XML-Based Information Mediation for Digital Libraries

Chaitanya Baru, Vincent Chu, Amarnath Gupta, Bertram Ludäscher,
Richard Marciano, Yannis Papakonstantinou, Pavel Velikhov
University of California, San Diego, La Jolla, CA 92093
{baru,gupta,ludaesch,marciano}@sdsc.edu {vchu,yannis,pvelikho}@cs.ucsd.edu

ABSTRACT

We demonstrate a prototype distributed architecture for a digital library, using technology being developed under the MIX Project at the San Diego Supercomputer Center (SDSC) and the University of California, San Diego. The architecture is based on XML-based modeling of metadata; use of an XML query language, and associated mediator middleware, to query distributed metadata sources; and the use of a storage system middleware to access distributed, archived data sets.

1. INTRODUCTION

The *Mediation of Information using XML (MIX)* project at SDSC and the Database Lab at UCSD employs a *mediator-wrapper* approach, based on XML, to model and query information in distributed, heterogeneous information sources. The technology being developed includes the *XML Matching And Structuring (XMAS)* query language and the *Blended Browsing and Querying (BBQ)* interface. The mediator middleware, MIXm, employs XML as the common model for data exchange and provides integrated *views* defined over distributed, heterogeneous information sources. The mediator views are expressed in XMAS, which is a declarative, XML-based query language. To facilitate user-friendly query formulation, and for optimization purposes, MIXm employs XML DTDs as a structural description (in effect, a "schema") of the exchanged data. The novel features of the overall system include:

- Data exchange and integration is solely based on XML, i.e., instance and schema information is represented by XML documents and XML DTDs, respectively. XML queries are denoted in XMAS, which builds upon ideas of previous query languages for semistructured data, like XML-QL, MSL, Yat, and UnQL. Additionally, XMAS features powerful grouping and order constructs for generating new integrated XML "objects" from existing ones.
- A Blended Browsing and Querying (BBQ) graphical user interface, which is driven by the mediator *view DTD*, and which integrates browsing and querying of XML data. Complex queries can be constructed in an intuitive way, in a manner reminiscent of the Query-by-Example (QBE) system. Due to the nested nature of XML data and DTDs, BBQ provides graphical means to specify the nesting and grouping of query results.
- Support for demand-driven query evaluation, i.e., the BBQ interface generates queries (or, *incremental* parts of a query) based on a user's particular navigation into the mediated view.

2. MIX ARCHITECTURE

The main module of MIXm middleware is the query processor, which resolves the client queries with the mediator view definitions, resulting in a set of unfolded queries against the underlying sources. Where necessary, wrappers are used on top of sources. XMAS queries can be simplified based on the given or inferred DTDs. Evaluation and further optimization of XMAS queries is accomplished by translating them into the *XMAS algebra*. Like in TSIMMIS, views are computed at runtime and not pre-computed and materialized. Thus, the XML views in MIX are *virtual*. In addition, MIXm employs a *lazy* approach to query evaluation, i.e., evaluation is driven by the client's navigation into the virtual XML view corresponding to the query result. This is accomplished by implementing the Document Object Model (DOM) API for *Virtual XML documents* (DOM-VXD).

BBQ in a Nutshell. The graphical user interface BBQ allows the construction of complex XML queries in an intuitive way, based on the DTDs of the mediator view. The results returned via MIXm are displayed in a separate window and can be browsed in the usual way. Since the answer is itself a (virtual) XML document, it can also be queried using the same interface. In this way, BBQ smoothly integrates browsing and querying of XML data.

XMAS Queries. The output of BBQ is a XMAS query. In addition, since BBQ supports the generation of complex XML queries, it can also be used by the "mediation engineer" as a design tool for the mediator view. The syntax of XMAS queries resembles XML-QL. For example, the XMAS query in Figure 2 extracts the title, type, and image identifier for all art pieces which are paintings, and constructs the corresponding XML answer elements.

3. A DIGITAL LIBRARY PROTOTYPE

The MIX technology is being used in a pilot project to implement a prototype architecture for digital library collections as part of the California Digital Library (CDL). CDL is a tenth library for the University of California (UC). A collaborative effort of the nine campuses, organizationally housed at the University of California Office of the President, CDL is responsible for the design, creation, and implementation of systems that support the shared collections of the University of California [CDL].

