Eleventh ERCIM Database Research Group Workshop: Metadata for Web Databases

by Karl Aberer and Brian J Read

The latest in the series of EDRG workshops was held at GMD Birlinghoven Castle near Bonn on 26 May 1998. The workshop was held in conjunction with the spring meeting of ERCIM from 25 - 29 May 1998 at GMD's headquarters in Sankt Augustin. The topic "Metadata for Web Databases" attracted significant interest among members of the working group, resulting in a very full day of presentations and discussion.There were sixteen participants from ten institutes.

In the context of Web data management, database systems are mostly used in an isolated way as data sinks or sources. Data management services that exploit and support the connectivity of the Web require the interaction and co-operation of different data management components on the Web. To enable this the Web needs to be equipped with the metadata on structure and behaviour of Web data that these components require. Thus the workshop was intended to address such questions as the extraction, modelling and querying of metadata, so adding semantics to the use of web data.

Keith Jeffery (CLRC-RAL) introduced the workshop topic by presenting an overview of the nature of metadata in databases, distinguishing its various purposes, and classifying it into three main kinds: schema, navigational and associative. Capturing metadata from the web presents problems as virtual pages generated from database queries are invisible to the large web crawlers. The limitations of HTML, and indeed XML, in managing metadata were discussed in this and several subsequent talks.

Yannis Stavrakas (NTU-Athens) expanded on the nature of metadata for web-based information systems. He distinguished three perspectives corresponding to the atomic level (information within a page or document), the local level (the structure of a site and links between documents), and the global information space of the whole web.

Terje Brasethvik (IDI/NTNU-Trondheim), currently in Paris, described his work with Arne Sølvberg on a Referent Model of Documents classified by semantic metadata. In this approach to sharing information on the web, they are developing a modelling language and editor to capture the meaning of documents.

Giuseppe Sindoni (Rome III University), currently visiting RAL, presented work from Paolo Atzeni's Rome group on a logical model for metadata in web bases. Their Araneus Data Model with the Penelope language embeds the schema within HTML. Turning to XML is potentially attractive, but that too has limitations for data modelling.

Three research projects were covered in the afternoon session. Menzo Windhouwer (CWI) described the work with Martin Kersten on the Acoi project. This is developing a feature detector engine to classify multimedia objects, especially images. The Acoi web robot has already stored in a database details extracted from over two hundred thousand images.

Thomas Klement (GMD) spoke about the ICE (Information Catalogue Environment) project. This concerns metadata for multidimensional categorisation and navigation support on multimedia documents. It includes an interesting use of dynamic menus to explore hypercube structures stored in an object-relational database.

The last presentation was from Donatella Castelli (CNR-Pisa) about supporting retrieval by “relation among documents” in the ERCIM Technical Reference Library (ETRDL) based on the Dienst system and the Dublin core. This provided an interesting discussion on the possible semantics of a relationship defined between documents.

The workshop concluded with a lively panel and discussion session on the future research direction of EDRG and also its rôle in the EC Fifth Framework Programme. A relevant component of the latter is “Creating a User Friendly Information Society”, especially Key Actions relating to application domains. This suggested that future workshops might be targeted towards an application area (such as transport, environment or health) instead of a technical topic. CWI emphasised semantic indexing of the web in an ambitious research agenda and cautioned against being too much influenced by funding considerations.

Karl Aberer - GMD-IPSI
Tel: +49 6151 869935
E-mail: aberer@darmstadt.gmd.de

Brian J Read - CLRC
Tel: +44 1235 446492
E-mail: b.j.read@rl.ac.uk

METADATA: AN OVERVIEW AND SOME ISSUES

Keith G Jeffery

Head Information Systems Engineering Division, CLRC-RAL, UK

kgj@rl.ac.uk

1. INTRODUCTION

1.1 Problem Statement

There are large problems for information systems today. There is a need to somehow manage / exploit the explosion of information appearing on multiple WWW sites with very variable standards of data quality and currency - and of course there is the need to know such data sources exist. This has the major aspects of:

(a) data quality

(b) query quality

(c) answer quality

(d) integration of heterogeneous sources

Let us consider each briefly in turn.

1.1.1 Data Quality

Data quality can be improved by better data collection facilities (including help and explanation with examples) and better validation controlled by constraints, with automated conversions of unit values if required or necessary.

1.1.2 Query Quality

Query quality can be improved not only by classical query optimisation using knowledge about the database size and structure, but also by assisting the user to formulate the query best to meet the requirement - by means of online help, explanation, examples.

1.1.3 Answer Quality

The answer to a query commonly includes values and structures that are unfamiliar to the user; explanations and help, hyperlinked descriptions of units, precision, calibration or of similar terms could help the user to understand better the results.

1.1.4 Integration of Heterogeneous Sources

First, there is the need to know that a source exists and to know something of its characteristics. Heterogeneous data sources commonly have disparate schemas and there is a need to understand the differences, even when apparently reconciled by one of the integration techniques.

1.2 The solution - Metadata

For all of the above to be realised, there is one essential and common ingredient: metadata. Let us consider briefly how it may be used in each of the cases:

1.2.1 Metadata for Data Collection

Metadata is necessary for validation through schema and constraints, using value-sets and domain range limits and even more sophisticated logic tests [GoGlJe93]. It is necessary for online help / explanation, and - in the form of a multilingual thesaurus - for translation.

1.2.2 Metadata for Queries

Metadata is necessary for validation through schema and constraints, for online help / explanation, for translation, and for optimisation: both user assistance in proposing more appropriate terms (synonyms, super- / sub-terms) and performance optimisation [JeHuKaWiBeMa94] since the metadata stores the structural indexes into the databases, optimal access paths, optimal query segmentation and distribution for parallelism and minimal network transfers.

