ISYWEB : an XML-based Architecture for Web Information Systems

Moussa Lo

Laboratoire d'Analyse Numérique et d'Informatique (LANI)
Université Gaston Berger - B.P. 234 - Saint-Louis (Sénégal)
lom@ugb.sn

Amrane Hocine

Département d'Informatique - Université de Pau
B.P. 1155 - 64013 Pau Cedex (France)
Amrane.hocine@univ-pau.fr

Abstract

The development of Internet technologies made migrate the information systems towards the Web. This migration has profited mainly of an evolution of the database technologies. Nevertheless, it requires a new architecture of information system permitting to bring solutions for the problems of heterogeneous data integration, relevant information retrieval and diffusion of integrated data on the Web. We propose ISYWEB, an XML-based architecture for web information systems providing models, methods and tools to solve problems posed by the migration of information systems towards the Web. ISYWEB relies on an XML data repository called dataweb; we define a dataweb as a collection of structured and semi-structured data. In this paper, we describe the ISYWEB architecture and show how it allows to (1) integrate heterogeneous data sources into an XML data repository, (2) provide relevant XML information retrieval based on domain knowledge, and (3) design and realize web applications to publish the XML data on the Web.

1. Introduction

With the development of Web technologies, it is more and more necessary, in particular for enterprises, to exchange, integrate, store, interrogate and publish data coming from heterogeneous sources : databases, text documents, web documents, etc. Currently, the Web is the most used support to diffuse information. More and more information systems are developed above the Web platform taking benefits of all its technologies. These kind of information systems, called web-based information systems [2] can be considered as important information bases integrating several heterogeneous data sources accessible through the Web. The development of such systems needs to rely on a rich data model to provide powerful methods and tools for data integration, relevant information retrieval and a better use of retrieved data for exchange, sharing and treatment. Web-based information systems profited of an evolution of databases technologies allowing : (i) the definition of semi-structured data models to integrate heterogeneous data more or less regular [29]; (ii) the use of database concepts to design web data management systems [30], [27]; (iii) the definition of new information retrieval methods adapted to Web [31], [32]. The emergence of XML (eXtensible Markup Language) [1] as standard for the representation, the exchange and the diffusion of data on the Web allowed the development of new data integration systems. The migration of information systems towards the Web needs however an architecture able to provide, at the same time, data integration, information retrieval and data diffusion solutions.

We have mainly be interested by the environmental domain in which applications present the particularity to need a conservation/observation approach. In a such approach, the conservation aspect means that there are acquisition and conservation of data or documents and that this information nest updated but not filed. This information is then restored to end-users. The observation aspect expresses the fact that the users of such systems can carry out observations (measurements, synthesis, research, extraction, etc.) [11]. Our objective is to provide models, methods and tools for the development of web information systems based on a conservation/observation approach. We present in this paper an architecture entirely based on XML to develop Information Systems For the Web (ISYWEB) allowing the development of web based information systems according to a conservation/observation approach. The section 2 gives an overview of the ISYWEB architecture while the section 3 describes its components: the XML Documents Repository (XDREP), the XML Information Retrieval System (SIRX) and the XML Web Site Management System (XWEB). In section 4, we present the work related to the problems we approach : data integration, XML information retrieval, design of web applications.

2. ISYWEB : an all-XML Architecture

2.1.Towards a warehousing approach and XML-based architecture for web information systems

Web-based information systems use the web technologies to retrieve information from sources connected to the Web and to present the information in a Web hypermedia presentation [28]. Therefore, the migration of information systems towards the Web create problems of data integration, relevant information retrieval, and diffusion of integrated data. The

mediation systems focus on data integration and allow to exploit several heterogeneous information sources through a data integrated view. However, they rely on a virtual approach which not facilitate data preparation after their integration for web diffusion or the addition of an information retrieval module. In addition, the advantages provided by the XML language as data representation and exchange format on the Web make it impossible to circumvent for realization of web-based information systems. Moreover, XML provides many advantages [14]. XML allows a natural representation of data coming from any source. It provides the opportunity to develop wrappers allowing to transform any data source with an unspecified format into XML. XML allows also interoperability between the different components of a system by offering a data standard syntax. Our approach is to rely the development of web information systems on an architecture entirely based on XML.

