THE TEXT ENCODING INITIATIVE

Edward Vanhoutte & Ron Van den Branden

Centre for Scholarly Editing and Document Studies

Royal Academy of Dutch Language and Literature - Belgium

Koningstraat 18

9000 Gent

Belgium

edward.vanhoutte@kantl.be

ron.vandenbranden@kantl.be

KEYWORDS: text-encoding, markup, markup languages, XML, humanities

ABSTRACT

The result of community efforts among computing humanists, the Text Encoding Initiative or TEI is the *de facto* standard for the encoding of texts in the humanities. This article explains the historical context of the TEI, its ground principles, history, and organisation.

INTRODUCTION

The Text Encoding Initiative (TEI) is a standard for the representation of textual material in digital form through the means of text encoding. This standard is the collaborative product of a community of scholars, chiefly from the humanities, social sciences, and linguistics who are organized in the TEI Consortium (TEI-C <http://www.tei-c.org>). The TEI Consortium is a non-profit membership organisation and governs a wide variety of activities such as the development, publication, and maintenance of the text encoding standard documented in the *TEI Guidelines*, the discussion and development of the standard on the TEI mailing list (TEI-L) and in Special Interest Groups (SIG), the gathering of the TEI community on yearly members meetings, and the promotion of the standard in publications, on workshops, training courses, colloquia, and conferences. These activities are generally open to non-members as well.

By ‘TEI Guidelines’ one may refer both to the markup language and tag set proposed by the TEI Consortium and to its documentation online or in print. Informally ‘TEI Guidelines’ is often abbreviated to ‘TEI’. In this article ‘TEI Guidelines’ is used as the general term for the encoding standard. The *TEI Guidelines* are widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. Since the TEI is expressed in terms of the eXtensible Markup Language (XML) and since it provides procedures and mechanics to adapt to one’s own project needs, the *TEI Guidelines* define an open standard that is generally applicable to any text and purpose.

The article first introduces the concepts of text encoding and markup languages in the humanities and then introduces the TEI encoding principles. Next, the article provides a brief historical survey of the TEI Guidelines and ends with a presentation of the Consortium's organisation.

TEXT ENCODING IN THE HUMANITIES

Since the earliest uses of computers and computational techniques in the humanities at the end of the 1940s, scholars, projects, and research groups had to look for systems that could provide *representations* of data which the computer could process. Computers, as Michael Sperberg-McQueen has reminded us are binary machines that 'can contain and operate on patterns of electronic charges, but they cannot contain numbers, which are abstract mathematical objects not electronic charges, nor texts, which are complex, abstract cultural and linguistic objects.' (1, p. 34) This is clearly seen in the mechanics of early input devices such as punched cards where a hole at a certain coordinate actually meant a I or 0 (true or false) for the character or numerical represented by this coordinate according to the specific character set of the computer used. Because different computers used different character sets with a different number of characters, texts first had to be transcribed into that character set. All characters, punctuation marks, diacritics, and significant changes of type style had to be encoded with an inadequate budget of characters. This resulted in a complex of 'flags' for distinguishing upper-case and lower-case letters, for coding accented characters, the start of a new chapter, paragraph, sentence, or word. These 'flags' were also used for adding analytical information to the text such as word classes, morphological, syntactic, and lexical information. Ideally, each project used its own set of conventions consistently throughout. Since this set of conventions was usually designed on the basis of an analysis of the textual material to be transcribed to machine readable text, another corpus of textual material would possibly need another set of conventions. The design of these sets of conventions were also heavily dependent on the nature and infrastructure of the project, such as the type of computers, software, and devices such as magnetic tapes of a certain kind that were available.

Although several projects were able to produce meaningful scholarly results with this internally consistent approach, the particular nature of each set of conventions or encoding scheme had lots of disadvantages. Texts prepared in such a proprietary scheme by one project could not readily be used

by other projects; software developed for the analysis of such texts could hence not be used outside the project due to an incompatibility of encoding schemes and non-standardization of hardware. However, with the increase of texts being prepared in machine-readable format, the call for an economic use of resources increased as well. Already in 1967, Michael Kay argued in favour of a 'standard code in which any text received from an outside source can be assumed to be.' (2, p. 171) This code would behave as an exchange format which allowed the users to use their own conventions at output and at input. (2, p. 172)

MARKUP LANGUAGES IN THE HUMANITIES

Descriptive markup

Whereas markup languages in use in the typesetting community were mainly of a procedural nature, that is they indicate procedures that a particular application should follow, the humanities were mainly considered with descriptive markup that identify the entity type of tokens, and with referential markup that refers to entities external to the documents, e.g. in order to encode characters outside the current characters set. Unlike procedural or presentational markup, descriptive markup establishes a one to one mapping between logical elements in the text and its markup.