1.2.3 Metadata for Answers

Metadata from the schema and associated metadata as domain ontology information (in a KBS) is necessary for answer consolidation, for online help / explanation, and translation [MaBeWiJeKaHu95].

1.2.4 Metadata for Integration

Metadata can catalogue sources of information at a high level so that they become visible. The well-known web indexing systems such as [AltaVista] or [Excite] do this in a very general way. An example in the field of CRIS (Current Research Information Systems) is the Bergen system [BergenCRIS] which points to structured informaiton systems for CRIS. Metadata, when used with an inferencing mechanism, is the key resource to find matching data structure and content despite heterogeneous representations and languages so allowing integration across heterogeneous data sources [JeHuKaWiBeMa94], [KoJe95].

Similarly, metadata provides the information necessary for customisation of standard products allowing integration into the desktop / office environment.

2. METADATA

2.1 Introduction

At present most Information Systems make very limited use of metadata and - since it supports all the user-friendly easy-to-use features and extends the range of available information features outlined above - perhaps this explains why these Information Systems have been less successful the few information systems really using metadata. Having outlined above how useful metadata could be for Information Systems, let us consider exactly what metadata is. The aim is to decide what kinds of metadata are useful for Information Systems and how best to generate, maintain and use metadata for the benefit of end-users of Information Systems.

Metadata is data about data. It is therefore of great utility:

(a) to any Information System which aspires to be more than a simple, inflexible unfriendly information source - use of metadata can allow dynamic optimisation and flexibility and allow integration over heterogeneous distributed information;

(b) to any end-user requiring help, explanation, data quality assurance, assistance in finding relevant information, assistance in integrating information from heterogeneous sources.

Distributed RDBMSs use metadata extensively [Je96]. WWW indexing systems (such as [AltaVista], [ExCite]….) are based on sophisticated metadata. Metadata is clearly of great importance. Perhaps the earliest use of metadata was in computerised library catalogue systems based on IR techniques where the catalogue card record is metadata describing the 'real data' in the book or other primary publication. Sadly this same field of endeavour is where metadata has hardly been developed further, and yet this is the very area of Information Systems technology where metadata could exert the greatest leverage.

2.2 Standards

There have been many attempts to standardise metadata structure and content for specific application areas. In the world of libraries the [MARC] standard for catalogue records allows some interworking. Unfortunately there are more than 50 major variants and so interworking is not as easy as one might expect. Similarly, in many scientific areas -e.g. space science, particle physics - there are metadata standards. In the world of commerce there is [EDI] / [EDIFACT]. There have been attempts to agree a standard European Patient Medical Record. Perhaps the most successful is in the field of Engineering: the EXPRESS language describing the STEP data exchange format with commercial support [STEPTOOLS].

The increasing requirement for interworking among systems handling grey electronic literature has caused the internet community to propose as a metadata standard the [Dublin Core] and, subsequently to provide convertors between the standards [UKOLN]. In the field of CRIS a common metadata form for exchange has been proposed [van Woensel] and is now used for metadata catalogs in the ERGO Project [Finch].

The great spread of WWW has increased dramatically the requirement for metadata standards to allow a global browsing and querying capability. The creation of [W3C] (World Wide Web Consortium) provided the forum for intense work on metadata [W3Cmetadata]. The main results have been PICS (Platform for Internet Content) which allows categorisation of WWW information in a way similar to film censorship, and following the Netscape MCF (meta content framework) and Microsoft XML-Data proposals, the W3C standard named RDF (Resource Description Framework - which is XML based) has gained widespread acceptance and subsumes PICS.

2.3 Kinds

Here we propose that there are three main kinds of metadata: schema, navigational and associative.

2.3.1 Schema

Schema metadata is an intensional description of extensional instances. Typically a schema consists of: database {name, size, security authorisations}, attributes {name, type, constraints}. Some of the constraints concern the attribute domain, some are inter-attribute and as such may express relationships.

The schema intension has a formal logic relationship to the data instances. This is important in ensuring data quality. It also provides a formal basis for systems.

*2.3.2 **Navigational***

Navigational metadata provides information on how to get to an information resource. Mechanisms include: filename, DB name + navigational algorithm, DB name + predicate (query), URL (Uniform Resource Locator), URL + predicate (query) or various combinations of them. They may also be obtained via a web-indexing mechanism (such as [AltaVista], [ExCite]…) which themselves make extensive use of metadata. Navigational metadata has no formal logic relationship to the data instances.

2.3.3 Associative

Associative metadata provides additional information for application assistance. The assistance may improve performance, accuracy or precision of the system and / or provide assistance to the end-user through a domain aware supportive user interface. The main kinds of associative metadata are:

(a) descriptive: catalogue record (e.g. [Dublin Core])

(b) restrictive: content rating (e.g. PICS) or security, privacy (cryptography, digital signatures) [W3C]

(c) supportive: dictionaries, thesauri, hyperglossaries [VHG], domain ontologies e.g. [PROTÉGÉ]

Associative metadata usually does not have a formal logic relationship to data instances although there may be systematic association relationships.

3. METADATA AND DATAWEB

3.1 Dataweb Technology

In order to combine the benefits of universal access (WWW) with the benefits of data managed and with structure and quality in a database various teams have worked on linking WWW and Database systems. CLRC-RAL was early into this field and experimented with several techniques since 1993, currently basing the departmental web [DCIweb] on Microsoft ASP technology. Now much of the information available over the web is held and managed within databases linked to the web through CGI (Common Gateway Interface) and scripts in a language such as Purl or Tcl.

3.2 The Problem

The data in these structured databases behind a web interface is essentially invisible to web indexing systems such as [AltaVista] or [Excite]. Since this is usually structured, managed, high quality data its use might be preferable to authored html pages. The problem is how to make it visible to web-indexing or information-cataloguing systems, in a way that is universally acceptable and utilised.

