Society of Archivists Data Standards Group

A Guide to Archival and Related Standards

Standards applicable to archives; for the digital delivery of repository guides, finding-aids, and images of material from collections.

Title
Text Encoding Initiative (TEI): Guidelines for Electronic Text Encoding and Interchange

Name of Standards Developing Organisation
The TEI Consortium

Current version
(version 1.9.1, updated 5 March 2011)

Replaces

Abstract
TEI P5 Guidelines for Electronic Text Encoding and Interchange describe an open source general purpose Extensible Mark-up Language (XML) conformant schema that provides a standard format for characterising the structure and content of electronic representations of textual information, for machine searching and data exchange. The Guidelines focus on the characterisation of text in digitised material although they are also applicable to born-digital materials. Originating in the humanities research domain the schema is modular to enable it to be comprehensively applicable to a range of materials. It is both flexible and extensible to accommodate the different scholarly purposes of humanities text studies.

Description
TEI is a well-established methodology, with the first Schema and Guidelines released in 1990. Since then a large body of expertise has developed across the humanities; and library and archives domains. Notable implementers are the Oxford Text Archive developed by the University of Oxford and the Electronic Text Center at the University of Virginia.

TEI provides a powerful XML schema and a set of supporting guidelines for marking-up textual documents for machine searching and data exchange. It defines several hundred XML elements and attributes to enable the mark-up of text documents of any kind. It is not expected that all the available tags will be used by an implementation, and tags are included for:

- a range of textual genres such as verse, drama, hand-written manuscripts and transcribed speech;
- syntactical devices such as figures, tables, characters and glyphs;
- description of hierarchical and non-hierarchical structures;
• linguistic and narrative characterisations;
• identification of names, dates and places.

In *The Guidelines* these elements and attributes are organised into modules for specific applications. However, in practice the most applicable elements and attributes for the intended implementation may be used from any module. Therefore implementation requires an analysis of both a project's aims and objectives and the actual materials to be marked-up, before a profile of the TEI tag-set to be used can be defined, or required customisations identified.

Help with defining an applicable profile of TEI is included in *The Guidelines* through an introduction to XML and advice on technical implementation and customisation. A core set of header tags, a set of generic tags which can appear at any point in the structure, and a default document structure are defined. A full data dictionary to all model classes, elements, attributes, data types and macros that can be used is included.

The use of the XML Schema Language means that TEI is flexible enough for other XML schemas, such as MathML (for mathematical notation) or MusicXML (for music notation) to be referenced from inside a TEI document. This enables the standard XML schema for specific genre, not included in TEI, to be used. Meanwhile a TEI document can be embedded within other XML documents such as EAD (Encoded Archival Description) or METS (Metadata Encoding Transmission Standard). This makes it possible for collections of digital representations of textual documents, which are both structured and text searchable, to be embedded in a digital catalogue of an archive or a digital repositories.

The TEI Consortium provides practical help with implementation through the provision of both on-line and face-to-face tutorials. There are also a number of online tools available to help with different aspects of implementation. *Roma* is a Web-based tool which helps with the development of a profile or customisation. A list of authoring and editing tools is maintained and style-sheets are provided to allow transformations for Web delivery of the data.

There is an active user base which maintains an archived discussion list and a wiki of help, advice, activities, meetings and examples. This is supported through 10 Special Interest Groups (SIGs) for different user bases or specialist applications. There is an annual conference and a newly launched online journal to facilitate sharing. Membership of the TEI Consortium brings a number of benefits including discounts on digitisation costs, software, and training.

The TEI P5 schema, *Roma* tool, additional tools and style-sheets are freely available from SourceForge at <http://tei.sourceforge.net>. All other information regarding TEI and the TEI Consortium can be found at <http://www.tei-c.org/index.xml>.

Next month
Next month we will look at the metadata structure standard Dublin Core (ISO 15836:2009) and the Dublin Core Metadata Initiative’s (DCMI) content authority files, DCMI Vocabularies, for use with Dublin Core.

Sarah Higgins
Aberystwyth University