseaux HyWebMap créés ou téléchargés. Ceci est valable pour assurer l'homogénéité des documents et en même temps pour permettre à plusieurs utilisateurs de travailler ensemble en se basant sur une structure prédéfinie.

Comme le montre le schéma précédent, un réseau HyWebMap est constitué de deux parties :

- la première permet de définir les métadonnées du réseau (Titre, auteur, description, catégories, dates de création, de modification et de lecture du réseau...)
- la deuxième partie permet de présenter le contenu du réseau (les nœuds). Chaque nœud a un titre, un auteur, une URL, des mots clés et un espace d'annotations et de remarques.

3-2 Métadonnées des réseaux d'information

Une métadonnée est définie souvent comme étant une donnée sur une donnée. Il s'agit plus précisément, d'un ensemble structuré d'informations décrivant une ressource quelconque. Les éléments des métadonnées que nous avons définis [figure 4] sont :

- Identifiant : c'est un identifiant du réseau qui doit être attribué automatiquement par le système ;
- Titre : Titre du réseau ;
- Description : Un résumé décrivant le réseau ;
- Auteur : auteur du réseau ;
- Catégorie : thème plus général du réseau ;

- Dates : Cet élément comporte trois dates pour chaque réseau : date de création, date de modification et date de lecture.

```
<HYWEBMAP>
<HEADER>
<IDENTIFIANT>1</IDENTIFIANT>
<TITRE>HyWebMap Test</TITRE>
<DESCRIPTION>Ce réseau représente un dossier assez général sur XML et l'informatique</DESCRIPTION>
<AUTEUR>Abderrazak MKADMI</AUTEUR>
<CATEGORIE>Informatique et structuration des données</CATEGORIE>
<DATES CREATION="2002-12-17" MODIFICATION="2003-01-10" LECTURE="2003-01-10"/>
</HEADER>
</HYWEBMAP>
```

Figure 4 : Métadonnées des réseaux d'informations

3-3 Conversion des documents HyWebMap en XML

Après une réflexion faite au niveau de l'architecture de HyWebMap qui ressemble en fin de compte à un schéma XML, nous avons développé une interface permettant à des utilisateurs de créer des réseaux d'information à partir d'un schéma XML représenté graphiquement sous format HyWebMap. Le même schéma présenté ci-dessus a été présenté sous un format permettant le renseignement de tous les champs manuellement (Figure5).

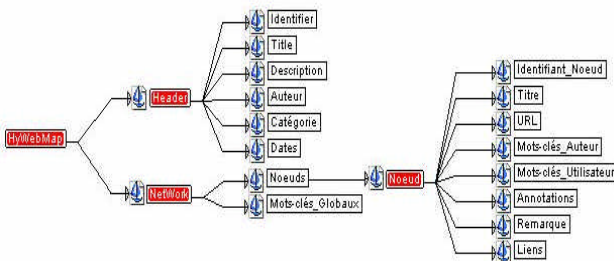


Figure 5 : Réseau HyWebMap à partir d'un schéma XML

À partir de ce graphique, 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 (ex. Figure4).

3-4 Stockage des réseaux XML

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 (Xindice), plutôt que dans une base de données relationnelle et ce pour les raisons suivantes :

- Xindice (base de données XML native open source) 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.

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 XPath comme langage de requête et de navigation dans des documents XML.

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

3-5 Apport de XML pour les réseaux HyWebMap

En utilisant XML comme format de description des documents, les réseaux HyWebMap sont devenus ainsi plus compréhensibles et plus échangeables entre utilisateurs. 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, et plus échangeables 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.

4 Conclusion et perspectives

Au terme de ce parcours rapide des voies ouvertes par l'adoption progressive de XML par nos travaux au sein du laboratoire Paragraphe, il apparaît clairement que les enjeux de cette adoption vont bien au-delà de la publication des réseaux et des documents électroniques.

XML, conçu pour expliciter et valider la structure des documents au fil même du contenu de ces documents par un balisage flexible et peu contraignant, nous permet de rendre nos réseaux d'information beaucoup plus expressifs et beaucoup plus sémantiques. Il nous permet aussi et surtout d'échanger des documents et des données avec d'autres utilisateurs sans aucune contrainte.