4. CONCLUSION

The key to the Future of Information Systems is Metadata. However, there are serious issues to be addressed:

(a) standard form for metadata: the W3C RDF is general and uses XML as the language - is this sufficient?

(b) sub-forms of metadata by application domain: will they all be based on the same basic data model and language to allow cross-domain interoperation?

(c) Progressively dataweb technology is being adopted; is there a standard mechanism for making such structured and hopefully quality information sources visible on the web through metadata?

References

[AltaVista] http://www.altavista.com/

[BergenCRIS] http://www.nsd.uib.no/english/research/eucris/

[DCIWeb] http://www.dci.clrc.ac.uk/

[Dublin Core] http://purl.oclc.org/metadata/dublin_core/

[EDI] http://www.dsii.com/edihome.html

[EDIFACT] http://www.r3.ch/sjwg/

[ExCite] http://www.excite.com/

[Finch] For details of this project please contact peter.finch@lux.dg13.cec.be

[GoGlJe93] Goble C A;Glowinski A;Jeffery K G: 'Semantic Constraints in a Medical Information System' Proceedings BNCOD-11 'Advances in Databases' July 1993 pp40-57 Edited by Worboys,M and Grundy,A F; Lecture Notes in Computer Science Series 696, Springer Verlag, 1993

[Java] http://www.javasoft.com/

[Je96] Jeffery,K.G: ‘Distributed Database: The Issues’ Invited Lecture DATASEM’96. Datasem’96 Proceedings pp 65-92, Ed Sbornik Prednasek, CS-Compex a.s., Olomoucka 84, 61800 Brno, Czech Republic. October 1996

[JeHuKaWiBeMa94] Jeffery,K G; Hutchinson,E K; Kalmus,J R; Wilson,M D; Behrendt, W; Macnee, C A: 'A Model for Heterogeneous Distributed Databases' Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag 1994

[KoJe95] Kohoutkova, J; Jeffery, K G : Hypermedata: Interoperability for Healthcare Systems ' Proceedings MBB’95 Slovakia, September 1995

[MaBeWiJeKaHu95] Macnee,C A; Behrendt, W; Wilson,M D; Jeffery, K G; Kalmus, J R; Hutchinson, E K: 'Presenting Dynamically Expandable Hypermedia' Information and Software Technology 37 (7) pp 339-350 1995

[MARC] http://minos.bl.uk/services/bsds/nbs/marc/commarcm.html

[PROTÉGÉ] http://smi-web.stanford.edu/projects/protege/Hpkb-web/MusenWestKickoff/tsld001.htm

[STEPTOOLS] http://www.steptools.com

[UKOLN] http://www.ukoln.ac.uk/metadata/interoperability/

[vanWoensel] van Woensel, l: ‘CERIF Manual’ October 1988 ; for more information contact lieve.vanwoensel@lux.dg13.cec.be

[VHG] http://www.venus.co.uk/vhg/

[W3C] http://www.w3.org/

[W3Cmetadata] http://www.w3.org/Metadata/

"Different Perspectives of Metadata for Web-based Information Systems"

Panos Vassiliadis, Yannis Stavrakas

National Technical University of Athens, Greece, {pvassil, ys}@dbnet.ece.ntua.gr

Abstract. Metadata can be of extreme value during the lifecycle of a Web-based Information System (i.e. during its design, development, maintenance, and evaluation phases). Metadata can also provide machine understandable information for applications which communicate without human intervention on the WWW. In this paper, we argue that the metadata of Web-based Information Systems can be of added value, if presented at three levels of granularity (global, local and atomic), further organized from the viewpoint three different perspectives: conceptual, logical and physical. To prove the feasibility of this approach, we have also implemented a model which respects the aforementioned separation, in the ConceptBase system.

1 Introduction

In the recent years, the notion of *enterprise-wide computing* has prevailed in the development of Information Systems. Enterprise-wide computing is based on the notion of bringing most of the computing power of an organization on the desktop. This is mainly achieved through the linkage of different separate networks into an *interconnected network*, by employing mainly the Internet [LL96]. Thus, the information is dispersed over a large number of interconnected repositories. The *World Wide Web* (WWW) is an information retrieval tool, mainly used to provide a uniform interface for the stored information to the users. *Intranets* are an enterprise-secure part of Internet technologies, providing security and performance apart from the other benefits of Internet access.

Nevertheless, the access to the information is not always an easy task, neither from the part of the information provider, nor from the part of the user. Problems of authoring, presentation, security and performance have to be resolved from the part of the former, whereas problems of data location, relevance and credibility have to be resolved from the part of the latter. Search engine technologies are employed to simplify the process of seeking meaningful content in the WWW, yet the result is not always satisfactory.

Metadata can be of extreme value during the overall life cycle of an information system. During the *design* phase, metadata can provide the designer with valuable information about the structure and meaning of the various concepts and structures he has already created. During the *software development* and *maintenance* phase of an information system, metadata play the role of a guide through the numerous pre-stored structures and hardcoded values of such a system. As for the *assessment* of an Information System, metadata are most useful in increasing the interpretability of the components and information of the system to the evaluator. Finally, it is obvious that in the presence of rich metadata schemes, the *reverse engineering* task of a system is a less painful procedure.

Apart from the information they provide to people, metadata play a critical role for the *interoperability* between different systems, since they contribute to the automatization of the communication of information between these systems. Metadata provide machine understandable information for applications which communicate without human intervention and can be used for the automatic processing of Web resources in application areas such as resource discovery, cataloging, intelligent knowledge agents, content rating, electronic commerce, etc.

