METADATA & INFORMATION MANAGEMENT ISSUES IN XML-BASED MEDIATION.

Boris Rousseau, Eric Leray, Mícheál ÓFoghlú
Telecommunications Software Systems Group, Waterford Institute of Technology, Waterford, Ireland
Email: brousseau@tssg.org, eleray@tssg.org, mofoghlu@tssg.org

Keywords: Information Retrieval, Native XML Database, Web Services, XML queries

Abstract: The advancement in XML-based mediation has made a significant impact on the area of resource discovery. Search engines have now been provided with new ways to improve resource discovery and new tools to customise resulting content. In the early days of XML, this work was undertaken within the context of the European funded project GESTALT (Getting Educational System Talk Across Leading Edge Technologies). Building on this experience, new improvement came from the European funded project GUARDIANS (Gateway for User Access to Remote Distributed Information And Network Services). However, due to the lack of support for native XML databases and XML querying languages, search facilities were limited. This paper builds upon the achievements of both projects and proposes a solution for XML querying in XQuery.

1 INTRODUCTION

As the World Wide Web continues to grow so is the requirement for better **information management** and mediation tools. This is even more crucial in information and resource discovery where users are expecting resources delivery to be efficient and accurate through personalised and fast search requests. At present, a few solutions are available to the public, but as the number of users and Internet sites continues to increase the problem still remains. A complete solution requires the emergence of new standards and technologies such as the eXtensible Markup Language (XML) [1] that leverage the Web both as a flexible way to store resources and describe them for more advanced searching.

The work presented in this paper covers the development of an XML-based information system for metadata content management. The paper will present in two steps the context in which this work on information systems was developed, first by looking at the emergence of XML as an information mediation language and then by presenting the evolution of metadata, insisting on the technical challenges the XML has posed within these standards.

This section covers the major initiatives for improving searching over the Internet. The main body of the paper first presents the background of the work carried out in an IST European project, then it presents the work carried by the writers in developing such an XML-based information system. The argumentation will be supported by the presentation of technology solutions among which SOAP [2], Oracle XML Database [3] and XQuery [4] and research development examples based on European projects.

2 THE EMERGENCE OF XML IN RESOURCE DISCOVERY

2.1 The Ascent of XML

As the Web evolved, people and companies found themselves extending the HTML (Hyper-Text Markup Language) tagset to perform special tasks. A rich marketplace of server-side-includes and macro pre-processing extensions to HTML demonstrated that users understand the benefit of using local markup conventions to automate their in-house information management practices. And the cost of “dumbing down” to HTML became more apparent when more organisations tried to go beyond information dissemination to **information exchange**. The fundamental problem is that HTML is not extensible. A new tag potentially has ambiguous grammar (is it an element or does it need an end-tag?), ambiguous semantics (no metadata

about the ontology it is based on), and ambiguous presentation (especially without stylesheet hooks). For example, a book publisher would rather use tags such as <title>, <author> and <isbn> than the generic <P> paragraph tag. So, in 1996, the World Wide Web Consortium set out to find a way to introduce the power and flexibility of SGML into the Web domain. As they saw it, SGML [5] - in particular the cut down version they named XML - offered three significant benefits that were missing in HTML:

- *Extensibility*: Authors can define new elements, containers and attribute names at will.
- *Structure*: An XML Schema or a Document Type Definition can constrain the information model of a document. For example, a Chapter might require a Title element, an Author list and one or more Paragraphs.
- *Validation*: Every document can be validated. Furthermore, *well-formedness* can establish conformance to the structure mandated by a schema.

2.2 Impact of XML on Metadata Standards

A key aspect in networked educational systems is to define, as precisely as possible, the resources and services offered to potential users. Information on offered courses, related contents, targets audience or technical requirements should be made available in a way that permits discovery, searching, and eventually access. This need for better definition of content has been carried over by standards in the work on Metadata.

Metadata [6] is traditionally defined as data about data, or information about information, and is used to describe document contents and structure, and to provide information about accessibility, organisation of data, relations among data items, and the properties of the corresponding data domains. Metadata is also useful to provide textual description for non-textual objects, for example, to enable the representation of multimedia document properties in a structured way simplifying document management and retrieval.

