The Text Encoding Initiative Anno 2005: An Orientation and Workshop

Lou Burnard[1], Matthew Driscoll[2]

[1]Information and Support Group, Oxford University Computing Services
13 Banbury Road, Oxford OX2 6NN, England
e-mail: lou.burnard@computing-services.oxford.ac.uk
[2]Department of Scandinavian Research, University of Copenhagen
Njalsgade 136, 2300 Copenhagen, Denmark
e-mail: mjd@hum.ku.dk

Abstract

The Text Encoding Initiative is an international and interdisciplinary standards project established in 1987 to develop, maintain and promulgate hardware- and software-independent methods for encoding humanities data in electronic form. Initially the TEI was jointly sponsored by three established international professional associations (the Association for Computers and the Humanities, the Association for Computational Linguistics and the Association for Literary and Linguistic Computing), which established a small management committee, and appointed two editors to co-ordinate the enthusiastic participation of more than a hundred scholars worldwide. Its remit was to attempt a complete definition of current practice and to produce recommendations or Guidelines for the creation and usage of electronic texts in key linguistic and literary disciplines. The first research phase of the TEI came to an end in 1994 with the publication of TEI P3, which over the next few years was to become the reference standard for the building of the digital library. At the start of the current century, the TEI re-established itself as a membership consortium, jointly hosted by four institutions, two on either side of the Atlantic, and managed by a Board of Directors and its technical work overseen by an elected Council. There are currently nearly 100 members of the Consortium, among them universities, research libraries, academic and other publishers, both non-profit and commercial, as well as scholarly societies and research projects concerned with the design, production and delivery of structured electronic text. One does not, of course, need to be a member of the Consortium to use the TEI, and indeed there are thousands of users worldwide. In 2002 the first major revision of the Guidelines, known as TEI P4, was published. This was a "maintenance release", seeking only to bring the Guidelines up to date with changes in the technical infrastructure, most notably in the use of the W3C's Extensible Markup Language (XML) as its means of expression, rather than the ISO standard SGML used by earlier editions. Since then, work has been proceeding on a complete revision of the TEI Guidelines, TEI P5, initial releases of which are now available from Source Forge. In this presentation we will sketch some of the changes introduced with this new release and will also show how P5's modular construction facilitates the customisation of the TEI scheme for use in a wide variety of literary and linguistic study, and speculate about its likely implications for future encoding work consequent on the internationalisation effects currently under way. The session is intended as "an orientation and workshop", meaning that there will be both a theoretical and a practical aspect. Previous experience with TEI mark-up would be a decided advantage, but is not required.

Keywords: text encoding; XML, description and transcription of primary sources

Plan for the Workshop

Session 1 (14.00-15.30)

a) Introduction to the TEI: where it came from, what it covers, and how it can be used (LB)

b) Basic components of TEI (LB)
This section provides an introduction to the basic concepts underlying TEI mark-up and presents mechanisms for dealing with such things as text and document structure, character and glyph mark-up, direct speech, figures, tables and lists, and bibliographical elements.

c) Basic markup techniques for the transcription of primary sources (MJD)
In this sections we will look in greater detail at how to encode features of manuscripts and other hand-written sources such as abbreviations, scribal corrections, deletions and additions, text layout and editorial emendations.

Session 2 (16.00-17.30)

a) Manuscript description (MJD)
This section will look at the new Manuscript Description module, which can be used to provide detailed descriptive information about handwritten primary sources. Although originally developed to meet the needs of cataloguers and scholars working with medieval manuscripts in the European tradition, the scheme presented here is general enough that it can also be extended to other traditions and materials, and is potentially useful for any kind of inscribed artefact.

b) Mark up of biographical and prosopographical material (MJD)
This section will present work currently under way within the TEI for marking-up biographical and prosopographical data, in other words information on people including such things as birth and death, marriage and family relations, social origins, place of residence, education, occupation, religion, experience of office and so on.

c) Using Xaira to index TEI documents (LB)
The section will introduce Xaira, the text searching software originally developed at Oxford University Computing Services for use with the British National Corpus but now entirely re-cast as a general purpose XML search engine. It will operate on any corpus of well-formed XML documents but is best used with TEI-conformant documents. Xaira has full Unicode support, which means that it can be used to search and display text in any language.