For all these reasons, the presence of metadata describing *Web-based Information Systems* (WbIS) is of great importance. Currently, although several approaches have been made, there is no commonly agreed model for the representation of meta-information about Web-based Information Systems. In this paper we will (a) try to determine a framework for different *perspectives* of metadata for WbISs, (b) describe an implementation of such a model and (c) show how this framework can be used for the efficient querying of a metadata repository.

The structure of this paper is organized as follows: in Section 2, we first give a general architecture of a WbIS, along with a discussion on the problems occurring during its lifecycle. Furthermore, the role of metadata in such an environment is also discussed. In Section 3, we present our approach towards the general structure if a metadata repository for the WWW. In Section 4, we go into more detail on the implementation of our metadata model in the ConceptBase system. In Section 5, related work is presented and finally, in Section 6, conclusions and topics for future research are presented.

2 Description of the problem and contributions

2.1 Architecture of a Web-based Information System

Researchers and practitioners seem to agree on a generic architecture of a Web-based Information System. We can consider a WbIS as the synthesis of information from several sources in the operational environment of an Intranet. Each *Web site* consists mainly of a hypermedia database and a Web server who processes requests for information from the hypermedia database and returns answers to the clients. The hypermedia database[1] is the central data repository of a Web site. It stores chunks of information in the form of *resources* connected by links established by the user. These resources can contain text, graphics, sound, full-motion video, or executable programs and can be either *static* (i.e. having contents which are populated by the time of authoring) or *dynamic* (i.e. whose content is dynamically populated -for example, through a database query- each time a user asks for them). The most popular way to present text to users is through the use of a *markup language*. In a markup language, text is annotated with *tags*, having either presentational or/and semantical nature. Currently, the vast majority of text documents found in the WWW are authored with the use of the *HyperText Markup Language* (HTML) [RHJ97]. In the future, it is anticipated that the new emerging standard of *eXtensible Markup Language* (XML)[BPS98] will gain more ground, basically because it allows each author to define his/hers own personal tags. In either case, *the combination of these resources produces a virtual space of information, organized in a graph. The nodes of the graph represent individual resources and the arcs the user-established links.* In Figure 1, this basic architecture is presented.

The search for information is not done in a pre-determined organized fashion, rather than in a "navigational" way, by following the pre-established links. Clients request information found in the Intranet and receive answers, in the form of retrieved *Web pages*.

The information of a WbIS can be examined at three different levels of granularity: (a) at the *atomic* level each resource creates its own *atomic information space*, characterized from its content, structure and physical characteristics, (b) the interconnection of these resources produces a *local information space* for a Web site and (c) the synthesis of all the local information spaces of the Web sites produces the *global information space* of the WbIS.

[1] The term "database" should not be confusing. The vast majority of Web sites store information under a simple file system.

One should note that the information space (in all three levels) is *virtual*, since it can consist of both dynamic and static resources. Furthermore, the three levels can be considered to be interconnected -in which case it is interesting to model their interconnection- or in isolation -in the cases where we deal with only one of them.

