**CUMMINGS — TEI Simple**

<1 figure>

**TEI Simple: Power, Economy, and a Processing Model for Encoders and Developers**

Cummings, J., Rahtz, S., Pytlik Zillig, B., Mueller, M. and Turska, M.

TEI Simpleis a Mellon-funded project developing a customisation and extension of the Guidelines of the Text Encoding Initiative. Its intent is to create a highly constrained and prescriptive subset of the TEI suitable for a straightforward representation of standard early modern and modern books and a processing model documentation for easy web presentation. The project will also develop a mapping to non-TEI ontologies, and a method of indicating the status and richness of a digital text. The outputs from TEI Simple will be integrated into the TEI infrastructure and supported by the TEI Technical Council. The project runs from September 2014 to July 2015. As DH2015 is scheduled near the end of this project, it is an ideal location to both announce the project to the wider DH community while simultaneously reporting on the status of the outputs.

The project is directed by Sebastian Rahtz (University of Oxford), Brian Pytlik Zillig (University of Nebraska-Lincoln), and Martin Mueller (Northwestern University). Other project staff include Magdalena Turska (DiXiT Project / University of Oxford), Lou Burnard (consultant), and James Cummings (University of Oxford). The Advisory Committee consists of Pip Willcox (Bodleian Library, Oxford), Suzanne Haaf (Deutsches Textarchiv, Berlin), Matthias Goebel (University of Gottingen), and James Cummings (University of Oxford).

All of the work for the project is undertaken openly through a public github repository: https://github.com/TEIC/TEISimple.

**TEI Simple Objectives**

TEI Simple has the following high-level objectives:

1. Definition of a new *highly constrained* and *prescriptive* subset of the TEI Guidelines suited to the representation of early modern and modern books. The degree of detail supported will be sufficient to encompass, at a minimum, the current practices of the TCP’s EEBO, ECCO, and Evans collections plus those of other major European initiatives such as Text Grid or the German Text Archive (DTA) and the Consortium Cahiers in France.

2. Developing and implementing processing rules for TEI Simple and creating a notation (as an extension to TEI’s ODD metalanguage1) for specifying such rules, referencing web standards such as XPath, CSS, and XSL FO.

3. Formal mapping of the elements used by TEI Simple to the Conceptual Reference Model of the International Council of Museums (CIDOC CRM), allowing for full interoperability with the Europeana Data Model, in order to facilitate the participation of projects in the Europeana repositories.

4. Definition and implementation of machine-readable descriptions of the encoding status and richness of TEI texts, providing ‘TEI Performance Indicators’, which help to document expectations for the encoded text.

5. Full integration of TEI Simple into the TEI Guidelines and infrastructure with ongoing maintenance by the TEI Technical Council.

More information concerning each of these objectives is detailed below.

**TEI Simple Subset**

This subset of TEI Simple attempts to remove ambiguity for both encoders and developers. While simple does not necessarily mean ‘small’, nor does it mean ‘simplistic’, the decision was made to base the selection of elements on TEI texts from a set of large archives or text collections. This included

• Text Creation Partnership: Evans, ECCO, EEBO (including the unreleased phase 2).

• Oxford Text Archive: All TEI P5 files.

• Deutsches Textarchiv.

• Documenting the American South.

• CESR.

• OBVIL: corpus critique.

The decision was also made to concentrate primarily on the <text> element, not the metadata stored in the <teiHeader>. The constraining of TEI elements, and limiting of encoding options, meant that the 546 (as of TEI P5 version 2.7.0) elements could be limited to 104 (at time of writing) not including <teiHeader> elements. It is anticipated that the current list of elements may change over the course of development of TEI Simple, but the elements selected were noticed to generally fall into a particular set of categories:

castlist <actor>, <castGroup>, <castItem>, <castList>, <role>, <roleDesc>

character <g>

editorial <abbr>, <add>, <addSpan>, <am>, <choice>, <corr>, <del>, <desc>, <ex>, <expan>, <gap>, <handShift>, <orig>, <reg>, <sic>, <space>, <subst>, <supplied>, <unclear>

