HAL
open science

Modélisation, interrogation et génération de documents sur le Web

Mourad Ouziri

To cite this version:

Mourad Ouziri. Modélisation, interrogation et génération de documents sur le Web. Revue nationale Konex, 2003, 1 (1). sic_00000389

HAL Id: sic_00000389

https://archivesic.ccsd.cnrs.fr/sic_00000389

Submitted on 18 Feb 2003

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modélisation, interrogation et génération de documents sur le Web

Mourad Ouziri
Laboratoire d'Ingénierie des Systèmes d'Information
INSA – Lyon
7 av Jean Capelle 69621 Villeurbanne Cedex
mail : mouziri@lisi.insa-lyon.fr

Résumé

Nous proposons dans ce papier de modéliser les documents du Web en utilisant un formalisme de représentation de connaissances sur le Web appelé les TopicMaps. Ce formalisme permet de représenter à la fois les données, leurs structures ainsi que la sémantique dans une base de connaissances. Ces informations peuvent être explorées par navigation et/ou requêtes dans une interface adaptée et dynamiquement adaptative. Cette interface est conviviale et ne présente à l'utilisateur que les informations pertinentes ce qui facilite la tâche d'exploration. A la fin de l'exploration, un document personnalisé contenant toutes les informations requises est généré automatiquement pour l'utilisateur final.

1 Introduction

Le Web [Bern-94] représente une source de données et d'informations interrogée par un grand nombre d'internautes. Les requérants du Web sont de profils très variés et ont donc des objectifs différents. Dans ce contexte, le document joue un rôle central. D'une part, il constitue le support d'information le plus familier aux internautes et représente, d'autre part, le format standard d'échange de données sur le Web.

Trois types de sources de données peuvent être considérés pour le Web : les sources de données fortement structurées, les sources de données semi-structurées et les sources de données non structurées. La première classe est représentée par les bases de données (relationnelles ou objets). Les

données sont créées conformément à un schéma bien défini pendant la phase de conception. Une source de données semi-structurées ne possède pas un schéma de données explicitement représenté dans la source. Le schéma de données est implicite et est véhiculé avec les données. Par conséquent, une même entité peut avoir plusieurs structures différentes dans une même source de données. Ce type de sources de données est généralement représenté en utilisant le format XML. La troisième catégorie est celle des données non structurées. Une source de données non structurée ne présente aucune notion de schéma de données. Ces sources représentent l'information en utilisant soit des images soit des phrases exprimées en langue naturelle. Ces sources sont représentées sur le Web en utilisant le format HTML. Elles peuvent être transformées, en utilisant des techniques d'indexation et d'annotation, en sources de données semi-structurées. C'est pour cela que nous ne considérons dans nos travaux que les deux premiers types de sources.

Ces hétérogénéités rendent difficile la conception de langages d'interrogation des données du Web. En effet, un langage de requêtes est basé sur un modèle de données sur lequel sont définis ses opérateurs algébriques et ses types (abstraits) de données. Or, la communauté du Web n'a encore pas trouvé un consensus sur un modèle de données unique, fiable et performant pour représenter des sources de données hétérogènes.

Nous proposons dans ce papier un modèle de données pour modéliser les données sur le Web et son langage d'interrogation sous-jacent. Nous proposons d'utiliser un modèle assez riche afin de représenter les données, leurs structures ainsi que la sémantique portée. Nous avons, pour cela, utilisé les TopicMaps qui représentent un formalisme issu de l'intelligence artificielle pour la représentation des connaissances. Nous avons étendu et adapté les TopicMaps pour faciliter l'interrogation. Etant donné que la navigation hypertexte est la façon la plus naturelle pour rechercher l'information sur le Web, nous avons proposé une interface dynamiquement adaptative pour l'interrogation des données par navigation, requêtes et mots clés. Le résultat d'une session d'interrogation est un document XML

représentant les informations demandées par le requérant et qui est ensuite transformé dans un format compréhensible par l'utilisateur final.