Early attempts

Some sort of standardization of text encoding for the encoding and analysis of literary texts was reached by the COCOA encoding scheme originally developed for the COCOA program in the 1960s and 1970s (3), but used as an input standard by the Oxford Concordance Program (OCP) in the 1980s (4) and by the Textual Analysis Computing Tools (TACT) in the 1990s (5). For the transcription and encoding of classical Greek texts, the Beta-transcription/encoding system reached some level of standardized use (6).

The Standard Generalized Markup Language (SGML)

The call for a markup language that could guarantee reusability, interchange, system- and software-independence, portability and collaboration in the humanities was answered by the publication of the Standard Generalized Markup Language (SGML) as an ISO standard in 1986 (ISO 8879:1986) (7) Based on IBM's *Document Composition Facility Generalized Markup Language*, SGML was developed mainly by Charles Goldfarb as a metalanguage for the description of markup schemes that satisfied at least seven requirements for an encoding standard (8, p. 28-29):

1. The requirement of comprehensiveness;
2. The requirement of simplicity;
3. The requirement that documents be processable by software of moderate complexity;
4. The requirement that the standard not be dependent on any particular characteristic set or text-entry devise;
5. The requirement that the standard not be geared to any particular analytic program or printing system;
6. The requirement that the standard should describe text in editable form;
7. The requirement that the standard allow the interchange of encoded texts across communication networks.

In order to achieve universal exchangeability and software and platform independence, SGML made use exclusively of the ASCII codes. As mentioned above, SGML is not a markup language itself, but a metalanguage by which one can create separate markup languages for separate purposes. This means that SGML defines the rules and procedures to specify the vocabulary and the syntax of a markup language in a formal Document Type Definition (DTD). Such a DTD is a formal description of, for instance, names for all elements, names and default values for their attributes, rules about how elements can nest and how often they can occur, and names for re-usable pieces of data (entities). The DTD allows full control, parsing, and validation of SGML encoded

documents. By and large the most popular SGML DTD is the Hypertext Markup Language (HTML) developed for the exchange of graphical documents over the internet.

A markup scheme with all these qualities was exactly what the humanities were looking for in their quest for a descriptive encoding standard for the preparation and interchange of electronic texts for scholarly research. There was a strong consensus among the computing humanists that SGML offered a better foundation for research oriented text encoding than other such schemes. (8) (9)

From the beginning, however, SGML was also criticized for at least two problematic matters: SGML's hierarchical perspective on text, and SGML's verbose markup system. (9) These two issues have since been central to the theoretical and educational debates on markup languages in the humanities.

The eXtensible Markup Language (XML)

The publication of the eXtensible Markup Language (XML) 1.0 as a W3C recommendation in 1998 (10) brought together the best features of SGML and HTML and soon achieved huge popularity. Among the power XML borrowed from SGML are the explicitness of descriptive markup, the expressive power of hierarchic models, the extensibility of markup languages, and the possibility to validate a document against a DTD. From HTML it borrowed simplicity and the possibility to work without a DTD. Technically speaking, XML is a subset of SGML and the recommendation was developed by a group of people with a long standing experience in SGML, many of whom were TEI members. As Steven DeRose pointed out, XMLs advantages to the TEI community were substantial by its provisions to capitalize the large investments in SGML expertise, data, and software: XML is simple enough to be understood and supported by browsers so that the structural information of TEI documents can be delivered to the end user with cheap ubiquitous tools; XMLs approach allows servers to deliver subtrees directly without the need to ship whole documents over network connections; XML separates the validation function of DTDs from the parse-enabling

function so that there is no need to re-validate every document on every reading; XML spreads the notion of decriptive markup to a wide audience that will thus be acquainted with the concepts articulated in the TEI Guidelines. (11, p. 19)

TEI: GROUND RULES

Guidelines

The conclusions and the work of the TEI community are formulated as guidelines, rules, and recommendations rather than standards, because it is acknowledged that each scholar must have the freedom of expressing their own theory of text by encoding the features they think important in the text. A wide array of possible solutions to encoding matters is demonstrated in the TEI Guidelines which therefore should be considered a reference manual rather than a tutorial. Mastering the complete TEI encoding scheme implies a steep learning curve, but few projects require a complete knowledge of the TEI. Therefore, a manageable subset of the full TEI encoding scheme was published as TEI Lite, currently describing 145 elements. (12) Originally intended as an introduction and a didactic stepping stone to the full recommendations, TEI Lite has, since its publication in 1995, become one of the most popular TEI customizations and proves to meet the needs of 90% of the TEI community, 90% of the time.

The ground rules that are discussed in this section apply to the most recent version of the TEI at the time of writing, i.e. TEI P5. See the section on TEI:History for more details about P5 and previous versions of the TEI Guidelines.