2.2. Architecture overview

Our architecture relies on a warehousing approach and is built around an XML repository called dataweb (fig. 1) . A dataweb is defined as an XML documents repository built from data resulting from heterogeneous sources. We propose a dataweb model which is based on a global source. The global source integrates data from structured and semi-structured sources in XML format. A catalog of metadata which allows the management of the dataweb (consultation, interrogation, update) supplements it. For the information retrieval, we integrate into the dataweb a relevant information retrieval system. This system extends the results obtained in the traditional information retrieval systems to XML documents. To take account of the semantics of the data and to improve the search for a non expert user in particular, we use a concepts base which integrates a knowledge domain. The model of concepts base we propose supports on the use of a domain thesaurus and on the concept of unit semantic in order to associate a semantic structure to the logic one of the XML documents. For the diffusion, we propose a method based on a declarative approach. It allows the creation of various views of the dataweb according to several categories of users (non-experts, experts, decision makers, etc.). The data contained in the dataweb are diffused through one (or several) media base(s) represented in XML. We have opted for an all-XML architecture with the use of XML as a uniform format to represent heterogeneous data but also to represent all data manipulated in the system.

3. ISYWEB Components

ISYWEB is composed by the following components : the XML Documents Repository Management System (XDREP), the XML Information Retrieval System (SIRX) and the XML Web Site Management System (XWEB).

3.1 XDREP

XDREP allows integration of data from heterogeneous sources. It relies on the concept of dataweb that is defined as an XML documents warehouse built from data resulting from heterogeneous sources. The dataweb model we propose is based on a global source which integrates data from structured and semi-structured sources in XML format. A catalog of metadata which allows the management of the dataweb (consultation, interrogation, update) supplements it [3].

Dataweb data model

We use the concept of Information Unit (IU) to model dataweb data coming from the same information source. Hence, a dataweb can be defined as a set of IU. Data of an IU are stored and managed as XML documents. An Information Unit is a couple (d, md) where *d* is an XML document obtained from a dataweb data source, and md is a meta-document. A meta-document is a tuple (*n,p,u,v,s*) which comprises necessary metadata to get and manage an IU where : *n* is the source name, p the source owner, u the source URL, v the view to extract from the considered source and to integrate into the dataweb, and s the DTD of the XML document (d). We consider two kinds of information units : (i) the IU, we call internal, which are got from a semi-structured source (represented in XML); data of such IU are static and this kind of IU need very small updates; (ii) the IU, we call external, got from structured data (relational databases); data of such IU are represented in XML by means of wrappers before their storage in the dataweb, they are dynamic and this kind of IU need regular updates. We propose to make update periodically. The view to extract from a relational data source can be expressed in SQL; in the case of an XML source, an XML query language (like XQuery) or XSLT can be used to express the view. The notion of view is important here because it allows to overcome the problem of data proprietarism. Often, data's owners don't want to share some parts of their data; so we provide them the possibility to only integrate data they want to share. After the construction of the dataweb, IU data (documents and meta-documents) are stored in two bases of XML documents : a global source for documents and a metadata catalog for meta-documents.

Dataweb construction method

The dataweb construction is done through a declarative method in three stages (i) definition of IU : we obtain the metadata catalog; (ii) generation of the global source from the catalog; (iii) definition of a global schema from the DTD stored in the catalog, this schema is used by the to allow end users to perform queries written in Xquery over the global source.