The pilot project is based on a collection of high resolution images of art pieces obtained from the Art Museum Image

Consortium [AMICO]. The prototype architecture of the digital library is shown in Figure 1. Clients use the BBQ interface to query a mediated view of the metadata corresponding to art pieces. The basis for this view is an XML DTD derived from the metadata specification provided by AMICO. For information sources that do not directly support the AMICO DTD, a wrapper is required to map the source schema to the AMICO DTD. For example, this approach is needed in order to integrate US MARC sources into the integrated view at the mediator.

The architecture approach is based on XML DTDs being available for the metadata describing collections and objects in a collection. This information is queried using BBQ and XMAS. In response to a query, the system returns links to the actual images corresponding to the particular art pieces. The user can follow these links to retrieve the high resolution images. However, at this point the system invokes an authentication/authorization scheme based on X.509 certificates and user information kept in a Lightweight Directory Access Protocol (LDAP) system. The certificate is used to authenticate the user (currently, users must be within the UC system). For authenticated users, the system looks up an LDAP service at CDL to determine the user's access privileges prior to serving the full image.

The high resolution images are retrieved via a distributed storage system middleware, called the Storage Resource Broker (SRB), which provides transparent access to archival storage systems in a distributed environment [SRB]. The SRB system is also developed at SDSC.

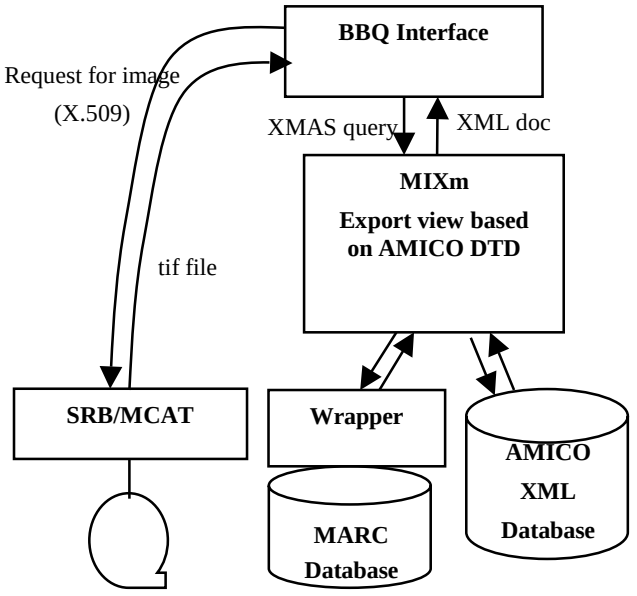

Figure 1. Prototype Architecture for the CDL/AMICO Collection

```
CONSTRUCT <amico_paintings>
                    <title> $Title </>
                    <type> $Type </>
                    <imagelocation> $Img </>
                  </amico_paintings>
WHERE <amico_objects>
            <amico_object>
            <OTY__object_type> $Type </>
            <OTG_object_title_group>
              <OTN__object_title_name> $Title </>
            </OTG_object_title_group>
            <RIG_related_images>
              <RIL__related_image_identifier_link> $Img </>
            </RIG_related_images>
          </amico_object>
        </amico_objects>
IN  "http://www.npaci.edu/DICE/AMICO/Demo/amico-objects.xml"
AND substr("painting", $Type)
```

Figure 2. Sample XMAS query

4. ACKNOWLEDGMENTS

The CDL/AMICO testbed is a joint project between SDSC and CDL and involves several other participants including, Joan Gargano, Robert Brandriff, and Ken Weiss, from CDL; Susan Jurist from the UCSD Library; and, Reagan Moore, Arcot Rajasekar, and Wayne Schroeder, from SDSC.

5. REFERENCES

[1] AMICO: Art Museum Image Consortium. http://www.amico.org/.

[2] CDL: The California Digital Library. http://www.cdlib.org/.

[3] MIX: Mediation of Information using XML. http://www.npaci.edu/DICE/MIX. http://www.db.ucsd.edu/Projects/MIX.

[4] SRB: The SDSC Storage Resource Broker. http://www.npaci.edu/DICE/SRB.