Text structure

The TEI Guidelines (13) define a set of rules to mark up the phenomena in a wide range of texts in a descriptive fashion. This means that for example, encoders can (and should) not just indicate *that* a bit of text is printed in italics, but *why* this is the case: either because it appears to be a title in a

bibliographical reference, or a technical term, a foreign word, a regular word with rhetorical emphasis, and so on. Texts that are enriched with intelligible meta-information not only can be preserved and reused more easily by humans, but also can be processed more intelligently by computers.

The TEI guarantees this potential by imposing a common structure for all texts:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt>
    <title><!--Title--></title>
   </titleStmt>
   <publicationStmt>
    <p><!--Publication Information--></p>
   </publicationStmt>
   <sourceDesc>
    <p><!--Information about the source--></p>
   </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text xml:id="text1">
  <body>
   <p>This is the first paragraph</p>
  </body>
 </text>
</TEI>
```

NOTE: This is an example of a TEI XML text, representing both information and meta-information. Information (plain text) is contained in *XML elements*, delimited by *start tag*s (eg: <TEI>) and *end tag*s (eg: </TEI>). Additional information to these XML elements can be given in *attributes*, consisting of a name (eg: xml:id) and a value (eg: text1). *XML comment*s are delimited by start markers (<!--) and end markers (-->).

This example, as any TEI text, is recognisable as a TEI text by the outermost <*TEI*> element, which is declared in the dedicated TEI namespace (*http://www.tei-c.org/ns/1.0*). The TEI considers texts units of information that are composed of two mandatory parts:

- a header (<*teiHeader*>) containing descriptive meta-information. This should minimally document following aspects of the electronic file itself (<*fileDesc*>):
    - the title statement (<*titleStmt*>), providing information about the title, author and others responsible for the electronic text
    - the publication statement (<*publicationStmt*>), providing publication details about the electronic text
    - a description of the source (<*sourceDesc*>), documenting bibliographic details about the electronic text's material source (if any)
- the actual text (<*text*>) containing meta-information about the text's structure and the actual text. This should minimally contain a text body (<*body*>). The body contains lower-level text structures like paragraphs (<*p*>), or different structures for text genres other than prose: lines for poetry, speeches for drama.

Apart from simple texts, TEI provides means to encode composite texts, either by grouping structurally related texts in a <*group*> element inside <*text*>, or treating them as a corpus of diverse texts, using <*teiCorpus*> as the outermost element.

TEI Modules

As illustrated by this example, a significant part of the rules in the TEI Guidelines apply to the expression of descriptive and structural meta-information about the text. Yet, the TEI defines concepts to represent a much wider array of textual phenomena, amounting in a total of 503 elements and 210 attributes. These are organised into 21 modules, grouping related elements and attributes:

1. The TEI Infrastructure

   Definition of common datatypes and modular class structures used to define the elements and attributes in the other modules.

2. The TEI Header (54 elements)

   Definition of the elements that make up the header section of TEI documents. Its major parts provide elements to encode detailed metadata about bibliographic aspects of electronic texts, their relationship with the source materials from which they may have been derived, non-bibliographic details, and a complete revision history.

3. Elements Available in All TEI Documents (78 elements)

   Definition of elements and attributes that may occur in any TEI text, of whatever genre. These elements cover textual phenomena like paragraphs, highlighting and quotation, editorial changes (marking of errors, regularisations, additions), data-like structures (names, addresses, dates, numbers, abbreviations), cross-reference mechanisms, lists, notes, graphical elements, bibliographic references, and passages of verse or drama.

4. Default Text Structure (33 elements)

   Definition of elements and attributes that describe the structure of TEI texts, like front matter and title pages, text body, and back matter. These may contain further divisions, possibly introduced by headings, salutations, opening formulae, and/or concluded by closing formulae, closing salutations, trailing material and postscripts.

5. Representation of Non-standard Characters and Glyphs (11 elements)

   Definition of specific provisions for representing characters for which no standardised representation (such as defined by the *Unicode Consortium* <http://www.unicode.org/>) exists.

6. Verse (4 elements)

   Definition of specific elements and attributes for dedicated analysis of verse materials, such as caesurae, metrical systems, rhyme schemes, and enjambments.

7. Performance Texts (16 elements)

   Definition of specific elements and attributes for dedicated analysis of drama materials. These include provisions for encoding specific phenomena in front and back matter, like details about performances, prologues, epilogues, the dramatic setting, and cast lists. Other drama-specific structures include speeches and stage directions. For multimedia performances, elements for the description of screen contents, camera angles, captions, and sound are provided.

8. Transcriptions of Speech (12 elements)

   Definition of elements and attributes for (general purpose) transcription of different kinds of spoken material. These cover phenomena like utterances, pauses, non-lexical sounds, gestures, and shifts in vocal quality. Besides this, specific header elements for describing the vocal source of the transcription are provided.

9. Dictionaries (35 elements)