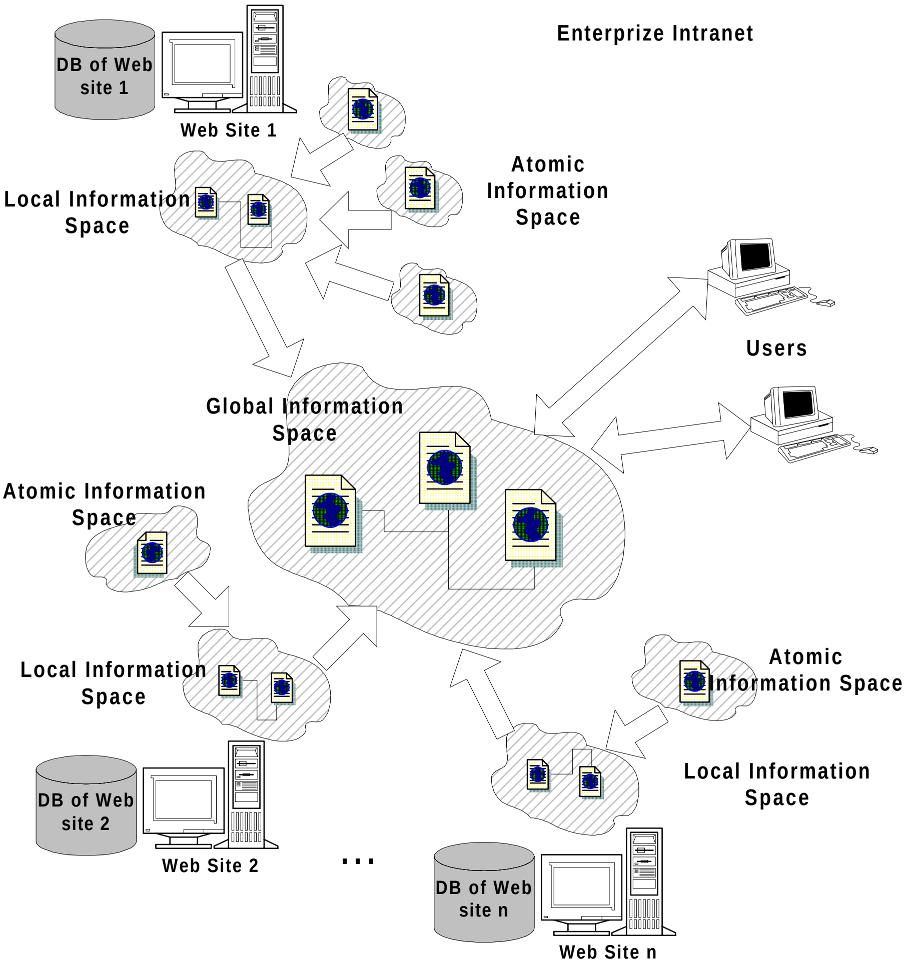

Figure 1. A generic architecture of a Web-based Information System

2.2 Problems during the life cycle of a WbIS

There are several processes taking place during the life cycle of a Web-based Information System. During the *design* phase, the system must be specified with respect to its structure, software and hardware support. The designer has to come up with solutions to problems such as "How can I provide the users with navigational paths which have a meaning for them?", "How can I construct my Web site consistently with the context of the overall context of the global information space of my Intranet?" or "How can I design the system so that it is easily maintained?". During the *software development* and *maintenance* phase of an information system, the administrator usually deals with problems regarding the location of information. For example, a typical question would be "What is the exact path to the index page of the department of electrical engineering?". The information providers, which usually are the people constructing the Web pages, are involved in the *Web authoring* procedure. They usually have to deal with similar problems, involving the location and presentation of information. Questions like "What (and where) is the template for the pages of my department?", "Why is the link I just added invalid?" are most common to people involved in such an effort. The users of the information system, are involved in an information retrieval process, which can usually make them nervous when dealing with

Web sites having a complex structure, bad performance, strange outlook or hard to use navigational paths.

Apart from this trivial tasks, performed in a regular fashion, a major process taking place in the World Wide Web is the process of automatic processing of Web resources. Search engines *catalogue* and *rate* Web sites for their content by applying sophisticated techniques, in the absence of a standard formalism for the description of this information. The most common problem, which furthermore has to be automatically dealt is the answer to the question: "What is this page really talking about?". No matter what techniques are employed, it is hard to imagine the case where no improvement of the classification of resources can be achieved.

2.3 The role of Metadata in a WbIS

All of the aforementioned problems have a common source: the lack of knowledge about the content, structure and location of the information of a Web site. Therefore, a major claim of both researchers and practitioners is that *rich metadata schemes could help eliminate or reduce this kind of problems.*

Metadata is instrumental in transforming raw data into knowledge. Without metadata, stored information is reduced to a meaningless data repository (also known as "data graveyard"). Several proposals exist for metadata formalisms, and more specifically for metadata customized for Web-based Information Systems. We can classify the proposals in two major categories: on the one hand, proposals at a high level such as the ones of the Metadata Coalition [Meta97] or MicroSoft [BHS+97] which deal with non-special purpose metadata specifications; on the other hand, proposals, which are still in a "work in progress" status, mainly proposed by the W3 Consortium, which are customized for Web resources [LS98], [GB97]. In the former approach, "metadata is data about data", or "information about enterprise data" [Meta97]. In the latter [W3C98], metadata are "machine-understandable information about web objects". We will mainly focus on the latter approach, since it is closer to the problem we discuss.

Yet, as we shall prove in Section 5 ("Related Work") the W3C approach suffers a serious problem: being definitely "machine-oriented", it results in a mixed representation of information, which is hard for a person to understand. One of the contributions of this paper is the introduction of a clear separation of metadata for WbISs in three perspectives (conceptual, logical and physical) for all their levels of granularity. Note that we do not propose a new formalism, or a detailed standard. We just claim that *clearly separated semantics in the presence of a repository with enhanced query capabilities, definitely make the work of both people and software agents much easier.*

3 Metadata for the Web

In this section, we will describe the basic architecture of a metadata repository for Web-based Information Systems. We distinguish three levels of detail for this task: the *global level*, used to describe the WbIS as a whole, the *local level,* used to describe each Web site and the *atomic level* dealing with each specific resource. It is important to note that each of the three levels can be viewed either isolated from its higher levels or in relationship with them. Furthermore, each of these levels can be viewed from three different perspectives: *conceptual, logical* and *physical*. The conceptual model provides a description of information with a way which is very close to the perception of the user; the physical model deals -more or less- with the details of storage and access of the information. Between these two extremes, the logical level hides several implementation issues but can be directly represented in a computer system in order to be able to both give an intuition of the data organization to the users and be close to the physical details of this organization. In the sequel, we present the models for each level and perspective in more detail.

3.1 The conceptual perspective

In the conceptual perspective, an abstract, high-level representation of the Information domain is provided. The most commonly used model for the conceptual representation of an application domain is the Entity-Relationship (E-R) model [Ch76]. Since the E-R model is a well-documented, established and powerful model, we choose to use it for the conceptual representation of a WbIS for all its levels. As we will see in the sequel, entities mostly provide patterns for the organization of information in Web pages, whereas relationships mostly determine navigational paths between these pages. We will distinguish between the different levels of granularity of the conceptual perspective.

At the upper level, the global conceptual model exists. The global conceptual model represents the overall information domain covered by the distributed WbIS. At each site, a local conceptual model exists, representing the information found at the specific site. Finally, each resource, e.g. Web page, can have its own conceptual model for the high level representation of its contents and structure. Our approach is open, in order to incorporate the cases where the information of an organization is distributed across various Web sites -which by definition are semantically related to each other. *We believe that the global conceptual model should be the central reference model of the organization.* Instead of considering it as the union/merge of the conceptual models of the lower levels, we argue that the inverse approach should be taken: all other models should be expressed with respect to the central model. This approach provides the organization with the potential to detect anomalies in the design at each level, since a global conceptual model can capture the relationships between different parts of information, distributed in different Web sites.

The same rule can be applied for the relationship between the conceptual models of Web sites and their resources. The relationship between two different Web pages, for example, can be modeled at the local level and not at the atomic one -therefore it is wiser to deal with the conceptual content of the specific pages in the context of an overall, local conceptual content, rather to do it the other way around. In Figure 2 the conceptual perspective of a WbIS is presented.