Nous ne présentons pas dans ce papier un système permettant de rechercher tout type d'informations pour n'importe quel utilisateur mais un système d'interrogation de données du Web appartenant à un domaine particulier (médical, judiciaire, etc.) pour des utilisateurs de ce domaine. En effet, le système conçu permet à des utilisateurs (représentés dans le système par leurs profils) de naviguer dans une interface adaptative pour interroger des données multi-sources relatives au domaine d'application.

2 Etat de l'art

Durant cette dernière décennie, plusieurs groupes de chercheurs ont mené leurs travaux dans le but de développer un modèle de données pour les données sur le Web et un langage d'interrogation. Au départ, le Web est considéré comme une collection de documents HTML liés par des liens hypertextes. En partant de ce modèle de données, deux types d'interrogation peuvent être cités. Le premier est représenté par les moteurs de recherche tels que Google et WebCrawler. Ces moteurs de recherche utilisent des index construits par des agents appelés Robot ou Spider [Hein-96] [Kost-95]. Un Robot est un programme qui parcourt automatiquement tous les documents HTML du Web en suivant les liens hypertextes et indexe ces documents par des mots-clés qu'il extrait à partir de sections HTML définies a priori (balises Title, Meta, H1, etc.). Ces index sont utilisés par les moteurs de recherche pour rechercher les documents pertinents par rapport à une expression de mots-clés. La deuxième façon d'interroger le Web consiste à utiliser les langages de requêtes déclaratifs de type SQL ou OQL[Aroc-98] [Mend-97]. WebSQL [Mend-97] modélise le Web comme une base de données relationnelle virtuelle composée de deux tables qui représentent les documents du Web ainsi que leurs liens. WebOQL [Aroc-98] modélise le Web par un graphe ayant deux types d'arcs: les arcs internes qui représentent les liens internes des documents et les arcs externes qui représentent les liens entre des documents différents. Le problème de ces systèmes est qu'ils ne prennent pas en considération la sémantique des données du Web véhiculée par les documents. Plusieurs

travaux ont été menés dans ce sens afin de concevoir des systèmes qui effectuent la recherche d'information non seulement suivant les liens structurels mais aussi qui tiennent compte du sens véhiculé par les données. Ces travaux s'inscrivent dans le cadre du Web sémantique [Koiv-01]. Pour représenter la sémantique, les documents du Web sont annotés en utilisant des concepts d'une ontologie ou d'un thésaurus. Les documents sont donc interrogés et parcourus suivant ces annotations [Habr-99] [Khei-95]. Les chercheurs de l'université de Stanford ont développé un SGBD pour le Web appelé Lore [McHu-97]. Dans ce système, le Web est considéré comme une grande base de données semi-structurées représentée par un graphe appelé OEM (Object Exchange Model) [Papa-95]. Les données stockées dans Lore sont interrogées en utilisant le langage Lorel [Abit-97]. Depuis que XML est devenu un standard pour la représentation et notamment pour l'échange de données sur le Web, plusieurs recherches ont été entreprises pour modéliser [Flor-99] [Shan-99], interroger [Deut-99] [Robi-98] et visualiser les données XML [Xform-02].

Les interfaces de navigation sur le Web deviennent de plus en plus intelligentes dans le sens où elles s'adaptent automatiquement par rapport à l'utilisateur. L'élément fondamental de ces interfaces est le profil de l'utilisateur [Brus-96] [Lang-99] (et parfois le modèle de tâches [Garl-99]). Le profil de l'utilisateur est construit soit, a priori, en demandant à l'utilisateur de répondre à un questionnaire qui définit ses préférences soit se construit d'une façon incrémentale pendant le processus de navigation. Généralement, ces modèles sont utilisés pour filtrer les informations (documents et liens) non pertinentes trouvées lors de la recherche d'informations.