   Definition of elements and attributes for representing dictionaries, with provisions for unstructured and structured dictionary entries (possibly grouped). Dictionary entries may be structured with a number of specific elements indicating homonyms, sense, word form, grammatical information, definitions, citations, usage, and etymology.

10. Manuscript Description (63 elements)

Definition of specific header and structural elements and attributes for the encoding of manuscript sources. Header elements include provisions for detailed documentation of a manuscript's or manuscript part's identification, heading information, contents, physical description, history, and additional information. Dedicated text elements cover phenomena like catchwords, dimensions, heraldry, watermarks, and so on.

11. Representation of Primary Sources (16 elements)

    Definition of elements and attributes for detailed transcription of primary sources. Phenomena covered are facsimiles, more complex additions, deletions, substitutions and restorations, document hands, damage to the source material and illegibility of the text.

12. Critical Apparatus (13 elements)

    Definition of elements and attributes for the representation of (different versions texts as) scholarly editions, listing all variation between the versions in a variant apparatus.

13. Names, Dates, People, and Places (50 elements)

    Definition of elements and attributes for more detailed analysis of names of persons, organisations, and places, their referents (persons, organisations, and places) and aspects of temporal analyses.

14. Tables, Formulæ, and Graphics (6 elements)

    Definition of specific elements and attributes for detailed representation of graphical elements in texts, like tables, formulae, and images.

15. Language Corpora (14 elements)

    Definition of elements and attributes for the encoding of corpora of texts that have been collected according to specific criteria. Most of these elements apply to the documentation of these sampling criteria, and contextual information about the texts, participants, and their communicative setting.

16. Linking, Segmentation, and Alignment (11 elements)

Definition of elements and attributes for representing complex systems of cross-references between identified anchor places in TEI texts. Recommendations are given for either in-line or stand-off reference mechanisms.

17. Simple Analytic Mechanisms (10 elements)

    Definition of elements and attributes that allow the association of simple analyses and interpretations with text elements. Mechanisms for the representation of both generic and particularly linguistic analyses are discussed.

18. Feature Structures (28 elements)

    Definition of elements and attributes for constructing complex analytical frameworks that can be used to represent specific analyses in TEI texts.

19. Graphs, Networks, and Trees (12 elements)

    Definition of elements and attributes for the analytical representation of schematic relationships between nodes in graphs and charts.

20. Certainty and Responsibility (2 elements)

    Definition of elements for detailed attribution of certainty for the encoding in a TEI text, as well as the identification of the responsibility for these encodings.

21. Documentation Elements (35 elements)

    Definition of elements and attributes for the documentation of the encoding scheme used in TEI texts. This module provides means to define elements, attributes, element and attribute classes, either by changing existing definitions or by creating new ones.

Each of these modules and the use of the elements they define are discussed extensively in a dedicated chapter of the TEI Guidelines .

Using TEI

In order to use TEI for the encoding of texts, users must make sure that their texts belong to the TEI namespace (http://www.tei-c.org/ns/1.0) and adhere to the requirements of the text model proposed by the TEI. In order to facilitate this adherence, it is possible (and strongly suggested) to associate TEI texts with formal representations of this text model. These formal *structural grammars* of a TEI compatible model of the text can be expressed in a number of ways, commonly referred to as a *TEI schema*. Technically, a TEI schema can be expressed in a variety of formal languages such as *Document Type Definition* (<http://www.w3.org/TR/REC-xml/#dt-doctype>), *W3C XML Schema* (<http://www.w3.org/XML/Schema>), or the *RELAX NG* schema language (<http://www.relaxng.org/>). It is important to notice that no such thing as 'the TEI schema' exists. Rather, users are expected to select their desired TEI elements and attributes from the TEI modules, possibly with alterations or extensions where required. In this way, TEI offers a stable base with unambiguous means for the representation of basic textual phenomena, while providing standardised mechanisms for user customisation for uncovered features. It is a particular feature of TEI that these abstract text models themselves can be expressed as TEI texts, using the documentation elements defined in the dedicated module *Documentation Elements*. A minimal TEI customisation file looks as follows:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>A TBE customisation</title>
        <author>The TBE Crew</author>
      </titleStmt>
```

```
        <publicationStmt>
          <p>for use by whoever wants it</p>
        </publicationStmt>
        <sourceDesc>
          <p>created on Thursday 24th July 2008 10:20:17 AM by the form at http://www.tei-
c.org.uk/Roma/</p>
        </sourceDesc>
      </fileDesc>
    </teiHeader>
    <text>
      <front>
        <divGen type="toc"/>
      </front>
      <body>
        <p>My TEI Customization starts with modules tei, core, header, and textstructure</p>
        <schemaSpec ident="TBEcustom" docLang="en" xml:lang="en" prefix="">
          <moduleRef key="tei"/>
          <moduleRef key="header"/>
          <moduleRef key="core"/>
          <moduleRef key="textstructure"/>
        </schemaSpec>
      </body>
    </text>
</TEI>
```