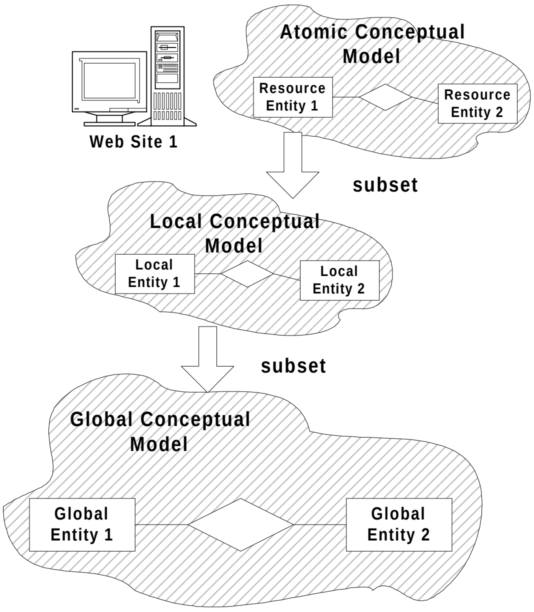

Figure 2. The Conceptual perspective of a WbIS

3.2 The logical perspective

In the logical perspective, one should try to determine how the conceptual model of a WbIS can be organized in abstract structures, independent from the physical attributes, yet close to the actual representation of the information.

In the *atomic level* we consider that a logical schema can exist for a resource. We will focus on the logical schemata of Web pages. The logical schema of a Web page defines its internal structure. In all markup languages for documents, tags are embedded in the text, either for semantical or for presentational reasons. HTML has mainly presentational tags; consequently the logical structure of HTML documents (denoted by the parse tree of each document) mainly concerns their presentation. On the other hand, XML is based on a logical scheme (also denoted by the parse tree of each document) which is oriented more towards the semantical structure of the document; in that sense it is very close to the conceptual perspective of the document.

As far as the *local level* of a Web site is concerned, although there exist several proposals for logical models, there does not seem to be a common agreement on one of them. Consequently, we tried to derive a common platform for such models. Our approach is based on the assumption that *a logical model, based on page patterns and navigational paths can (and must) be the cornerstone of the design of a WbIS.* This has already been the approach followed in several other proposals, e.g. [ISB95], [GMP95]. In the sequel of this section, we will describe the minimum requirements we believe a logical model for a WbIS should fulfill and present the minimalistic approach we have taken towards this specifications.

Our logical model consists of *page patterns* and *link patterns* between them. The page patterns can be derived from the entities of the conceptual model, as subsets, conjunctions or unions of different entities. This stems from the fact that the Web pages are not normalized relational tables rather than structures used for the presentation of information to the user; consequently, their content may combine information from various underlying entities. The link patterns can be derived from the relationships -possibly transitive- between entities. The direction and cardinality of a link are important, as well as a description of the link. The links between pages can be of different types (HTML 4.0 already supports two different types of links, namely "anchor" and "link"). In our quest for an open architecture, we do not constraint the model to a single type of links, rather present a generic. In the future, it is anticipated that more types of links will exist in the WWW. XML [BPS98] and its XLink specification [MD98] specify the way links might be provided for the Web pages, if any XML compliant language is used (instead of HTML). Our approach can be customized to these specifications easily through a specialization mechanism.

We believe that a logical model is a key component in the design process of a WbIS, and is mainly used by the people involved in this process. Thus, it should be concerned, on the one hand with the representation of conceptual entities and on the other with their internal structure. Consequently, we attach an inherent structure to each page pattern: a *Document Type Declaration (DTD)* is used to define the syntax of markup constructs through of element definitions, to deal with the combination of information within the page pattern.

A logical model with the characteristics we have prescribed, offers several advantages to the designer and administrator of a WbIS application. First of all, it is oriented towards the presentation of mixed information from various sources and the support of navigational paths from page patterns -in other words, it is very close to the real world structure of the WbIS. Second, such a model is high level enough to avoid the problem of mixed representation of schema and instance information. *This is the main reason for which we have chosen to separate our models in different levels: schema information should be present at the local model, whereas instance information should be present at the atomic level.* Furthermore, under this light, updates on specific pages do not always have to directly affect the logical model of the application; on the other hand sometimes this might be inevitable.

As far as the *global level* is concerned, the objective of a logical model should be to *detail the global search space of the organization*, based on the global enterprise conceptual model. To achieve this goal we provide a set of *GlobalQueries*. Each global query represents a possible request for information and is related to a set of concepts of the global conceptual model. Obviously, the set of global queries grows incrementally over time. We do not advocate that one should predict all the possible queries from the first moment of the WbIS operation. Rather, we claim, that based on the global conceptual model, one can create a first basis for search on the global information space. The answers to the queries (which can be either page patterns, or atomic resources) can be obtained by relating the involved global conceptual entities to the respective patterns or pages.