L'ajout d'une couche XML au système HyWebMap nous permet d'avoir une base de documents flexibles, extensibles et inter opérables. Il permet aussi de développer un travail collaboratif entre différents utilisateurs et différents concepteurs de sites Web, et d'encourager d'autres à utiliser la technologie XML et d'en profiter de sa souplesse.

Notre prochaine étape consistera à faire le lien entre ces espaces d'informations sémantiques dans le moteur de recherche « K-Web Organizer » créé lui aussi par l'équipe du Laboratoire Paragraphe et qui représente un système de supervision de travail collaboratif qui permet de créer entre autres des communautés d'« utilisateurs – lecteurs - auteurs » fédérés autour de ressources documentaires thématiques.

5 Références bibliographiques

[Anrabassan00] ANBARASAN, E. – Tim Berners-Lee : « j'ai fait un rêve ». – in : Courrier de l'UNESCO, 09-2000.

http://www.unesco.org/courier/2000_09/fr/dires.htm.

[Ben Hamadou01] BEN HAMADOU A. - Élaboration d'outils pour le filtrage d'informations et génération automatique de résumés. – in CIDE 2001, Toulouse, 23-25 octobre 2001.

[Beers-Lee 01] BERNERS-LEE T., HENDLER J., LASSILA O. - The Semantic Web. – in : *Scientific American*, May 2001, p35-43.

[Bouhaï02] BOUHAÏ, N. – 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

[Bourdoncle00] BOURDONCLE F., BERTIN P. - Recherche d'aiguilles dans une botte de liens, La recherche, n°328, février 2000.

[Clement95] CLEMENT J. – Du texte à l'hypertexte : vers une épistémologie de la discursivité hypertextuelle. – in : « Hypertextes et hypermédiat : Réalisations, outils, méthodes », Hermès, Paris, 1995.

[Dkaki99] DKAKI T. – Collecte, prétraitement et traitement des informations issues du Web dans un environnement coopératif. – in « Solaris », n°5, Presses Universitaires de Rennes, 1999.

[Dufour01] DUFOUR C., BERGERON P. – Lorsque systèmes d'information Web et professionnels de l'information se rencontrent. – in « Proc. of the 29th Annual Conference of the Canadian Association for Information Science, 27-29 May, 2001, Québec.

[Feret01] FERRET O., GRAU B., HURAUT-PLANTET M., ILLOUZ G., JACQUEMIN C. – Comment trouver « LA » réponse. – in « 3^{ème} congrès du chapitre français de l'ISKO, 5-6 juillet 2001, Paris, pp. 159-168.

[Garlatti99] GARLATTI S., IKSAI S. - Documents virtuels personnalisables pour des systèmes d'informations en ligne, IHM99, Montpellier, 1999.

[Hudon01] HUDON M. – Structuration du savoir et organisation des collections dans les répertoires du Web. – in : BBF, Paris, T. 46, n°1, 2001, pp. 57-62.

[Lelu99] LELU A., HALLAB M., RHISSASSI H., BOUYAHI S., BOUHAÏ N., HE H., QI C., SALEH I. – Projet NeuroWeb : un moteur de recherche multilingue et cartographique. – in : H2PTM'99, 23-24 septembre 1999, Paris.

[Mkadmi03] MKADMI A., BOUHAÏ N., LANGLOIS M. – Partage des modèles XML : une solution pour les échanges électroniques professionnels, Conférence « Journées Francophones de la Toile » (JFT'2003), Tours, 30 juin, 1 & 2 juillet 2003, École Polytechnique de l'Université de Tours.

[Morizio99] MORIZIO C. – Zapper, chercher, lire des documents électroniques. – in : BBF, T44, n°5, 1999, pp.48-51.

[Papy01b] PAPY F., SALEH I., BOUHAÏ N. – Chercher et réorganiser l'information sur le Web. – in : Colloque : «Hypermédias et Apprentissages», 9-11 avril 2001, Grenoble.

[Urso99] URSO, Patrick ; FAURE, Jérôme. - Le XML pour structurer la recherche d'information. - in *Technologies Internationales*, n° 54, mai 1999.