Besides the common minimal TEI structure (header and text), a TEI customisation file has one specific element which defines the TEI schema (<schemaSpec>). A TEI schema must minimally include the modules which define the minimal TEI text structure: the *TEI infrastructure* module, the *core* module with all common TEI elements, the *header* module defining all teiHeader elements, and the *textstructure* module defining the elements representing the minimal structure of TEI texts.

In the vein of *Literary Programming* <http://www.literateprogramming.com/>, a TEI customisation file not only contains the formal declaration of TEI elements inside <schemaSpec>, but may also contain prose documentation of the TEI encoding scheme it defines. Consequently, TEI customisation files are commonly called *ODD files* (One Document Does it all), because they serve as a source for the derivation of

- a formal TEI schema
- human-friendly documentation of the TEI encoding scheme

In order to accommodate the process of creating customised TEI schemas and prose documentation, the TEI has developed a dedicated piece of software called *Roma* <http://www.tei-c.org/Roma/>. This is a dedicated ODD processor, offering an intuitive web-based interface for the creation and basic editing of ODD files, generation of according TEI schemas and prose documentation in a number of presentation formats.

A TEI schema, stating all structural conditions and restraints for the elements and attributes in TEI texts can then be used to automatically validate actual TEI documents with an XML parser. Consider, for example, following fragments:

[A]	[B]
<TEI xmlns="http://www.tei-c.org/ns/1.0"> <teiHeader> <fileDesc>	<TEI xmlns="http://www.tei-c.org/ns/1.0"> <text> <body>

```
    <titleStmt>
      <title>A sample TEI document</title>
    </titleStmt>
    <publicationStmt>
      <publisher>
        Taylor &amp;amp; Francis
      </publisher>
      <pubPlace>London</pubPlace>
      <date when="2008"/>
    </publicationStmt>
    <sourceDesc>
      <p>No source, born digital</p>
    </sourceDesc>
  </fileDesc>
 </teiHeader>
 <text>
  <body>
    <p>This is a sample paragraph, illustrating a
<name type="organisation">TEI</name>
document.</p>
  </body>
 </text>
</TEI>
```

```
    <p>This is a sample paragraph, illustrating a
<orgName>TEI</name> document.</p>
  </body>
 </text>
</TEI>
```

When validated against a TEI schema derived from the previous ODD file, file [A] will be recognised as a valid TEI document, while file [B] won't:

- The TEI prescribes that the <teiHeader> *must* be present in each document, and that it precede the <text> part.
- The minimal set of TEI modules does not include the specialised <orgName> element. Although it is a TEI element, using it requires selection of the appropriate TEI module in the ODD file (in this case, the module for *Names, Dates, People, and Places*).

TEI: HISTORY

Poughkeepsie Principles

Shortly after the publication of the SGML specification as an ISO Standard, a diverse group of 32 humanities computing scholars gathered at Vassar College in Poughkeepsie, New York in a two-day meeting (11 & 12 November 1987) called for by the Association for Computers and the Humanities (ACH <http://www.ach.org>), funded by the National Endowment for the Humanities (NEH), and convened by Nancy Ide and Michael Sperberg McQueen. The main topic of the meeting was the question how and whether an encoding standard for machine-readable texts intended for scholarly research should be developed. Amongst the delegates were representatives from the main European text archives and from important North American academic and commercial research centres. Contrary to the disappointing outcomes of other such meetings in San Diego in 1977 or in Pisa in 1980, this meeting did reach its goal with the formulation and the agreement on the following set of methodological principles – the so called *Poughkeepsie Principles* – for the preparation of text encoding guidelines for literary, linguistic, and historical research (14, p. 132-133) (15, p. E.6-4) (16, p. 6.):

1. The guidelines are intended to provide a standard format for data interchange in humanities research.
2. The guidelines are also intended to suggest principles for the encoding of texts in the

same format.

3. The guidelines should
    a) define a recommended syntax for the format,
    b) define a metalanguage for the description of text-encoding schemes,
    c) describe the new format and representative existing schemes both in that metalanguage and in prose.
4. The guidelines should propose sets of coding conventions suited for various applications.
5. The guidelines should include a minimal set of conventions for encoding new texts in the format.
6. The guidelines are to be drafted by committees on
    a) text documentation
    b) text representation
    c) text interpretation and analysis
    d) metalanguage definition and description of existing and proposed schemes

coordinated by a steering committee of representatives of the principal sponsoring organizations.