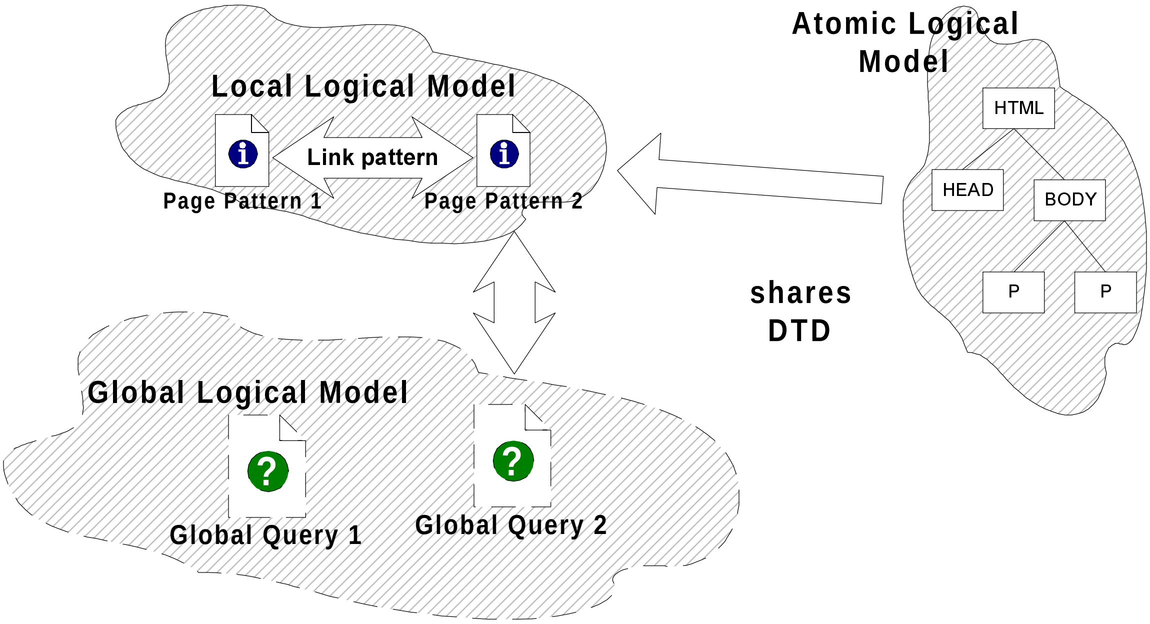

Figure 3. The logical perspective of a WbIS

3.3 The physical perspective

In the *physical* perspective, the objective is the description of the physical characteristics of the WbIS. At the *atomic level*, the physical instance of any resource on the WWW covers two issues: on the one hand, it deals with the location of the resource in the WWW; on the other hand it deals with its physical characteristics (e.g. size, date of last modification, type, transfer protocol).

As far as the *local level* is concerned, there exist two main groups of information about the physical characteristics: on the one hand there is the specification of hardware and middleware supporting the Web site (machinery, Web server, proxy, DBMS etc.). On the other hand, there is a description of the physical characteristics of the local information space (e.g. the directory structure served from the Web server).

Finally, at the *global level*, there may not exist at all any global physical interconnection other than the Internet. In this case, there is no need for physical description of the WbIS at the global level. If, on the other hand, the different Web sites are connected with private lines (e.g. in the sense of a Virtual Private Network), then there can also be a model for the physical description of the overall organization. In Figure 4, the physical perspective is graphically presented.

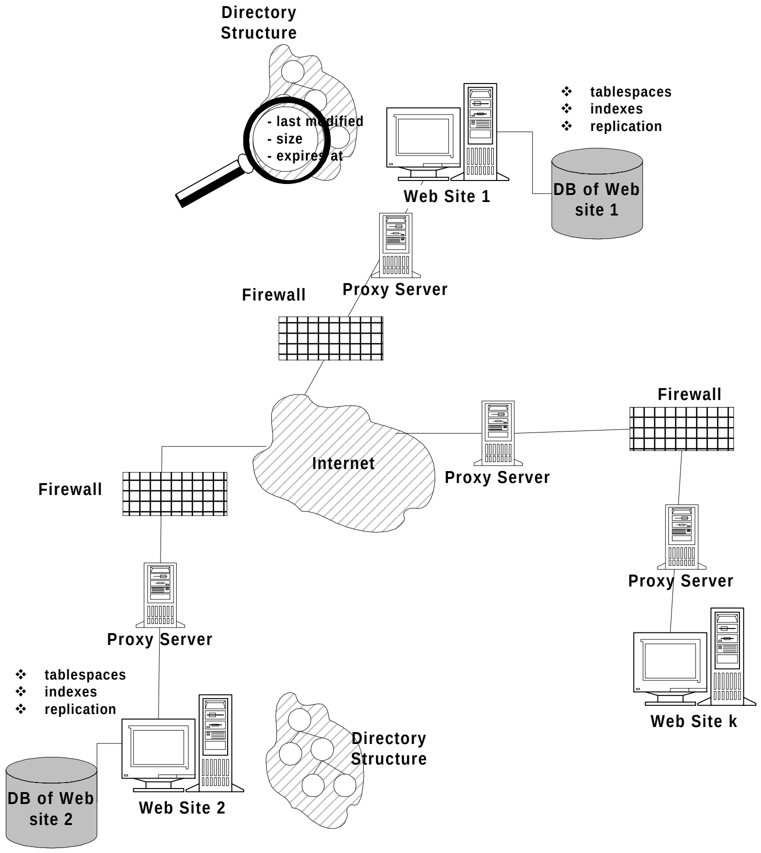

Figure 4. The physical perspective of a WbIS

4 The ConceptBase implementation

The model for WbIS metadata is represented in Telos, a metadata modeling language. Its implementation in the ConceptBase system [JGJ+95] provides query facilities, and definition of constraints and deductive rules. Telos is well suited because it allows to formalize specialized modeling notations by means of metaclasses. The Telos representation can provide the user with the facility to examine the structure and test the quality of the represented Information System.

The set of Telos classes employed corresponds to the description presented in the previous section. A detailed presentation is beyond the scope of this paper. Here, we will only demonstrate the power of the query language in order to present the usefulness of our approach.