3 Navigation et interrogation de données

3.1 Architecture du système

Nous décrivons dans cette section l'interface du système d'accès par navigation/requêtes aux données via le Web. A l'inverse des moteurs de recherche, le système interroge non seulement des documents mais aussi les bases de données sur le Web. Dans notre conception, nous ne

considérons que les documents XML et nous supposons que tout autre type de documents (texte, HTML, image, vidéo, etc.) peut être transformé en XML en utilisant les techniques d'annotation et d'indexation.

La première étape consiste à conceptualiser (abstraction des données) les sources de données afin d'extraire le modèle conceptuel des données. Dans une deuxième étape, chaque source est décrite dans une base de connaissances (BC) TopicMap (TM) [ISO-99]. Ces BC sont ensuite fusionnées en un TM global. L'interface de navigation/interrogation est construite à partir du TM fusionné. L'utilisateur peut ainsi naviguer et interroger les données multi-sources via cette interface sans se soucier des structures des données ni de leur emplacement.

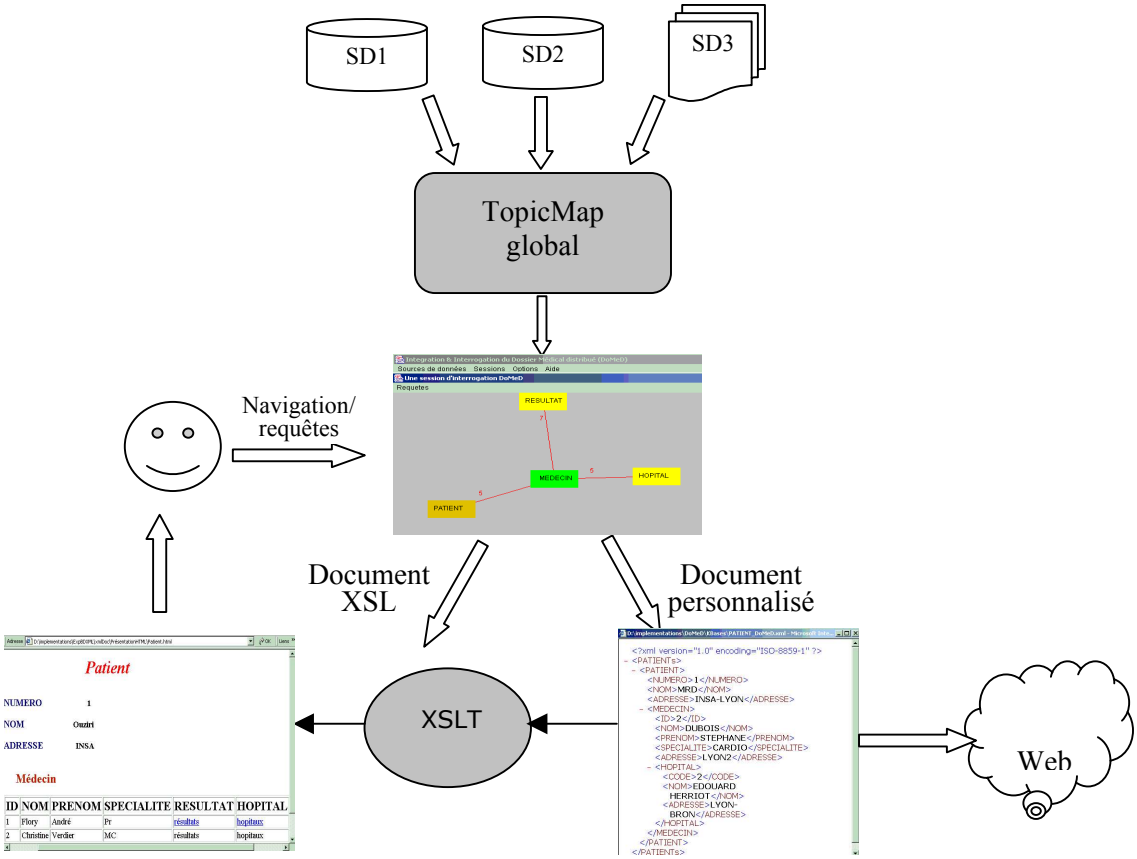


Figure 1 : Système d'interrogation et de construction de documents personnalisés

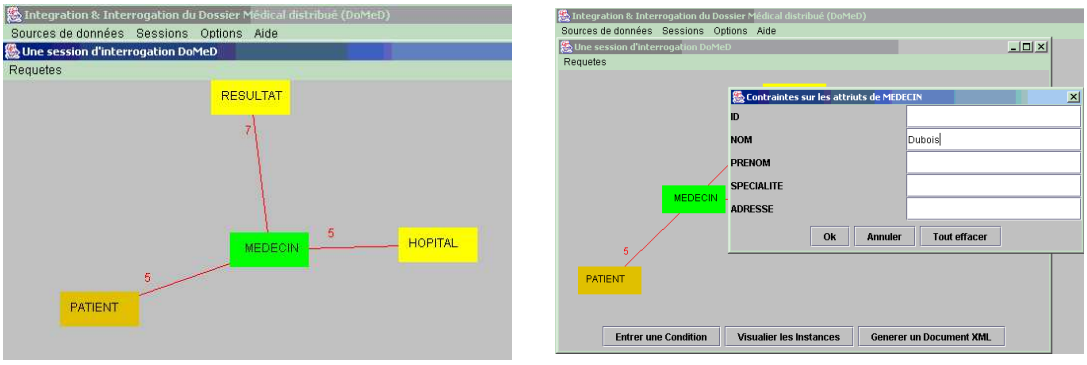
3.2 Interface de construction de documents personnalisés

Comme montré dans la figure 1, cette interface est générée à partir de la base de connaissances globale qui représente les sources de données. Les particularités de cette interface sont les suivantes :

1. Elle permet à l'utilisateur de consulter les données multi-sources par trois modes d'interrogation à savoir par navigation, par requêtes déclaratives et recherche par mots-clés. Etant donné que la navigation est le mode d'exploration le plus naturel et le plus familier aux utilisateurs, cette interface permet de naviguer dans la carte de concepts en visualisant les concepts d'une façon incrémentale au fur et à mesure que l'utilisateur avance dans