interpretation <author>, <date>, <foreign>, <hi>, <measure>, <name>, <num>, <q>, <quote>, <ref>, <rs>, <seg>, <time>

linguistic <c>, <pc>, <s>, <w>

pictures <figDesc>, <figure>, <graphic>

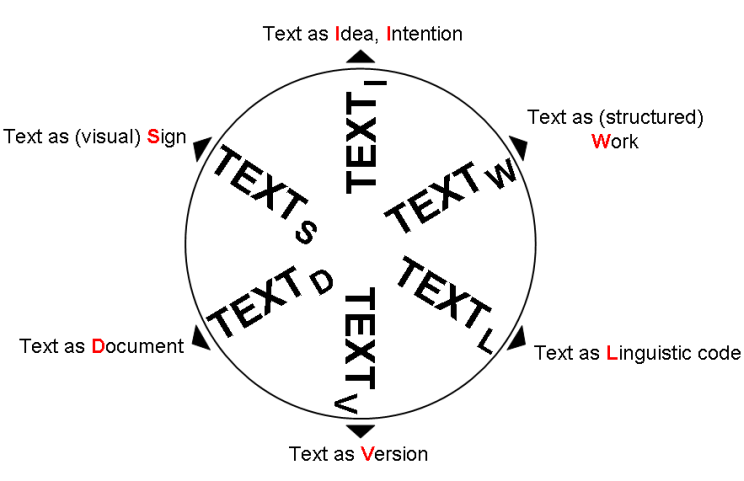
structure <ab>, <address>, <addrLine>, <anchor>, <back>, <body>, <bibl>, <cb>, <cit>, <div>, <floatingText>, <formula>, <front>, <fw>, <group>, <head>, <item>, <l>, <label>, <lb>, <lg>, <list>, <listBibl>, <milestone>, <note>, <p>, <pb>, <sp>, <speaker>, <spGrp>, <stage>, <text> , <TEI>, <teiCorpus>, <title>

table <cell>, <row>, <table>

titlepage <publisher>, <pubPlace>, <docAuthor>, <docDate>, <docEdition>, <docImprint>, <docTitle>, <imprimatur>, <titlePage>, <titlePart>

wrapper <argument>, <byline>, <closer>, <dateline>, <epigraph>, <opener>, <postscript>, <salute>, <signed>, <trailer>

We aim to link this simple taxonomy to the distinctions made in Patrick Sahle’s ‘Text Wheel’, and investigate the relationship of this to default behaviour in the Processing Model.

****

In addition, TEI Simple has customised the available attribute values for a number of attributes. It has removed the @rend and @style attributes, preferring to use @rendition to document the appearance of the original object. In this attribute it uses a closed list of values that its implementations know about in the form of a private URI of ‘simple’: (e.g., ‘simple:bold’).

**TEI Simple Processing Model**

A ‘cradle to grave’ processing model is at the heart of the TEI Simple project. The TEI Simple processing model offers a bridge across the divide between encoders and developers: the aim is to lower the access barriers to working with TEI-encoded texts in various web environments.

The TEI Simple project has developed a notation by which a TEI profile records the intended processing for documents meeting that profile. The TEI Simple Processing Model notation provides a way to document the intended output rending in the TEI customisation profile (TEI ODD file). This is done by means of a proposed <model> element for use in the context of TEI ODD at a fairly high level and in an abstract manner. The agreement on notation here, though, enables the generation of document-specific transformations. While the implementation of TEI Simple uses an open function library as a method of documenting the further processing in the case of TEI Simple documents, this same processing model notation will be incorporated into the TEI infrastructure where it will be of benefit for those wishing to use a high-level form of output specification, and who wish to develop their own function library.