The following query discovers the conceptual objects of the global conceptual schema with no other object pointing to them -i.e. not mapped to something in the Web Sites.

```
Individual NotImplementedGlobalConcepts in QueryClass isA WEntityType with
  attribute, constraint
       c: $ exists gs/WGlobalConceptualSchema (gs hasConcepts this) and
             not (exists le/WEntityType ls/WLocalConceptualSchema
                    (ls hasConcepts le) and (le relevantConcepts this)) $
end
```

The next query discovers the entity names related to each web resource.

```
Individual EntitiesForPage in GenericQueryClass isA WEntityType with
  parameter
      p: WResource
  attribute, constraint
      c: $ exists als/WAtomicLogicalSchema (p logical als) and
            (als hasConceptual this) $
end
```

Finally, although Telos is based on a deductive object base formalism, our approach can be also exported to other formalisms, e.g. semi-structured data, or the RDF approach proposed from the W3C [LS98]. We do not advocate that ConceptBase should definitely be the central metadata repository of a WbIS; rather we argue that rich metadata schemes can provide extended query facilities.

5 Related Work

A lot of effort has been spent on the standardization of meta data. A promising approach is the *Metadata Interchange Specification* [Meta97], proposed by the Metadata Coalition, an open group of companies such as IBM, Sybase, Informix, etc. The Metadata Interchange Specification (MDIS) should provide a standard access mechanism and a standard application programming interface to control and manage metadata with interchange specification-compliant tools. The MDIS is character-based and platform-independent. Furthermore, it is extensible because the contents of what is considered as metadata will be evolved. In a commercial approach, the repository system of Microsoft [BHS+97] is found. This metadata repository is extensible and evolvable, so that customers and independent software vendors can store any kind of (meta-) data in the repository.

As far as proposals customized for the WWW are concerned, the major proposal is the one made by the W3 Consortium [LS98]. RDF - the Resource Description Framework- is a foundation for processing metadata, providing interoperability between applications that exchange machine-understandable information on the Web. RDF uses a graph-based model for representing metadata and a specific syntax for expressing and transporting this metadata in a manner that maximizes the interoperability of independently developed web servers and clients. The syntax employed uses the eXtensible Markup Language (XML). RDF has re-used the results of several previous efforts for the standardization of metadata out of which we can contribution to the employed model is found in [GB97].

A very interesting fact about this model is its simplicity. The model is based on a simple graph, where entities, concepts, attributes and types are all modeled as nodes on a graph and any relationship between them (including *isA*, *hasA* relationships) is modeled as a labeled arc. Research work in semi-structured data has been based -more or less- on the same basic assumption about an underlying model [Abit97]. Yet, the problem of mixed representation is also present in the research in semi-structured data, since the model is composed just of nodes and arcs.

6 Discussion and Conclusions

The main results of this paper can be summarized as follows: (a) a clear separation of different aspects of

Metadata for WbIS, as far as the level of granularity and perspective are concerned and (b) the proof of the usefulness of our approach through the query facilities of ConceptBase. Future work involves the research on quality aspects of WbISs as well as the evaluation and completion of our model.

7 References

[Abit97] S. Abiteboul. Querying semi-structured data. In *Proc. 6th International Conference on Database Theory,* Delphi, January 1997.

[BHS+97] P. A. Bernstein, B. Harry, P. Sanders, D. Shutt, J. Zander. The Microsoft Repository. In *Proc. 23rd International Conference on Very Large Data Bases*, Athens, August 1997. Morgan Kaufmann Publishers.

[BPS98] T. Bray, J. Paoli, C.M. Sperbreg-McQueen. Extensible Markup Language (XML) 1.0, W3C Recommendation, *Available at http://www.w3.org/TR/REC-xml*, February 1998.

[Chen76] Chen P. The Entity Relationship Model -Toward a Unified View of Data. In *ACM Transactions on Database Systems,* vol. 1, no. 1, March 1976

[GB97] R.V. Guha, T. Bray. Meta Content Framework Using XML. Note to the W3C. *Available at http://www.w3.org/TR/NOTE-MCF-XML-970624*, June 1997

[GMP95] F. Garzotto, L. Mainetti, P. Paolini. Hypermedia Design, Analysis and Evaluation Issues. In *Communications of the ACM,* vol. 38, no. 8, 1995

[ISB95] T. Isakowitz, E. Stohr, P. Balasubramanian. RMM: A methodology for the Design of Structured Hypermedia Applications. In *Communications of the ACM,* vol. 38, no. 8, 1995

[JGJ+95] M. Jarke, R. Gallersdörfer, M.A. Jeusfeld, M. Staudt, S. Eherer: ConceptBase - a deductive objectbase for meta data management. In *Journal of Intelligent Information Systems, Special Issue on Advances in Deductive Object-Oriented Databases*, 4, No. 2, 167-192, 1995.

[LL96] K. C. Laudon, J. P. Laudon. Management of Information Systems. *Prentice Hall*, 1996

[LS98] O. Lassila, R. R. Swick. Resource Description Framework (RDF) Model and Syntax. W3C Working Draft. *Available at http://www.w3.org/TR/1998/WD-rdf-syntax-19980216*, February 1998

[MD98] E. Maler, S. DeRose. XML Linkng Lankuage (Xlink). W3C Working Draft WD-xlink-19980303. Available at *http://www.w3.org/TR/1998/WD-xlink-19980303*, March 1998

[Meta97] Metadata Coalition, Metadata Interchange Specification, (MDIS version 1.1). *Available at http://www.he.net/~metadata/,* August 1997

[RHJ97] Dave Raggett, Arnaud Le Hors, Ian Jacobs. HTML 4.0 Specification. W3C Proposed Recommendation, PR-HTML40-971107. Available at *http://www.w3.org/TR/PR-html40/cover.html*, November 1997.

[SR96] Seligman L.and Rosenthal A. A Metadata Resource to Promote Data Integration. *Proceedings of IEEE Metadata Conference, Silver Spring, MD,* April 1996

[W3C98] W3C. Metadata: A W3C Activity. *Available at http://www.w3.org/Metadata/Activity.html*, February 1998