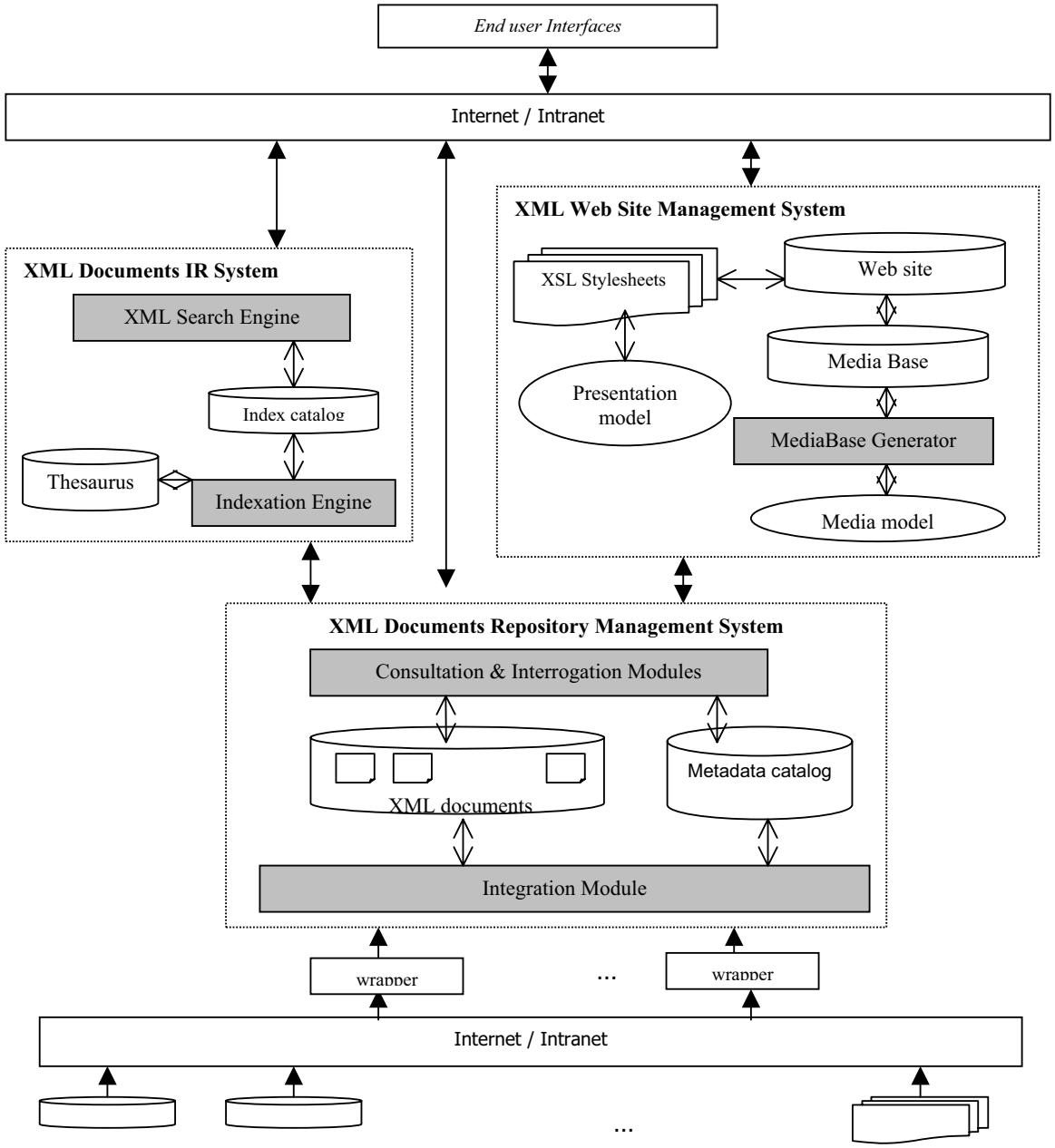

Figure 1: ISYWEB architecture

The dataweb global schema is linked to the sources schemas (DTD) through a Global-As-View approach. We have defined a semi-automatic method to facilitate dataweb construction [8]. The method is based on a relational-XML mapping [4] and on XSLT code generation. We then provide to dataweb administrator some facilities : interactive definition of information units, automatic generation of the metadata catalog from the IU definitions, automatic generation of the global source from the metadata catalog. The administrator defines each IU by fulfilling its attributes through a form. For the IU with XML sources, we propose a tool that assists the user when defining the view to extract from the source. Contrary to IU with databases sources, the views associated to XML sources are difficult to define because the user doesn't know the XML source structure. So, we provide help to user at two levels (i) during the definition of the view, the structure of the XML source is provided through a XML tree, so he can select in a interactive manner the XML elements and attributes to extract from the source ; (ii) the XSLT code representing the view to extract is automatically generated from the XML sub-tree defined by user. The metadata catalog is generated progressively of the IU definitions. To illustrate this approach, let us consider a simple example which consists to realize a dataweb

about publications of our research group. This dataweb has two IUs : an external IU having as data source a relational database containing information about the members of the group; and an internal IU having an XML document which contains a description of the group's publications.

