TEI By Example

Edward Vanhoutte
Ron Van den Branden
Melissa Terras

Centre for Scholarly Editing and Document Studies (CTB), Royal Academy of Dutch Language and Literature, Belgium, Gent, 9 July 2010

Last modified on: 2018-04-28
Licensed under a Creative Commons Attribution ShareAlike 3.0 License
TABLE OF CONTENTS

3. Markup Languages in the Humanities
   3.1 Procedural and Descriptive Markup
   3.2 Early Attempts
   3.3 The Standard Generalized Markup Language (SGML)
   3.4 The eXtensible Markup Language (XML)
3. Markup Languages in the Humanities

3.1 Procedural and Descriptive Markup

When human beings read texts, they perceive both the information stored in the linguistic code of the text and the meta-information which is inferred from the appearance and interpretation of the text. By convention, italics are, for instance, used as a code signalling a title of a book, play, or movie; a foreign word or phrase; or emphatic use of the language. Through their cognitive abilities, readers usually have no problems selecting the most appropriate interpretation of an italic string of text. Computers, however, need to be informed about these issues in order to be able to process them. This can be done by way of a markup language that provides rules to formally separate information (the text in a document) from meta-information (information about the text in a document). 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 (e.g. printing a string of text in italics), the humanities were also and mainly considered with descriptive markup that identifies the entity type of tokens (e.g. identifying that a string of text is a title of a book or a foreign word). Unlike procedural or presentational markup, descriptive markup establishes a one to one mapping between logical elements in the text and their markup. In order to achieve this, descriptive markup languages tend to formally separate information (the text in a document) from meta-information (information about the text in a document).

3.2 Early Attempts

Some sort of standardization of markup 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 \(^1\), but used as an input standard by the Oxford Concordance Program (OCP) in the 1980s \(^2\) and by the Textual Analysis Computing Tools (TACT) in the 1990s \(^3\). For the transcription and encoding of classical Greek texts, the Beta-transcription/encoding system reached some level of standardized use \(^4\).

3.3 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)\(^5\) 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.\(^6\)

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 enables 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.\(^7\)\(^8\) From the beginning, however, SGML was also criticized for at least two

problematic matters: SGML’s hierarchical perspective on text, i.e. the representation of text as a hierarchical tree structure, and SGML’s verbose markup system. These two issues have since been central to the theoretical and educational debates on markup languages in the humanities.

### 3.4 The eXtensible Markup Language (XML)

The publication of the eXtensible Markup Language (XML) 1.0 as a W3C recommendation in 1998 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.

Because of its advantages and widespread popularity, XML became the metalanguage of choice for expressing the rules for descriptive text encoding in TEI.

---