Before any standard was defined, information were mainly represented as plain text files, HTML, or structures in relational databases. But with the arrival of XML, the IEEE LTSC [11] provided initial work toward a data structure specification in 1996, known as the Learning Object Model (LOM) [11]. As its name implies this standard has been focusing on providing a specification for description (or metadata) of Learning Objects. Learning Objects being an entity, digital or non-digital which can be used, re-used or referenced during learning. This specification provided the learners and instructors with the ability to share, exchange information and enable personalisation of content for individuals. Based on this initial work, numbers of standards have emerged for metadata and the Semantic Web [7] such as Dublin Core, IMS Learning Resource Meta-Data Information Model, ADL SCORM and W3C's Resource Description Framework (RDF) [8]. The most famous is Dublin Core (DC) [9], a simple Metadata element set intended at facilitating discovery of electronic resources. It is compact and its structure is the result of a wide consensus. DC Metadata can be recorded and transferred using different methods including HTML, XML, RDF [10] and relational databases.

XML also has made a significant impact in personal profiles standards. In this area, two main standards have emerged: IEEE Public And Private Information (PAPI) and IMS Learner Information Package (LIP). The first is a specification for user records that serves as a mean to communicate user information among components of a distributed learning environment. A key feature of the PAPI Learner Standard [11] is the logical division, separate security, and separate administration of several types of learner information. The current specification splits the learner information into 6 areas such as personal information and preference information. The PAPI Learner Standard may be integrated with other systems, protocols, formats, and technologies.

The second is a specification that describes learner characteristics for the purpose of personalisation of content, to discover opportunities engaging the learning experience. Comparing to PAPI, LIP [12] divides the learner information into 9 areas such as interest and affiliation. This specification is much more detailed than PAPI and provides a complete profile for users. In GUARDIANS, the User Profile specification [13] used IMS LIP as a baseline, simplifying the structure and including a history element.

In the information management community, XML has revolutionised the way standards bodies have released their own specifications. However, each standards body provides their own model and view, which makes it difficult for designers and developers to choose between them. Nevertheless, standards bodies are now trying to put their efforts together to develop an industry-wide standard that ensures interoperability of learning. The major initiative in this domain is the one from IMS, which provides specifications for interoperability of services in distributed learning based upon earlier work (LOM, DC, RDF). Furthermore, this consortium supports the incorporation of IMS specifications into products and services worldwide.

3 XML-BASED SEARCH SERVICE

From the recognised results of GESTALT [14], and taking on board the advances in XML and in distributed systems, GUARDIANS [15] objective is to specify and implement an open architecture for delivery and management of online personal information services accessible via a range of technologies. However, the information services envisaged are beyond today's largely text-based offerings. They are expected to be increasingly composed of rich, interactive media, offering enhanced guidance to the user in navigating service offerings and responding to the user's understanding of the material provided. In GUARDIANS' context, XML is used extensively to represent the resources available, for instance both DC and LOM specification have been implemented.
An example of LOM document is as follows:

```
<lom>
 <general>
  <identifier>URN:id75</identifier>
  <title>
   <langstring>The Universe</langstring>
  </title>
  <language>en</language>
  <description>
   <langstring xml:lang="en"> a basic
introduction to our universe, planets and
stars that surround our planet earth.
   </langstring>
  </description>
 </general>
…
</lom>
```

Figure 1: Learning Object Metadata (LOM) example

The development of the XQuery based search service uses an example based on the "Search Facility", more specifically its XML search component. The search facility is an extensible component that can contain any number of search interfaces. For the purpose of GUARDIANS, both an XML and SQL search interface are implemented. As XML is the common language for communication in this distributed environment, the query is first formulated in a commonly understandable XML canonical language based on the ISO 11179 part 6 [16]. Such a query is then translated to either an XML or SQL query specific to the target repository. Resulting matches are passed back to the Search Facility for collection and ranking.