3.2. SIRX

Our objective is to provide relevant information retrieval over the XML data stored in the dataweb. In our architecture, this task is done by the SIRX sub-system [6]. We provide (i) keyword-based search for end users that have neither domain expertise and nor knowledge about the documents structure; and (ii) search of relevant portions of documents and their classification according to their degree of relevance compared to the request. So, our IR approach is based on (i) the indexation of portions of documents (i.e. the leafs of the XML document tree); (ii) the implantation of a search engine with classification of relevant portions of documents. We exploit the graph structure associated to any XML document. So, the idea consists to realize the indexation at the leaf level because in an XML document, information is at this level. The leaves are the atomic units and we call them elementary units (or sub-documents). Each elementary unit is identified in a unique manner by a path representing its position in the document. We use XPATH to describe this path.

In the classic IR systems, the keyword-based search is relies on index structures named inverse files. In these systems, an inverse file has this kind of structure *<keyword, document, frequency>,* that means the considered keyword indexes the document with the gived frequency. This indexation method allows to retrieve documents and not sub-documents. So, we extend the classic inverse files technique to allow the possibility to retrieve elementary units. An index is represented in the following general form : *<keyword, elementary unit, frequency>,* that means the considered keyword indexes the elementary unit with the given frequency. For the dataweb, the indexation and search processes are done in the overall collection of XML documents stored in the global source. The result of the indexation process is a set of XML documents, named documents-index. For each XML document, we obtain a document-index which represents the result of the indexation of its elementary units with the keywords. The keywords are chosen by the dataweb's administrator. The set of documents-index form the index catalog. The user's query is expressed in the form of keywords and SIRX, with the index catalog, makes a ponderation of these keywords, retrieve and classify the elementary units that satisfy this query. The search process allows to retrieve the relevant elementary units of the XML documents stored in the dataweb. These elementary units are ordered by their similarity degree which measures the pertinence of the portions of the documents XML with the user's query. To evaluate the similarity between the user's query and the sub-documents, we use vector space model. In this model, the queries and the elementary unit of documents are represented as vectors. The search engine gives the sub-documents by descending order of their similarity with the query since for the user, more the similarity measure between the vectors representing elementary units and query is high, more the pertinence of the elementary unit is important. The current SIRX prototype has been developed in Java [9].

3.3. XWEB

We use a declarative way to diffuse the content of the dataweb, through a Web site. That is to define the structure of the Web site as a view over the existing data, i.e. the base of XML documents. This base provides an uniform view over the underlying data sources [5]. To use such method, we use a media model at the conceptual level; a media model describes the media units of the dataweb and their navigational structure. A media unit (M.U.) is an " information unit which has a certain autonomy in a user's point of view (i.e. which has its own sense and so presents a coherent idea or a concept) and which merits to be solicited in many consultation steps" [7].

Media Model

In the dataweb, a M.U. is described in XML and constitutes the content of a Web page. The set of media units and their navigational links (i.e. a media base) is obtained from an algorithm of automatic generation based on the media model. This model, described by specific XML tags is presented in this section. We propose eight types of media units: Xobjects, navigational contexts, index, menus, links, texts, images and web pages. The Xobjects and the navigational contexts provide views over the base of XML documents. An Xobject is an extract of the base of XML documents, obtained from the application of a filter (or query). The definition of an Xobject requires the description of a view over the base of XML documents. A navigational context is a set of Xobjects concerning a given thematic, they are accessible from an index. It is obtained, like an Xobject, by the application of a filter on the base of XML documents. The definition of a navigational context comprises the description of a view over the base of XML documents and the description of an index. The view allows to obtain a set of Xobjects accessible from the index. A media unit with text type is free text added by the designer. The image type permits to insert a picture from a file. The links allow describing the hypertext links in the dataweb. One distinguishes three sort of link: the structural links, the context links and the referential links. The page units allow composing many

M.U. into global M.U. described in XML, which will be associated to stylesheets to generate the pages of the web site. A menu is a set of (referential) links to other media units.

Media Base

The media base implements the media model. To build a media base, we propose the following steps: (a) determine the "page units"; (b) determine the navigational units between page units; (c) describe formally in XML the media units to build; (d) generate the media base from the M.U. description and the base of XML documents. The generation of the media base results from a relatively complex process. The generation algorithm has as entry three types of information [10]: the base of XML documents; the XML description of the media units; a catalog allowing to establish a correspondence between the resources (images, destination URL of the hypertext links, filters, …) logic names and their physic names: we can for example precise that the M.U. corresponding to the university logo, named "uppa", corresponds to the file "uppa.gif". This catalog is also described in XML. The result of this algorithm is a collection of media units, essentially with "Xobjects" and "page units" types. To build the web pages from the media base, it remains to associate a media form to each media unit, by using stylesheets (CSS or XSL) for example.