la carte. Si le nombre de concepts (liens) de la carte est grand, l'utilisateur utilise l'un des deux autres modes d'interrogation : les requêtes ou les mot-clés. Si l'utilisateur connaît (complètement ou partiellement) la structure de la carte, l'utilisateur spécifie une requête de type SQL pour atteindre le concept désiré. Par exemple, la requête *Select Médecin Where this->*.Médecin* positionne la carte sur le concept Médecin en partant du concept courant, *this* (*this* est opérande de l'algèbre) et en suivant un chemin de longueur quelconque. Dans le cas où l'utilisateur n'a aucune idée de la façon dont la carte est structurée ou si le concept recherché se trouve dans plusieurs endroits dans la carte, on utilise dans ce cas la recherche par mots-clés (ou par concepts) qui trouvera toutes les occurrences du concept en question.
2. C'est une interface adaptée et dynamiquement adaptative. L'interface est adaptée à l'utilisateur dans le sens où le contenu et la structure de la carte de concepts visualisée sont adaptés en fonction des préférences et des droits de l'utilisateur. L'interface consulte le profil de l'utilisateur et construit une carte de concepts en conséquence. Pendant que l'utilisateur navigue, l'interface s'adapte en temps réel (dynamiquement) et visualise la carte de concepts en fonction des requêtes de l'utilisateur (adaptative). En fonction du résultat d'une requête, l'interface recalcule le contenu et la structure de la carte et ne visualise donc que les informations pertinentes. Si par exemple l'utilisateur souhaite explorer la

carte de concepts à partir du patient *Dupont*, toute la carte est structurée uniquement autour des données relatives à ce patient. Si la carte représente un lien entre le concept *Patient* et le concept *Médecin*, c'est que le patient *Dupont* a été traité par au moins un médecin. Si le patient *Dupont* n'a subi aucune radiographie, le concept *Radiographie* n'est donc pas représenté dans la carte.