7. Compatibility with existing standards will be maintained as far as possible.
8. A number of large text archives have agreed in principle to support the guidelines in their function as an interchange format. We encourage funding agencies to support development of tools to facilitate this interchange.
9. Conversion of existing machine-readable texts to the new format involves the translation of their conventions into the syntax of the new format. No requirements will be made for the addition of information not already coded in the texts.

For the implementation of these principles the ACH was joined by the Association for Literary and

Linguistic Computing (ALLC <http://www.allc.org>) and the Association for Computational Linguistics (ACL <http://www.aclweb.org/>). Together they established the Text Encoding Initiative (TEI) whose mission it was to develop the *Poughkeepsie Principles* into workable text-encoding guidelines. The Text Encoding Initiative very soon came to adopt SGML, published a year before as ISO standard, as its framework. Initial funding was provided by the US National Endowment for the Humanities, Directorate General XIII of the Commission of the European Communities, the Canadian Social Science and Humanities Research Council, and the Andrew W. Mellon Foundation.

TEI P1 and TEI P2

From the Poughkeepsie Principles the TEI concluded that the TEI Guidelines should:

- Provide a standard format for data interchange;
- Provide guidance for encoding of texts in this format;
- Support the encoding of all kinds of features of all kinds of texts studied by researchers;
- Allow the rigorous definition and efficient processing of texts;
- Provide for user-defined extensions;
- Be application independent;
- Be simple, clear, and concrete;
- Be simple for researchers to use without specialized software.

A Steering Committee consisting of representatives of the ACH, the ACL, and the ALLC appointed Michael Sperberg-McQueen as editor-in-chief and Lou Burnard as European editor of the Guidelines.

The first public proposal for the TEI Guidelines was published in July 1990 under the title *Guidelines for the Encoding and Interchange of Machine-Readable Texts* with the TEI document number TEI P1 (for Proposal 1). This version was reprinted with minor changes and corrections, as

version 1.1 in November 1990. (17) Further development of the TEI Guidelines was done by four Working Committees (Text Documentation, Text Representation, Text Analysis and Interpretation, Metalanguage and Syntax) and a number of specialist Working Groups amongst which groups on character sets, text criticism, hypertext and hypermedia, formulæ, tables, figures, and graphics, language corpora, manuscripts and codicology, verse, drama and performance texts, literary prose, linguistic description, spoken text, literary studies, historical studies, print dictionaries, machine lexica, and terminological data. The extensions and revisions resulting from this work, together with extensive public comment resulted in the drafting of a revised version, TEI P2, that was released chapter by chapter between March 1992 and the end of 1993(18) and that included substantial amounts of new material.

TEI P3

The following step was the publication of the TEI P3 *Guidelines for Electronic Text Encoding and Interchange* in 1994 (19) that presented a further revision of all chapters published under the document number TEI P2, and the addition of further chapters. A final revised edition of this P3 Guidelines correcting several typographic and other errors, and introducing one new element was published in 1999 (20). The publication of this 1,292 page documentation of the definitive guidelines defining some 439 elements marked the conclusion of the initial development work. With this work, the Poughkeepsie Guidelines were met by providing a framework for the encoding of texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content and treating both continuous materials ("running text") and discontinuous materials such as dictionaries and linguistic corpora.

TEI P4

Recognising the benefits for the TEI community, the P4 revision of the TEI Guidelines (21) was

published in 2002 by the newly formed TEI Consortium in order to provide equal support for XML and SGML applications using the TEI scheme. The chief objective of this revision was to implement proper XML support in the Guidelines, while ensuring that documents produced to earlier TEI specifications remained usable with the new version. The XML support was realized by the expression of the TEI Guidelines in XML and the conformation to a TEI conformant XML DTD. However, the TEI P4 generated a set of DTD fragments that can be combined together to form either SGML or XML DTDs and thus achieved backwards compatibility with TEI P3 encoded texts. In other words, any document conforming to the TEI P3 SGML DTD was guaranteed to conform to the TEI P4 XML version of it. This ‘double awareness’ of the TEI P4 is the reason why this version was called an ‘XML-compatible edition’ rather than an ‘XML edition’. This was achieved by restricting the revisions needed to make the P4 version with its 441 elements to error correction only. During this process of revision, however, many possibilities for other, more fundamental changes have been identified. Which led to the current TEI P5 version of the Guidelines.