4 XQUERY-BASED SEARCH SERVICE

4.1 Technology Solution Overview

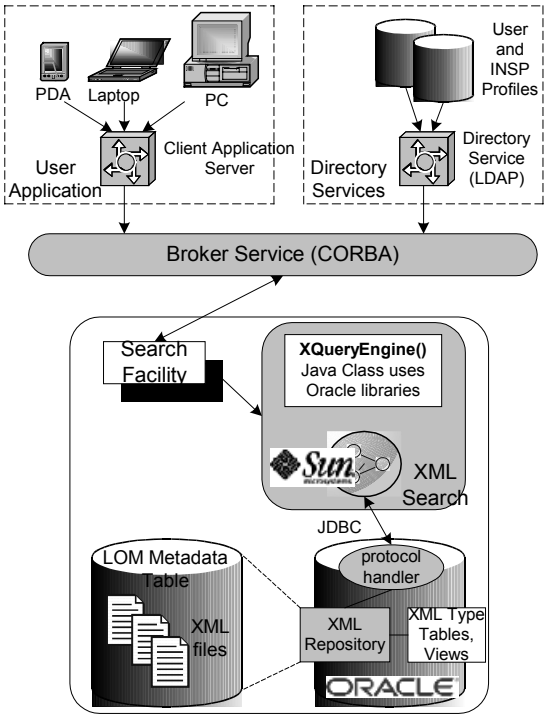

Figure 2: Technology solution diagram

The above figure shows the integration of the existing GUARDIANS architecture (upper part, down to the Broker) with the writers' solution. To formulate a search request, the user is required to provide some keywords and its user profile. From this information a query is formulated at the Search Facility using XQuery. Once formulated, this query is sent to a Web Service, which interrogates an Oracle 9i's Native XML Database with the Java DataBase Connection (JDBC) library. The response is sent back to the Search Facility for result collection and forwarded to the user application for further processing.

4.1.1 Database for XML Content Management

The initial impetus for XML may have been primarily to enhance this ability of remote applications to interpret and operate on documents fetched over the Internet. However, from a database point of view, XML raises a different exciting

possibility: with data stored in XML documents, it should be possible to query the contents of these documents. One should be able to issue queries over sets of XML documents to extract, synthesise and analyse their contents. XML documents fall into two broad categories: *data-centric* and *document-centric*. Data-centric documents are those where XML is used as a data transport. They include sales orders, patient records, and scientific data. Document-centric documents are those in which XML is used for its SGML-like capabilities, such as in user manuals, static Web pages, and marketing brochures.

In general, to store and retrieve the data in data-centric documents, an XML-enabled database that is tuned for data storage will be needed, such as a relational or object-oriented database, and some sort of data transfer software. To store and retrieve document-centric documents, a *native XML database* or *content management system* will be needed. When dealing with educational metadata documents, the focus is not on the computation of data held within the XML documents, but rather on the retrieval of appropriate documents.

Oracle Corporation released in early 2002 the second version of its advanced database for content management: Oracle9i Database. As part of this new database suite, Oracle9i supports a new system-defined data type called: XMLType. XMLType has built-in functions that offer a mechanism to create, extract and index XML data stored in the database. More recently Oracle has made available support for XQuery on its OTN website [17] via a refined Java API. Different other native or XML-enabled databases already support XQuery such as the Ipedo XML Database [18].

4.1.2 Web Services

Web Services [2] technology is based on the concept of Service-Oriented Architectures (SOA). The SOA is not new by any means. Many organisations have delivered different SOA-based frameworks for several years, including Microsoft, Sun and the OMG with COM, EJB and CORBA specifications respectively. It is only recently that open standards for SOAs have come onto the scene. Microsoft, Sun and IBM have put open standards at the heart of their approaches to SOA in the form of XML and SOAP. The Simple Object Access Protocol (SOAP) is a communication protocol that defines how self-describing XML data is transmitted from one point to another.

Web Services technology is an emerging technology, based on this protocol, driven by the will to securely expose business logic beyond the firewall. The technology offers a great potential to the world of distributed computing. Connecting to Web Services can be done quickly and painlessly with *the right tools*. However, the technology is still in its **infancy**, with many issues still to be addressed. One of the major issues is security. Although many Web Services may be public and available for all use, the most useful of them will be business related and therefore will need some method to restrict their access to only authorised users. Currently there is no definition of how security will be implemented, to what extent and how other services will dynamically support these layers of security. However, it is clear that security will be paramount for real world solutions.

Different software vendors offer Web Service products. For the purpose of the development of a search facility based on XQuery, the Java Web Services Developer Pack (Java WSDP) [19] has been utilised. The Java WSDP is an integrated toolset that allows development of Web Services. It provides all the standard implementation of the Web Service standard including WSDL, SOAP, ebXML and UDDI. It is really an all in one development pack to simplify building of Web Services using Java.

4.1.3 Supporting Information Retrieval with XQuery

Experts agree that XML will become the standard for data storage and retrieval in the next decade, however they cannot agree on how it will be implemented. Many believe that current database system (like Oracle, SQL Server and Sybase) will migrate their engines to store native XML, and that manufacturers will build optimisation, processing and manipulation capabilities onto this new engine. Others believe that database manufactures will leave engines intact and simply add an XML layer, allowing the engines to consume and emit the underlying data based on queries from existing XML languages like XSLT. With more data passed around as XML, and more systems designed to produce it, developers need a way to query XML sources for specific pieces of data from the data source.