4. Related work

Data integration

Mediation systems rely on the I3 architecture [12] and have for final objective to allow the interrogation of heterogeneous sources through an integrated view at the mediator level. Several data sources are transformed by software which serve as mediators between the language, the model and the concepts of the source and the global concepts divided by part of all the sources [13]. TSIMMIS [15] and YAT [16] rely on a common data model at the mediator level to represent the data coming from heterogeneous sources and allows their interrogation from queries expressed in a common query language. are representative of this approach. PICSEL [17] relies on a domain model expressed in a knowledge representation formalism. More recently, the popularity of XML as description and exchange format has incited its use as mediator tool in the integration systems. Some mediation systems based on XML have been developed recently (MIX [18], Agora [19]). But, all these systems are based on the classic mediator architecture that is not really adapted for some kind of applications like environmental ones.

XML Information Retrieval

Many works are done to propose information retrieval methods in XML documents. Most of the proposed methods exploit the XML data model and allow to retrieve relevant XML sub-documents. Theses methods treat however XML data in a database point of view by proposing an extension of the quer languages [20], [21], [22]. For an information retrieval system, we think that this kind of approach is not adapted for several reasons : it is difficult for end user to know the structure of XML documents, a non expert user may have some difficulties to express his query precisely.

Design of web applications

Several methodologies and conception models have been proposed for Web applications. We can distinguish the conception methodologies of hypermedia applications and the management systems of Web dynamic sites. The goal of the first works in the domain had to resolve the hypermedia systems's problems. That was to find methodology adapted to this type of applications, which are specific in relation to traditional ones. All these methodologies are founded on the separation between the domain analysis, the specification of the navigational space and the conception of the user interface. They use modeling techniques based on the one hand, on the Entity-Relationship model (HDM [23], RMM [24]); and on the other hand, on the object model (OOHDM [25]). The solutions proposed in those works could however be applied to the Web context and inspired many of the works done for the conception of Web applications. In addition to the problems linked to the hypermedia context, other problems linked to the Web specificity arise to Web application designer: integration of various data sources, interoperability, dynamic nature of the Web, need to couple with DBMS (Database Management Systems), etc. In this context, many systems (STRUDEL [30], Araneus [27], WebML [26]) are developed; however, they are all build on HTML. Our approach realized by the XWEB module is different from these systems by (i) the use of XML to represent the data of the site, during its designing stage (in a base of XML documents coming from the structural model), and also during its exploitation (in a media base coming from the media model) and (ii) the enormous possibility it allows to integrate after a relevant information retrieval system.

5. Conclusion and perspectives

We have presented an architecture for web information systems based on a warehousing approach. This architecture presents the originality to resolve the main problems created by the migration of information systems towards the Web : data integration, relevant information retrieval and web diffusion of integrated data. Furthermore, it completely relies on XML and, therefore, takes the advantages provided by this standard. A prototype has been realized in Java to

implement the three components of the architecture. This approach is currently used in the "SIC-Web Senegal" project which aims to develop a web-based platform to facilitate the integration, the management, the organization, the diffusion and the exploitation of data produced in the region of the Senegal river. Our perspectives focus mainly in the introduction of Semantic Web technologies to extend the dataweb model, for example to describe the metadata.

5. References

[1] T. Bray, J. Paoli & C. Sperbeg-MacQueen: Extensible Markup Language (XML) 1.0, W3C Recommandation, http://www.w3.org/TR/1998/REC-xml-19980210/.

[2] T. Isakowitz, M. Bieber, F. Vitali : Web Information Systems, Communications of the ACM, Vol. 41, N° 7, 1998.

[3] M. Lo : Dataweb basés sur XML: modélisation et recherche d'informations pertinentes, PhD Thesis, Université de Pau (France), 2002.

[4] A. Hocine, M. Lo : Modeling and information retrieval on XML-based dataweb, Proc. of First Biennal on Advanced in Information Systems, Izmir, Turquie, LNCS 1909, 2000.

[5] M. Lo, A. Hocine, P. Raffinat : A Designing Model of XML-Dataweb, Proc. of 6th International Conference on Object-Oriented Information Systems, Canada, 2001.

[6] S. Smadhi, M. Lo, A. Hocine : Repository de documents XML : Modélisation et recherche d'informations pertinentes, proceedings of 7th MCSEAI, Algérie, May 2002.

