Digital Data Preservation: a schema-driven model

WISP Competition 2005 – Technology Category

Student: Stacy Kowalczyk Co-Authors: Clare McInerney and Phil Mitchell

Digital Data Preservation – the problem

Librarians and archivists have long lamented the fact that much current research is in danger of being lost because it is sitting on computers under desks in offices – professors' completed research, librarians' bibliographies, curators' specialized collection data. Each of these files is different: in data, in format, and in technology. Not only is there no access to the data, there is little hope that the data can be preserved for future generations.   The Harvard University Digital Library Initiative thought this problem was ripe for research. A small team from the Harvard University Library Office for Information Systems was assembled to develop a solution to this digital data preservation problem.

The research question:

Can a system be developed to take an arbitrary data layout, store it in a preservation-quality format and provide access to the contents via the web without a programmer's involvement?

Stacy Kowalczyk, Phil Mitchell, and Clare McInerney developed the prototype system. Using XML schema technology, TED (TEmplated Database) allows a data owner to create a standard database for a collection, define a structured data format, and easily customize screens and parameters for search and display with minimal effort by either the data owner or the systems office.

The Solution

To have a system that could take any arbitrary data structure, the system had to be data independent. So the core solution was to abstract metadata dependencies out of the system into a template layer. Using XML Schema, the TED system takes a formal description of the metadata as input and automatically creates the query and display interface.   XML Schema is the key. TED relies on multiple schemas – the TED schema and the application data schemas. The TED schema describes all of the function points of the system. When the TED schema is used as markup to the application data schema, it becomes the instructions to build the interfaces to the system.

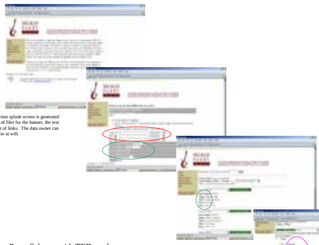
TED has three components – a data loading system with schema-driven indexing; a schema-driven data maintenance system used by the data owner to create, update, and delete instance documents in the database; and a schema-driven web query interface (what librarians call an online catalog). This poster focuses only on the last of these, the web query interface.

The TED web query interface is a Java servlet that runs in Tomcat. It uses a Schema Object Model (SOM) as well as a Document Object Model (DOM). The TED servlet parses the schema to create the user interface in HTML with cascading style sheets. TED uses an XML DBMS, Software AG's Tamino, for the datastore.   Because the underlying datastore is XML, considered a preservation-quality format, the data preservation issues are resolved.

Future Research

TED currently has 2, and soon to have 4, very different application data models running in production. Even with all of the efforts put into the system to be "easy to customize", it still requires a highly technical person to create the application data XML schema. A toolkit needs to be developed for the data owners to create their own schemas.

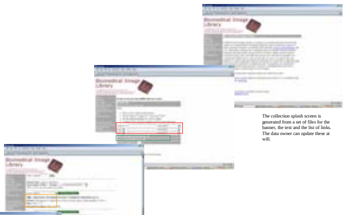
Milman Parry Collection



The collection splash screen is generated from a set of files for the banner, the text and the list of links. The data owner can update these at will.

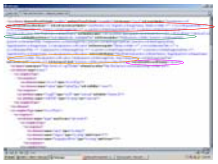
The TED Schema

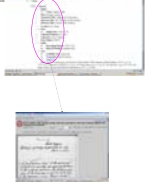
Biomedical Image Library (BIL) Collection



The collection splash screen is generated from a set of files for the banner, the text and the list of links. The data owner can update these at will.

Milman Parry Schema with TED markup





BIL Schema with TED markup