<teiPublisher>: Bridging the Gap Between a Simple Set of Structured Documents and a Functional Digital Library

Amit Kumar¹, Alejandro Bia², Martin Holmes³, Susan Schreibman¹, Ray Siemens⁴, and John Walsh⁵

¹ Maryland Institute for Technology in the Humanities (MITH), University of Maryland, USA
² Miguel de Cervantes DL & DLSI, University of Alicante, Spain
³ University of Victoria Humanities Computing and Media Centre, Canada
⁴ Department of English, Malaspina University College
⁵ Digital Library Program / University Information Technology Services, University of Indiana

Abstract. Digital Libraries are complex systems that take a long time to create and tailor to specific requirements [1]. Their implementation requires specialized computer skills, which are not usually found within humanities text encoding projects. Many encoders working on text encoding projects find they cannot take their work to the next level by transforming their collections of structured XML [2] texts into a publishable web searchable and browsable service. Most often these teams find the way to encode their texts with a high degree of sophistication, but unless they have funds to hire computer programmers their collections remain on local disk storage away from public access. <teiPublisher> is a novel tool designed with the aim of bridging the gap between simply having a collection of structured documents and having a functional digital library for public access via the web. The goal of this project is to build the tools to manage an extensible, modular and configurable XML-based repository which will house, search/browse on, and display documents encoded in TEI-Lite [3] on the World Wide Web. <teiPublisher> provides an administrative interface that allows DL administrators to upload and delete documents from a web accessible repository, analyze XML documents to determine elements for searching/browsing, refine ontology development, decide on inter and intra document links, partition the repository into collections, create backups of the entire repository, generate search/browse and display pages for users of the website, change the look of the interface, and associate XSL transformation scripts and CSS stylesheets to obtain different target outputs (HTML [4], PDF, etc.)

1 Background

Administrators of TEI¹ repositories working in SGML were limited to few databases, such as Dynaweb, to deliver their documents over the World Wide Web. While Dynaweb, to deliver their documents over the World Wide Web. While Dynaweb, to deliver their documents over the World Wide Web. While Dynaweb, to deliver their documents over the World Wide Web. While Dynaweb, to deliver their documents over the World Wide Web. While Dynaweb, to deliver their documents over the World Wide Web.

¹ This is an open source initiative which is being made available through SourceForge (http://teipublisher.sf.net)
² There are only three widely-used general-purpose markup vocabularies: HTML, DOCBOOK and TEI. HTML is more focused on presentational than on structural issues, while the other two are meant exclusively for structural markup. DOCBOOK is more adequate for manuals while TEI is mostly used to structure humanities contents. TEI stands for Text Encoding Initiative, and is run by the TEI-Consortium: http://www.tei-c.org/

© Springer-Verlag Berlin Heidelberg 2004
naweb was revolutionary in its day, throughout the late 1990s, advances in web technology and plugins for HTML rendered the Dynaweb look old fashioned.

With the release of the XML standard in 1998, industry experts predicted that there would be a proliferation of XML-aware software as programmers would find it easier to program applications to deliver XML over the Web. This has indeed come to pass. Over the past few years, a number of open source XML or native XML databases have been developed utilizing the XML:DB API

Programmers in the humanities computing community have begun using these databases for individual projects with great success. However, for projects that cannot afford programming support, the bar is still extremely high. Thus, a group of programmers and content developers teamed to create an extensible, modular and configurable XML-based repository entitled <teiPublisher>, that can house, search on, browse on, and display documents encoded in TEI-Lite. This is an open source initiative which is being made available to the digital library community to allow projects with limited programming support to mount their TEI-Lite encoded texts in a web-deliverable database. The date for the beta release of <teiPublisher> is June 2004.

2 Functionality and Features

<teiPublisher> utilizes the native XML database eXist and upon it, it generates a public interface for browsing and searching. Equally as important, it provides an administrative interface that will help repository administrators with limited technical knowledge to:

- establish an XML repository for TEI-Lite documents;
- upload and delete documents;
- analyze XML documents to determine elements for searching;
- develop ontology consistency and refine ontology development;
- index and store XML documents for efficient search and retrieval;
- generate search/browse, results and metadata display pages for users of the site;
- provide an extensible framework with plugin architecture;
- decide on inter and intra document links;
- partition the repository into collections;
- create backups of the entire repository;
- change the look of the interface;
- associate XSL transformation scripts to transform documents or metadata to different output formats;
- associate CSS stylesheets to control rendering.

Some of the features mentioned above, particularly ontology development, cannot be met by the software alone. Rather, <teiPublisher> provides a helper application to allow content creators to view the content of elements and attributes used in controlled vocabularies, and highlight semantic inconsistencies. It also assists in selecting elements and attributes which will ultimately be searched on.

---

2 Application programming interface for XML Databases: http://www.xmldb.org/xapi/
3 eXist: an Open Source Native XML database: http://exist.sourceforge.net/