The first standard approach to access these XML data sources was named XML Path Language (XPath). XPath was designed to allow **navigation** within an XML file and simple queries on a single file. Since, XPath was designed to navigate and "query" only a single XML data source at a time, using XPath effectively to query multiple data sources requires the developer to perform complex XML Document merges. A simplified example of such query used in GUARDIANS is as follows:

```
/lom/general/title/langstring[contains(text(
),"universe") ]
/lom/general/description/langstring[contains
(text(),"universe") ] or
/lom/classification/keyword/langstring[conta
ins(text (),"universe") ]
and (/lom/general/language[text()="en"])
```

Figure 3: XPath 1.0 query example

Such query has been tested successfully as part of the GUARDIANS system. However, XPath proved limited. Hence, from GUARDIANS experience, queries can be very long when trying to incorporate most of the user's preferences or trying to use SQL-like functions (joins, union, etc). Furthermore, results cannot be formatted requiring further processing to be displayed to a user. For this reason, different initiatives for XML query language were introduced. Out of these different initiatives the one that has been recently specified as a draft by the W3C is XQuery [4], which is designed to be broadly applicable across many types of XML data sources.

XQuery started life as Quilt [20], a collaborative effort of three working group members: J. Robie, D. Chamberlin and D. Florescu. XQuery was designed to solve the problem of multiple data sources by allowing complex queries across not only multiple XML Documents, but also between XML documents, relational databases and object repositories, and other unstructured documents. XQuery is a very **rich querying language**. It has primitives to allow iteration through data sources, as well as sorting, aggregation, and grouping functions. It allows connectivity between sources and restructuring of documents, based on defined criteria. More importantly, it includes a standard mechanism for extending the language with custom functions. This is very similar to the way relational databases have their own query and stored procedure languages. The difference is that XQuery's similar functionality will work across both XML data sources and relational data sources.

XQuery becomes more interesting when looking, beyond simple XPath, at the so-called FLWR. FLWR (pronounced "*flower*") is an acronym that stands for the four possible XQuery sub-expressions, this expression type can contain: FOR, LET, WHERE, and RETURN. This kind of expression is often useful for computing joins between two or more documents and for restructuring data.

```
FlwrExpr      ::= (ForClause | letClause)+
whereClause? ReturnClause
ForClause     ::= 'FOR' Variable 'IN Expr
   (',' Variable IN Expr)*
LetClause     ::= 'LET' Variable ':=' Expr
   (',' Variable := Expr)*
WhereClause   ::= 'WHERE' Expr
ReturnClause  ::= 'RETURN' Expr
```

Figure 4: FLWR Expressions in XQuery

An example based on GUARDIANS follows:

```
FOR  $r IN sqlquery("select metadata from
   guardians_lom")/ROW/METADATA
WHERE
   contains($r/lom/general/title/langstring,
   "universe") OR
   contains($r/lom/general/description/
   langstring, "universe") OR
   contains($r/lom/general/keyword/
   langstring, "universe")OR
   contains($r/lom/technical/format,
   "application/x-shockwave-flash")  OR
   (contains($r/lom/technical/requirement/
   type/value, "Browser")   AND
   contains($r/lom/technical/requirement/
   name/value, "Microsoft   Internet
   Explorer")) OR
   contains($r/lom/classification/taxonpath/
   taxon/entry/langstring,  "astronomy") OR
   contains($r/lom/educational/learning
   resourcetype/value/langstring,
   "Video/Animation") AND ($r/language="en")
RETURN <result> { $r/title/langstring }
   {$r/general/description/langstring}
 </result> ");
```

Figure 5: An Exact Search in XQuery using Oracle XDB

In this query, the first three conditions are taken from user-defined keywords. The rest is from preferences as informed in the user profile. It contains platform capabilities (format, browser) and interest preferences (astronomy) and language. Comparing to XPath, it looks pretty clear that the results will be an XML file with "result" as the root containing a title and a description. This formatting of results is very convenient for display purposes.

However from this experiment Oracle 9i seems heavyweight for the purpose of a search service. In fact, the whole database management system is installed with all its server-side facilities that were not required for this test. Furthermore, the fact that both Oracle and JWSDP used the same port for communication posed some problems of configuration.

For more details on how to implement the XQuery and Web Services consult the technical implementation page [21].

