TEI

Øyvind Eide, Unit for Digital Documentation, University of Oslo. Convener of the TEI Ontologies SIG together with Christian-Emil Ore.

What is TEI

TEI is a consortium of institutions and individuals from all over the world. The TEI is also a set of guidelines for the encoding of textual material, and it is a set of computer readable files. The guidelines and the computer readable files specify a set of rules documents have to adhere to in order to be accepted as TEI documents.

One of the main goals of TEI is to capture a wide range of intellectual work. It is used extensively in areas such as edition philology, but it is quite possible to create a detailed encoding scheme for e.g. archaeological grey documents and include it as a local extension of TEI, creating documents very different from any other TEI document (Eide 2006).

According to the TEI guidelines (TEI P5 section 1.2.2), there are three primary functions of the guidelines:

- guidance for individual or local practice in text creation and data capture;
- support of data interchange;
- support of application-independent local processing.

TEI is also an important discussion forum, or a set of fora, both through meetings and on-line media, as well as through articles and books. Few encoding problems will be faced that has not been discussed somewhere in the guidelines or in the literature surrounding TEI.

TEI encoding: From text to world

TEI is created for the encoding of texts. This means that the object of a TEI encoding is a text. To be able to find out what to encode and how to do it, the encoder have to read and understand the text. A very simple example is the word "Bergen" in a German text. Based on the interpretation of the context, the word may mean mountains or the place named Bergen.

Traditionally, the object of a TEI encoding has been the text as such. This means that the place name Bergen is encoded as a place name, e.g. <name type="place">Bergen</name>. But there have not been a strong tradition for building up models within TEI for the place to which Bergen is used as a name. From version 3 of the guidelines published in the late 1990's, I quote from the introduction to the chapter on names and dates:

"It should be noted however that no provision is made by the present tag set for the representation of the abstract structures, or “virtual objects” to which names or dates may be said to refer. In simple terms, where the core tag set allows one to represent a *name*, this additional tag set allows one to represent a *personal name*, but neither provides for the direct representation of a *person*. Appropriate mechanisms for the encoding of such interpretative gestures may be found in chapters 15 (‘Simple Analytic Mechanisms’) on p. 378 and 16 (‘Feature Structures’) on p. 394." (TEI P3, p. 483)

But in the almost ten years gone since the final version of P3 was published, the tools suggested for encoding persons and places as opposed to their names was clearly not sufficient for the users of TEI. In the current version of the TEI, P5 version 0.5, a similar paragraph is included at the same place in the guidelines, but with quite different wording:

"Finally, when this module is included, elements are provided to represent the abstract structures, or virtual objects to which names or dates may be said to refer. In simple terms, where the core module allows one to

represent a *name*, this module allows one further to represent a *personal name*. It also allows one to represent the *person* being named, and thus to encode biographical or other personal data for a wide range of applications, quite distinct from the names associated with such data." (TEI P5, chapter 20)

The line that is drawn between name and person in the citation from P3 above does not represent an absolute opposition. As an examples of this, present already in P3, consider the element *hand*. One may argue that the *hand* element is used in the description of the process of writing the document, and is thus external to the text in a way not commonly encoded in TEI apart from the specific interpretation chapters (15 and 16 in P3). Seen this way, it may be as close to a person as to a person's name.

A much clearer need for encoding information about persons was demonstrated through work on methods for the encoding of prosopographical data in TEI. Several element was included in P5 to encode information about persons, among them sex, faith and occupation. Elements for encoding events in a persons life is also included, such as birth and death. Work is currently being done to build up an an encoding system for places within TEI along the same lines, and a general event element is also being discussed.

TEI Ontologies SIG

In May and June 2004, there was a discussion on the TEI mailing list about prosopographical tags. This lead to a suggestion that detailed information about persons (physical and legal), dates, events, places, objects etc. and their interpretation could be marked up outside the text, and that this could be connected to on-going ontology work being done e.g. in the Museum community, such as the Conceptual Reference Model (CIDOC-CRM). The result of this was the establishment of a Ontologies SIG at the 4th annual members meeting of TEI in October 2004 (TEI Ontologies SIG).

During the meetings of the Documentation Standards Group of The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) in Gothenburg in 2006, liaison with other parties was discussed and the work in the TEI Ontologies SIG came up. The group expressed support for the work done by the Unit for Digital Documentation at the University of Oslo on the relationship between CIDOC-CRM and TEI, and hoped the work would continue. This support includes, of course, the work being done by other parties on the same topic.