TEI P5

In 2003 the TEI-Consortium asked their membership to convene Special Interest Groups (SIGs) whose aim could be to advise revision of certain chapters of the Guidelines and suggest changes and improvements in view of the P5. With the establishment of the new TEI Council, which superintends the technical work of the TEI Consortium, it became possible to agree on an agenda to enhance and modify the Guidelines more fundamentally which resulted in a full revision of the Guidelines published as TEI P5.(13) TEI P5 contains a full XML expression of the TEI Guidelines and introduces new elements, revises content models, and reorganises elements in a modular class system that facilitates flexible adaptations to users' needs. Contrary to its predecessor, TEI P5 does not offer backwards compatibility with previous versions of the TEI. The TEI Consortium will,

however, maintain and error correct the P4 Guidelines. This means that users still have the option between P4 and P5.

TEI: ORGANISATION

The TEI Consortium was established in 2000 as a not-for-profit membership organisation to sustain and develop the Text Encoding Initiative (TEI). The Consortium has hosts at Brown University, Oxford University, the University of Virginia, and institutes in Nancy. The Consortium is managed by a Board of Directors, and its technical work is overseen by an elected technical Council who take responsibility over the content of the TEI Guidelines.

The TEI charter outlines the consortium's goals and fundamental principles. Its goals are:

1. To establish and maintain a home for the Text Encoding Initiative (TEI) in the form of a permanent organizational structure.
2. To ensure the continued funding of TEI-C activities, for example: editorial maintenance and development of the TEI guidelines and DTD, training and outreach activities, and services to members.
3. To create and maintain a governance structure for the TEI-C with broad representation of TEI user-communities.

The Consortium honours four fundamental principles:

1. The TEI guidelines, other documentation, and DTD should be free to users;
2. Participation in TEI-C activities should be open (even to non-members) at all levels;
3. The TEI-C should be internationally and interdisciplinarily representative;
4. No role with respect to the TEI-C should be without term.

Involvement in the consortium is possible in three categories: voting membership which is open to individuals, institutions, or projects; non-voting subscription, which is open to personal individuals only, and sponsorship, which is open to individual or corporate sponsors. Only members have the

right to vote on consortium issues and in elections to the Board and the Council, have access to a restricted website with pre-release drafts of Consortium working documents and technical reports, announcements and news, and a database of members, Sponsors, and Subscribers, with contact information, and benefit from discounts on training, consulting, and certification. The Consortium members meet annually at a Members' Meeting where current critical issues in text encoding are discussed, and members of the Council and members of the Board of Directors are elected. The membership fee payable varies depending on the kind of project or institution and its location depending on where the economy of the member's country falls in the four-part listing of Low, Lower-Middle, Middle-Upper, and High Income Economies, as defined by the World Bank.

CONCLUSIONS

Computers can only deal with explicit data. The function of markup is to represent textual material into digital form through the explicating act of text-encoding. Descriptive markup reveals what the encoder thinks to be implicit or hidden aspects of a text, and is thus an interpretive medium which often documents scholarly research next to structural information about the text. In order for this research to be exchangeable, analysable, re-usable, and preservable, texts in the field of the humanities should be encoded according to a standard which defines a common vocabulary, grammar, and syntax, whilst leaving the implementation of the standard up to the encoder. A result of communal efforts among computing humanists, the Text Encoding Initiative documents such a standard in the TEI Guidelines. These guidelines are fully adaptable and customizable to one's specific project whilst enhancing this project's compatibility with other projects employing the TEI. Since over two decades, the TEI has been used extensively in projects from different disciplines, fields, and subjects internationally. The ongoing engagements of a broad user community through the organisation of the TEI Consortium consolidates the importance of the text encoding standard and informs its continuous development and maintenance.

FURTHER READING

Burnard, Lou; O'Brien O'Keeffe, Katherine; Unsworth, John (eds.) *Electronic Textual Editing.* MLA: New York, 2006. Preview: <http://www.tei-c.org/About/Archive_new/ETE/> (Accessed October 2008)

Cummings, James. The Text Encoding Initiative and the Study of Literature. In *A Companion to Digital Literary Studies;* Siemens, Ray, Schreibman, Susan, Eds.; Blackwell Publishing: Malden, MA, Oxford, 2007; 451-476. <http://www.digitalhumanities.org/companionDLS/> (Accessed October 2008)

Ide, Nancy; Véronis, Jean (eds.). *Text Encoding Initiative: Background and Context.* Kluwer Academic Publishers: Dordrecht, 1995. Reprinted from Computers and the Humanities **1995,** *29*.

Mylonas, Elli; Renear, Allen (eds.). Special Issue: Selected Papers from TEI 10: Celebrating the Tenth Anniversary of the Text Encoding Initiative. Computers and the Humanities **1999**, *33* (1-2).