5 CONCLUSION AND FUTURE WORK

It is important to point that unlike XML, XML Schema Definition (XSD) or the eXtensible Stylesheet Language Transformations (XSLT), XQuery is a draft and not a recommendation. As such, the vendor community is just beginning to create tools to make it easy to generate XQuery programs. Although XQuery is in its **infancy** and

not ready for widespread production use, it bears sufficient potential to make it a very relevant tool in the data manipulation arsenal. If the standard and the vendors' tools based on the standard can deliver on the promise of XQuery, then it might even become the Holy Grail of query languages. But most importantly, one should remember the way SQL has had to mature over more than 20 years and therefore not expect too much from XQuery in the draft stage it is at now.

From this experiment, future additions could be envisaged to provide an even better GUARDIANS search service. The major issue is that the XML search function was fulfilled using XPath, which proved to be very limited. Hence, the idea behind this paper will be to ultimately replace GUARDIANS XPath-based search with this work on XQuery. Currently, from this experience, an XML store has been implemented to manage and query (using XQuery) user profiles in another IST European project: AlbatrOSS [22].

From this work, XQuery offers lots of promises. It represents the confluence of document and databases research and is a powerful optimisable language. XQuery is useful for querying and managing XML and provides aggregation operators similar to SQL. However, one could still see some problems remaining: static typing, updating of the database and the syntax embedding in XML.

REFERENCES

[1] "*Extensible Markup Language (XML) 1.0*", W3C recommendation, Feb. 1998. Available at http://www.w3.org/TR/1998/REC-xml-19980210

[2] Curbera F., Duftler M., Khalaf R., Nagy W., Mukhi N., and Weerawarana S. "*Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI*" The IEEE Internet Computing, March-April 2002.

[3] The Oracle Corporation "*Querying XML in a Standard Way*", available at: http://otn.oracle.com/tech/xml/xmldb/htdocs/querying_xml.html

[4] "*XQuery: A Query Language for XML*", February 2001. World Wide Web Consortium, W3C Working Draft. Available at http://www.w3.org/TR/2001/WD-xquery-20010215

[5] Standard Generalized Markup Language (SGML), ISO8879, Information Processing-Text and Office Systems, 1986.

[6] Day, M., Summer 2001 "*Metadata in a nutshell*". Article published in Information Europe, p. 11. Information Europe quarterly magazine of EBLIDA (the European Bureau of Library, Information and Documentation Associations).

[7] Berners-Lee T., Hendler J. and Lassila O. May 17, 2001. "*The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*". Available at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[8] Luis E. Anido, Manuel J. Fernández, Manuel Caeiro, Juan M. Santos, Judith S. Rodríguez and Martín Llamas, May 2002. "*Educational metadata and brokerage for learning resources*", Computers & Education, Volume 38, Issue 4, Pages 351-374

[9] The Dublin Core Metadata Home Page, http://uk.dublincore.org/

[10] Decker S., Melnik S. et al. "*The Semantic Web: the Roles of XML and RDF*" The IEEE Internet Computing, September-October 2000.

[11] The IEEE Public and Private Information latest specification (2001), available at: http://www.edutool.com/papi/

[12] The Instructional Management Systems (IMS) Learner Information Package Specification. Available at: http://www.imsglobal.org/profiles/index.cfm

[13] Rousseau B., Browne P., ÓFoghlú M. "*User Profiling for Content Personalisation in Information Retrieval*". In Press.

[14] GESTALT (Getting Educational Systems Talking Across Leading-Edge Technologies), http://www.fdgroup.com/gestalt/

[15] The GUARDIANS (Gateway for User Access to Remote Distributed Information and Network Services), http://www.ist-guardians.tv/

[16] ISO 11179 Parts 1-6, Specification and Standardization of Data Elements from the ISO/IEC 11179 available online at: http://www.iso.ch/

[17] The Oracle Technology Network, available at: http://otn.oracle.com/tech/xml/content.html

[18] The Ipedo XML Database web site, available at: http://www.ipedo.com/

[19] The Java Web Services Developer Pack, available at: http://java.sun.com/webservices/webservicespack.html

[20] Chamberlin D., Robie J. and Florescu D. June 2000. "*Quilt: an XML Query Language*", Presented at XML Europe, Paris. Available at: http://www.almaden.ibm.com/cs/people/chamberlin/quilt_euro.html

[21] The technical implementation web page, available at: http://www.tssg.org/papers/ICEIS2003/lr_XQtechnical.pdf

[22] AlbatrOSS (Architecture for Location of Third Generation Operation Support System), http://www.ist-albatross.org/