[7] R. Deschamps: Bases de connaissances généralisées : une approche fondée sur un modèle hypertexte expert. Ph D Thesis of Toulouse University, France, 1995.

[8] S. Marques, C. Mendivil : Réalisation d'un outil d'aide à la construction d'un Dataweb, TER de Maîtrise d'Informatique, Université de Pau (France), 2002.

[9] P. I. Sall, M. Thiam : Développement en Java d'un Système de recherche d'informations dans une base de documents XML, TER de Maîtrise, UGB, 2004.

[10] V. Ellisalde & K. Rousseu-Salet : Modélisation objet et implantation en Java d'une Base Médiatique, Grand Projet, DESS IMOI, March 2001, Université de Pau (France).

[11] P. Dzeakou, P. Morand, C. Mullon : Méthodes et architectures des Systèmes d'Information sur l'environnement, Proc. of the 4th CARI, Dakar, Sénégal, october, 1998.

[12] G. Wiederhold: Mediation in information systems, ACM Computing Surveys, 27(2), pp 265-267, June 1995.

[13] J. Ullman : Information Integration Using Logical Views, Proceedings of 6th International Conference on Database Theory, LNCS 1186, 1997.

[14] V. Christophides: Community Webs (C-Webs): Technological Assessment and System Architecture, Research Report, INRIA, September 2000.

[15] A. Papakonstantinou, H. Garcia-Molina, J. Widom : Object Exchange accross heterogeneous information sources, In Proceedings of ICDE'95, 1995.

[16] S. Cluet and J.Siméon : Data integration based on data conversion and restructuring. Proceedings of WebDB'98, Valencia, Spain, March 1998.

[17] F. Goasdoué : A Knowledge Based Approach for Information Integration: The PICSEL System, In Declarative Data Access on the Web, Dagstuhl-Seminar-Report 251, 1999.

[18] C. K. Baru, A. Gupta, B. Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov, V. Chu : XML-based information mediation with MIX. Proceedings of ACM SIGMOD Conference on Management of Data, 2000.

[19] I. Manolescu, D. Florescu, D. Kossmann, F. Xhumari, D. Olteanu : Agora : Living with XML and Relational, The VLDB Journal, 2000.

[20] N. Fuhr, K. Grossjohann : XIRQL : An extension of XQL for information retrieval, Proceedings of ACM 2000 Workshop on XML and Information retrieval, 2000.

[21] D. Florescu, I. Manolescu, D. Kossman : Integrating Keyword Search in XML Query Processing, Proceedings of Nineth International WWW Conference, 2000.

[22] T. T. Chinenyanga, N. Kushmerick: Expressive retrieval from XML Information Document, Intern. Conf. on Research and Development in Information Retrieval, SIGIR 2001.

[23] F. Garzotto, L. Mainetti L. & P. Paolini : HDM : A model-based approach to Hypertext application design. ACM Transactions of Information Systems, 11(1),1-26, 1993.

[24] T. Isakowitz, E. Stohr & P. Balasubramanian : A methodology for the design of structured hypermedia applications. Communications of the ACM, (8)38, 1995.

[25] D. Schwabe and G. Rossi : OOHDM : The object-Oriented Hypermedia Design Model, Communication of ACM, August 1995.

[26] S. Ceri, P. Fraternali & A. Bongio : Web Modeling Language (WebML) : a modeling language for designing Web sites, WWW Conference, Amsterdam, May 2000.

[27] G. Mecca, P. Atzeni, P. Merialdo, A. Masci, and G. Sindoni: From Databases to Web-Base: The Araneus experience, Sigmod 98, 1998.

[28] R. Vdovjak, F. Frasincar, G.-J. Houben, P. Barna : Engineering Semantic Web Information Systems in Hera, in proceedings of WWW 2003 Conference, 2003.

[29] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, J. Widom : Lore : A Database Management System for Semistructured Data, SIGMOD Record, 26(3), 1997.

[30] M. Fernandez, D. Florescu, A. Levy and D. Suciu: A query languauge and Processor for a Web-Site Management System; In SIGMOD Record, 26(3), September 1997.

[31] D. Konopnicki, O. Shmueli : W3QS : A Query System for the WWW, In VLDB, Zurich, Suisse, 1995.

[32] A. O. Mendelzon, G. A. Mihaila, T. Milo : Querying the World Wide Web. In Journal of Digital Libraries, 1997.