3. Cette interface fournit un support de navigation sémantique des données multi-sources. En effet, la logique de navigation sur la carte de concepts consiste à construire une chaîne de concepts (et de leurs instances) sémantiquement liés par leurs associations. Par exemple, si l'utilisateur navigue en partant du concept *Patient*, après avoir spécifié la condition *nom=Dupont*, et visite successivement les concepts *Médecin* puis *Résultat* alors les instances $\{m_1, .., m_n\}$ du concept *Médecin* représentent les médecins qui ont traité le patient *Dupont* et celles du concept *Résultat* représentent les résultats des examens de *Dupont* produits par les médecins $\{m_1, .., m_n\}$.



(a) (b)

Figure 2 : Prises d'écrans de l'interface représentant une carte de concepts (a) et une fenêtre d'interrogation (b)

4. La carte de concepts est exactement cardinalisée. En effet, les liens entre les concepts représentent les associations entre ces concepts. La cardinalité d'une association (qui représente le nombre d'instances de cette association) entre deux concepts C_1 et C_2 est donc calculée et utilisée pour étiqueter le lien entre les boîtes représentant les deux concepts. Ces étiquettes augmentent la lisibilité de la carte de concepts,

donnent un aperçu a priori sur la distribution des données et permettent à l'utilisateur d'orienter sa navigation.

5. Etant donné que le document est le support le plus privilégié par l'utilisateur pour consulter l'information , à la fin d'une session d'interrogation et à la demande de l'utilisateur, le résultat (de la session de navigation/requêtes) est représenté dans un document personnalisé compréhensible par l'utilisateur final (figure 1). Ce document, généré automatiquement, contient l'ensemble des données sélectionnées par l'utilisateur lors de la session de navigation et d'interrogation. Ces données sont structurées conformément à l'organisation des données dans le TM global et respectant l'ordre dans lesquels ont été sélectionnées par l'utilisateur, c'est-à-dire, si le concept C_2 est sélectionné juste après le concept C_1 et sachant que les concepts sont sémantiquement liés, le document XML généré contiendra les deux balises BC_1 et BC_2 correspondantes aux deux concepts dans lequel la balise BC_2 est imbriquée dans la balise BC_1. Par conséquent, l'ordre de lecture des balises (qui peut exprimer la pertinence de la balise par rapport à l'utilisateur) dans le document respecte l'ordre dans lequel ces balises ont été sélectionnées (naviguées). Le document est créé initialement en XML et est transformé en un document compréhensible (et lisible) à l'utilisateur final en utilisant la feuille de style XSL conjointement créée.

4 Conclusion

La conception de systèmes d'exploration de données sur le Web est une tâche très difficile car les données sont volumineuses, hétérogènes et les profils des utilisateurs sont différents. Malgré les travaux qui consistent à améliorer la recherche d'informations sur le Web, son exploration reste difficile et rigide. En effet, la navigation sur le Web consiste à suivre les liens hypertextes prédéfinis par les concepteurs des documents Web. Le résultat de la recherche d'informations est éparpillé sur plusieurs documents Web et c'est à la charge de l'utilisateur de reconstituer le résultat global en combinant les segments de documents pertinents. L'utilisateur est souvent confronté à une surcharge d'informations et une multitude de possibilités de

navigations qui le rendent incapable d'orienter efficacement sa navigation, ce qui induit une perte de temps, de qualité et d'efficacité pour la recherche d'informations. Nous avons proposé un système d'interrogation/navigation des documents (bases de données) sur le Web. L'interface du système est dynamiquement adaptative en s'adaptant automatiquement en fonction du besoin de l'utilisateur spécifié lors de l'exploration. De plus, l'interface ne présente pas toute la carte de concepts mais seulement les concepts liés aux concepts courants et qui lui sont sémantiquement liés.