**TEI Simple Mapping to RDF**

Although simple presentation in web pages is an important aim of TEI Simple, it is also important to represent structural and semantic markup in the open data interchange format of RDF. The Europeana Data Model (EDM) and the Conceptual Reference Model of the International Council of Museums (CIDOC-CRM) are parts of an evolving ecosystem of metadata standards for cultural heritage documentation across the many languages and cultures of Europe. TEI ODD already has a notation (the <equiv> element) for expressing the relationship between TEI and RDF, and this has been used to map the elements from TEI Simple to CIDOC-CRM and FRBRoo.

**TEI Simple Performance Indicators**

Although TEI Simple is designed to be very constrained, the decisions as to what to mark up are still left to the encoder. Do they choose, for example, to explicitly identify names of people and places? Will they mark where spelling has been normalized? Will all the words be marked with part of speech information for linguistic analysis? For a simple example, are <name> elements not present in a text because they have not been encoded, or because there are no names in this text? This will affect the *query potential* of a corpus of texts, but cannot be determined simply by analyzing the markup. The TEI Simple project has created an extra level of metadata notation for enabling the automatic profiling of a text.

**TEI Simple Integration with TEI Infrastructure**

The outputs of the TEI Simple project are being fully integrated with the existing TEI infrastructure and thus are available for all TEI users whether using TEI Simple or not. The acceptance of TEI Simple by the TEI Technical Council is one of the success criteria for the funding received. As part of the integration with TEI Infrastructure the TEI Simple project has created a teitosimple XSLT conversion that checks texts conformance to TEI Simple, converts elements where possible, and maps attributes such @rend to known @rendition values.

**Further Development of TEI Simple**

This paper will conclude with a look at resources built on top of TEI Simple and what work might be developed from it. The TEI Simple project limited its scope to early modern and modern printed books. Under the aegis of the *DiXiT* project (an EU Marie Curie ITN) James Cummings and Magdalena Turska have been investigating TEI Simple as a starting point for the creation of scholarly digital editions of simple (but potentially multi-witness) manuscript materials.

**Note**

1. ODD is an acronym for ‘one document does it all’. The TEI ODD is an example of ‘literate programming’ and combines code and discursive documentation in a single document.

**References**

**Cummings, James.** (2008). The Text Encoding Initiative and the Study of Literature. In Schreibman, S. and Siemens, R. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 451–76, http://www.digitalhumanities.org/companionDLS/.

**Cummings, J. and Willcox, P.** (2013). Stationers’ Register Online: A Case Study of a Byte-Reduced TEI Schema for Digitization (tei\_corset). *Journal of the Text Encoding Initiative,* December, http://jtei.revues.org/926.

**Fischer, F.** (2008). The Pluralistic Approach—The First Scholarly Edition of William of Auxerre’s Treatise on Liturgy. *Jahrbuch für Computerphilologie,* http://computerphilologie.tudarmstadt. de/jg08/fischer.html.

**Pytlik Zillig, B. L.** (2013). Logging the Abbot: Reflection-Oriented XSLT Programming for Corpora Conversion and Verification. *Journal of the Text Encoding Initiative,* March, http://jtei.revues.org/722.

**Pytlik Zillig, B. L.** (n.d.). TEI Texts that Play Nicely: Lessons from the MONK Project. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science,* **1**(3), https://letterpress.uchicago.edu/index.php/jdhcs/article/view/81.

**Rahtz, S. and Cummings, J.** (2012). Kicking and Screaming: Challenges and Advantages of Bringing TCP Texts into Line with the Text Encoding Initiative. In Bodleian Libraries, University of Oxford, *‘Revolutionizing Early Modern Studies’? The Early English Books Online Text Creation Partnership in 2012: EEBOTCP 2012 Conference Proceedings,* http://ora.ox.ac.uk/objects/uuid%3Af9667884220b4ec9bb2fc79044302399.

**Sahle, P.** (2013). Band 9: Textbegriffe und Recodierung. In *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, 3 Bände, Norderstedt: Books on Demand, Schriften des Instituts für Dokumentologie und Editorik 7–9, http://www.ide. de/schriften/s79digitaleeditionsformen.