Schreibman, Susan; Rahts, Sebastian (eds.). Special Issue: TEI at 20. LLC. The Journal of Digital Scholarship in the Humanities **2009**, *24*.

Van den Branden, Ron; Vanhoutte Edward; Terras, Melissa. *TEI By Example*. <http://www.teibyexample.org> (Accessed October 2008)

REFERENCES

(1) Sperberg-McQueen, C.M.. Text in the Electronic Age: Textual Study and Text Encoding with examples from Medieval Texts. Literary and Linguistic Computing **1991**, *6* (1): 34-46.

(2) Kay, M. Standards for Encoding Data in a Natural Language. Computers and the Humanities **1967**, *1* (5): 170-177.

(3) Russel, D.B. *COCOA: A Word Count and Concordance Generator for Atlas.* Atlas Computer Laboratory: Chilton, 1967.

(4) Hockey, S. *Oxford Concordance Program Users' Manual.* Oxford University Computing Service: Oxford, 1980.

(5) Lancashire, I.; Bradley, J.; McCarty, W.; Stairs, M.; Woolridge, T.R. *Using TACT with Electronic Texts.* Modern Language Association of America: New York, 1996.

(6) Berkowitz, L.; Squiter, K. A. *Thesaurus Linguae Graecae, Canon of Greek Authors and Works.* Oxford University Press: New York/Oxford, 1986.

(7) Goldfarb, C.E. *The SGML Handbook.* Clarendon Press: Oxford, 1990.

(8) Barnard, D.T.; Fraser, C.A.; Logan, G.M.. Generalized Markup for Literary Texts. Literary and Linguistic Computing **1988,** *3* (1): 26-31.

(9) Barnard, D.T., R. Hayter, M. Karababa, G. Logan, and J. McFadden (1988b). SGML-Based Markup for Literary Texts: Two Problems and Some Solutions. Computers and the Humanities **1988**, *22* (4): 265-276.

(10) Bray, Tim; Paoli, Jean; Sperberg-McQueen, C.M. *Extensible Markup Language (XML) 1.0.*W3C Recommendation 10-February-1998. http://www.w3.org/TR/1998/REC-xml-19980210 (accessed September 2008)

(11) DeRose, Steven J. XML and the TEI. *Computers and the Humanities* **1999,** *33* (1-2): 11-30.

(12) Burnard, L.; Sperberg-McQueen. C.M. TEI Lite: Encoding for Interchange: an introduction to the TEI Revised for TEI P5 release. February 2006 <http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html>

(13) TEI Consortium (eds.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative Consortium. <http://www.tei-c.org/Guidelines/P5/ > (accessed October 2008)

(14) Burnard, L. Report of Workshop on Text Encoding Guidelines. Literary and Linguistic Computing **1988**, *3* (2): 131-133.

(15) Ide, N.M.; Sperberg-McQueen, C.M. Development of a Standard for Encoding Literary and Linguistic Materials. In *Cologne Computer Conference 1988. Uses of the Computer in the Humanities and Social Sciences.* Volume of Abstracts. Cologne, Germany, Sept 7-10 1988, p. E.6-

3-4.

(16) Ide, N.; Sperberg-McQueen, C.M. The TEI: History, Goals, and Future. Computers and the Humanities **1995**, *29* (1): 5-15.

(17) Sperberg-McQueen, M.; Burnard, L. (eds.). *TEI P1: Guidelines for the Encoding and Interchange of Machine Readable Texts*. ACH-ALLC-ACL Text Encoding Initiative: Chicago/Oxford, 1990. Available from http://www.tei-c.org.uk/Vault/Vault-GL.html (accessed October 2008)

(18) Sperberg-McQueen, M.; Burnard, L. (eds.). *TEI P2 Guidelines for the Encoding and Interchange of Machine Readable Texts Draft P2* (published serially 1992-1993); Draft Version 2 of April 1993: 19 chapters. Available from http://www.tei-c.org.uk/Vault/Vault-GL.html (accessed October 2008)

(19) Sperberg-McQueen, C.M.; Burnard, L. (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange. TEI P3.* Text Encoding Initiative: Oxford, Providence, Charlottesville, Bergen, 1994.

(20) Sperberg-McQueen, C.M.; Burnard L. (eds.). *Guidelines for Electronic Text Encoding and Interchange. TEI P3. Revised reprint.* Text Encoding Initiative: Oxford, Providence, Charlottesville, Bergen, 1999.

(21) Sperberg-McQueen, C.M.; Burnard, L. (eds.). *TEI P4: Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition.* XML conversion by Syd Bauman, Lou Burnard, Steven DeRose, and Sebastian Rahtz Text Encoding Initiative Consortium: Oxford, Providence, Charlottesville, Bergen, 2002. http://www.tei-c.org.uk/P4X/ (accessed October 2008)