La perspective de ce travail est de proposer un modèle de données pour la base de connaissances TopicMap et une algèbre qui supportent les spécificités de l'interface. Une première réflexion consiste à utiliser un modèle permettant de faire des raisonnements logiques sur les données ainsi que sur leurs descriptions.

5 Bibliographie

[Abit-97] S. Abiteboul, D. Quass, J. McHugh, J. Widom, The Lore query language for semistructured data. *International Journal on Digital Libraries, 1(1), p. 68-88, April 1997*

[Aroc-98] G. O. Arocena, A. O. Mendelzon, WebOQL: Restructuring Documents, Databases and Webs. *Proceedings of ICDE 98 , Orlando, Florida, February 1998*

[Bern-94] T. Berners-Lee, R. Cailliau, A. Loutonen, H. Nielsen, A. Secret, The World-Wide Web. *Communications of the ACM, v. 37, n. 8, August 1994, pp. 76-82*

[Brus-96] P. Brusilovsky, Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction 6, 2-3, p. 87-129, 1996*

[Deut-99] A. Deutsch, M. Fernandez, D. Florescu, A. Levy, D. Suciu, XML-QL: A Query Language for XML. *In 8th International World Wide Web Conference, Toronto, May 1999*

[Flor-99] D. Florescu, D. Kossmann, Storing And Querying XML Data using an RDMS. *In IEEE Data Engineering Bulletin, vol. 22(3), p. 27-34, 1999*

[Garl-99] S. Garlatti, S. Iksal, P. Kervella, Adaptive on-line information system by means of a task model and spatial views. *In Computer Science Report, Eindhoven University of Technoloy, 59-66, 1999*

[Habr-99] J. Habrant, A. Corbel, J . J. Girardot, Les réseaux sémantiques comme outils d'aide à la navigation sur le web. *Conférence Interfaces Homme Machine 99, Montpellier, novembre 1999*

[Hein-96] O. Heinonen, K. Hatonen, K. Klemettinen, WWW robots and search engines. *Seminar on Mobile Code, Report TKO-C79, Helsinki University of Technology, Department of Computer Science, 1996*

[ISO-99] ISO/IEC 13250, Topic Maps. *Dec ISO/IEC FCD, 1999*

[Khei-95] A. Kheirbek, Y. Chiaramella, Integrating hypermedia and information retrieval with conceptual graphs. *In HIM95, Konstanz, Germany, 1995*

[Koiv-01] M. R. Koivunen, E. Miller, W3C semantic Web activity. In *Semantic Web Kick-off Seminar, Finland November 2, 2001*

[Kost-95] M. Koster, Robots in the Web: threat or treat?. *NEXOR, http://web.nexor.co.uk/mak/doc/ robots/threat-or-treat.html, april 1995*

[Lang-99] P. Langley. User Modeling in Adaptive Interfaces. *In J. Kay, editor, Proceedings of the Seventh International Conference on User Modeling, pages 357--371. Springer, 1999*

[McHu-97] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, J. Widom, Lore: A database management system for semistructured data. *In SIGMOD Record, 26(3), p. 54-66, September 1997*

[Mend-97] A. Mendelzon, G. A. Mihaila, T. Milo, Querying the world wide web. *International Journal on Digital Libraries, 1(1), p. 54-67, April 1997*

[Papa-95] Y. Papakonstantinou, H. Garcia-Molina, J. Widom, Object Exchange accross heterogeneous information sources. *In Proc. Of the 11th International Conference on Data Engineering, p 251-260, Taipei, Taiwan, March 1995*

[Robi-98] J. Robie, J. Lapp, D. Schach, XML Query Language (XQL). *http://www.w3.org/TandS/QL/QL98/pp/wql.html, 1998*

[Shan-99] J. Shanmugasundaram, K. Tufte, G. He, C. Zhang, D. DeWitt, J. Naughton, Relational Databases for Querying XML Documents: Limitations and Opportunities. *In Proc of the 25th VLDB Conference, p. 302-314, Edinburgh, Scotland 1999*

[Xform-02] XForms 1.0. *W3C Working Draft 21 August 2002*