During the two and a half years since the establishment of the SIG, work has mainly been done in relation to the CIDOC-CRM ontology. Work was reported and discussed at the Digital Humanities conference in Paris in July 2006 (Ore 2006). Based on the discussions in Paris as well as presentations and discussions at a meeting in Heraklion in Greece in October 2006 and at the SIG meeting in Victoria in October 2006, a draft mapping of a set of TEI elements to CIDOC-CRM was published in January this year (Eide 2007). Relevant work in this area has also been reported by Conal Tuohy (2005).

Gregory Crane of the Perseus Project said at the SIG meeting in October 2006 that he would make sure work will be done on the integration between FRBR and TEI, mainly the TEI header. We hope to receive reports on this work on the mailing list of the SIG later this year.

The way forward

The work of the SIG will continue, as well as connected work in other TEI bodies. I will take part in a TEI work meeting next week to discuss a TEI encoding scheme for places.

In our work on a mapping system from TEI to CIDOC-CRM, we have identified some problems reducing the potential benefit from such work. The discussion below is based on a conference poster to be presented at the Digital Humanities conference in June this year.

If a mapping from a TEI document into a model conforming with CRM is to be created, it will naturally be

based on a general mapping of TEI elements to CRM we are currently developing. But in TEI, many elements are defined quite loose, and depending on the way they are used, they may be modelled differently in CRM. According to the TEI guidelines, tag usage may be described in the TEI header. Such descriptions should then help us in deciding which type of modelling is the most appropriate.

Ideally, such a specific mapping should be created based on an automatic reading of the TEI header. But an element description in a *tagUsage* element in the TEI header is in prose and will generally not be stringent enough to be understood by an automatic reading (TEI P5, sec. 5.3.4). Human interaction will be needed. It may be the case that use of the *equiv* element will make automatic creation of mappings possible, as a reference to a certain CRM class may be included as an external link (ibid, sec. 6.3.4).

A CRM conforming model based on the TEI document and populated with all instances of mapped elements should then be created. This model may be used as a query or a data mining system where the user looks for interesting structures in the CRM conforming model alone, as well as in combination with textual information collected from the TEI source document. But this model may also be used in connection with other CRM conforming models, such as museum databases. The connections will be based on regional or global object identification, such as authority lists of names and classification schema. The resulting "super model" may then be used as a data mining tool based on semantic integration between heterogeneous resources.

We are currently developing the building blocks for a system based on this method. We believe this will improve the usability of TEI documents as information sources as well as simplifying the process of adding more information, such as event elements, into such documents. A general observation from our work is that the more relevant information types is marked up in an TEI document, the easier it is to use automatic methods to generate CRM conforming models. But even a limited tagging with only names and dates marked up do help in such an automatic model generation.

Bibliography

CIDOC (2003). *Definition of the CIDOC Conceptual Reference Model* / Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. ISO/DIS 21127. URL: http://cidoc.ics.forth.gr/de nition_cidoc.html (as of 2006-11-13)

Eide, Øyvind and Jon Holmen (2006). "Reading Gray Literature as Texts. Semantic Mark-up of Museum Acquisition Catalogues". CIDOC 2006, Gothenburg. URL: http://www.edd.uio.no/artiklar/teknikk_informatikk/CIDOC2006/EIDE_HOLMEN_Reading_Gray_Literature.pdf (cheked 2007-02-02)

Eide, Øyvind and Christian-Emil Ore (2007). *Mapping of TEI to CIDOC-CRM. Version 0.1 2007-01-02*. URL: http://www.edd.uio.no/artiklar/tekstkoding/tei_crm_mapping.html (checked 2007-02-02)

Ore, Christian-Emil and Øyvind Eide (2006). "TEI, CIDOC-CRM and a Possible Interface between the Two." P. 62-65 in *Digital Humanities 2006. Conference Abstracts.* Paris, 2006.

TEI P3 (1999). *Guidelines for Electronic Text Encoding and Interchange* / Edited by C. M. Sperberg-McQueen and Lou Burnard. Chicago, Oxford, 1994. Revised Reprint, Oxford, May 1999.

TEI P5 (2006). *Guidelines for Electronic Text Encoding and Interchange*. Ver. 0.5. URL: http://www.tei-c.org/release/doc/tei-p5-doc/html/ (checked 2006-11-13)

TEI Ontologies SIG. *SIG:Ontologies Wiki.* URL: http://www.tei-c.org.uk/wiki/index.php/SIG:Ontologies (checked 2007-02-02)

Tuohy, Conal (2005). "Topic Maps @ NZETC." URL: www.nzetc.org/downloads/TM@NZETC.ppt (checked 2006-11-13)