*TEI Simple: towards an amenable TEI*

Sebastian Rahtz, Martin Mueller, Brian Pytlik Zillig

# Summary

TEI Simple aims to define a new *highly-constrained* and *prescriptive* subset of the Text Encoding Initiative (TEI) Guidelines suited to the representation of early modern and modern books, a formally-defined set of processing rules which permit modern web applications to easily present and analyze the encoded texts, mapping to other ontologies, and processes to describe the encoding status and richness of a TEI digital text.

# Background

The Text Encoding Initiative (TEI) has developed over 20 years into a key technology in text-centric humanities disciplines, with an extremely wide range of applications, from diplomatic editions to dictionaries, from prosopography to speech transcription and linguistic analysis. It has been able to achieve its range of use by adopting a *descriptive* rather than *prescriptive*  approach , by recommending *customization* to suit particular projects, and by eschewing any attempt to dictate how the digital texts should be rendered or exchanged. However, this flexibility has come at the cost of relatively limited success in interoperability. In our view there is a distinct set of uses (primarily in the area of digitized ‘European’-style books) that would benefit from a *prescriptive* recipe for digital text; this will sit alongside other domain-specific, constrained TEI customizations, such as the very successful *Epidoc* in the epigraphic community. TEI-Simple may become a prototype for a new family of constrained customizations. For instance, a TEI Simple MS for manuscript based work could be built on top of the ENRICH project, drawing on many of the lessons and some of the code for TEI Simple.

The TEI has long maintained an introductory subset (TEI Lite), and a constrained customization for use in outsourcing production to commercial vendors (TEI Tite), but both of these permit enormous variation, and have nothing to say about processing. The present project can be viewed in some ways as a revision of TEI Lite, re-examining the basis of the choices therein, focusing it for a more specific area, and adding a "cradle to grave" processing model that associates the TEI Simple schema with explicit and standardized options for displaying and querying texts. This means being able to specify what a programmer should do with particular TEI elements when they are encountered, allowing programmers to build stylesheets that work for everybody and to query a corpus of documents reliably.

This proposal, TEI Simple, will focus on interoperability, machine generation, and low-cost integration. The TEI architecture facilitates customizations of many kinds; TEI Simple aims to produce a complete 'out of the box' customization which meets the needs of the many users for whom the task of creating a customization is daunting or seems irrelevant. TEI Simple in no way intends to constrain the expressive liberty of encoders who do not think that it is either possible or desirable to follow this path. It does, however, promise to make life easier for those who think there is some virtue in travelling that path as far as it will take you, which for quite a few projects will be far enough. Some users will never feel the need to move beyond it, others will outgrow it, and when they do they will have learned enough to do so.

‘Comparability and interoperability with other resources’ are an increasingly important topic on various Digital Humanities agendas. Echoes of it are found in a recent ‘work set construction’ Mellon grant to the Hathi Trust Research Centre. Under the heading ‘Wissenschaftliche Sammlungen’ it is a major part of an ambitious DARIAH project anchored at the SUB Göttingen. Progress towards it may be slow, tedious, and partial, but ‘simplicity, interoperability, broad use and reuse’, and ‘comparability and interoperability with other resources’ are important goals to keep in mind for many purposes. For a lot of current and future users of the TEI the really important benefits come from the simple stuff, and beyond some level of complexity they begin to feel some sympathy with Andrew Prescott's not very kind phrase about ‘angels dancing on angle brackets.’[[1]](#footnote-2)

A major driver for this project is the texts created by phase 1 of the EEBO-TCP project, which will be placed in the public domain on 1 January 2015. Another 45,000 texts will join over the following five years, creating by 2020 an archive of 70,000 consistently encoded books published in England from 1475 to 1700, including works of literature, philosophy, politics, religion, geography, science and all other areas of human endeavor. When we compare the query potential of the EEBO TCP texts in their current and quite simple encoding with flat file versions of those text, it is clear that the difference in query potential is very high, especially if you add to that coarse encoding simple forms of linguistic annotation or named entity tagging that can be added in a largely algorithmic fashion. During 2012 and 2013 extensive work has been undertaken at Northwestern, Michigan and Oxford to enrich these texts and bring them into line with the current TEI Guidelines (where necessary working with the TEI to modify the Guidelines). TEI Simple will use this corpus as a point of departure and will provide its users with a friendlier environment for manipulating EEBO texts in various projects. But TEI Simple should not be understood as an EEBO specific project. We believe that, given the extraordinary degree of internal diversity in the EEBO source files, a project that starts from them can, with appropriate modifications, accommodate a wide range of printed texts differing in language, genre, or time and place of origin.

# Objectives

TEI Simple has the following high-level objectives:

1. Definition of a new *highly constrained* and *prescriptive* subset of the Text Encoding Initiative (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 DTA in Germany, and the Consortium Cahier in France.
2. Creation of a notation (as an extension to TEI's ODD metalanguage) for specifying processing rules for TEI encoded texts,[[2]](#footnote-3) referencing web standards such as XPath, CSS and XSL FO.
3. Reference implementations of processing rules defined for this TEI subset.
4. Formal mapping of the TEI elements used by Simple to the CIDOC CRM, allowing for full interoperability with the Europeana Data Model, in order to facilitate the participation of projects in the Europeana repositories.
5. Definition and implementation of machine-readable descriptions of the encoding status and richness of TEI texts, providing a “TEI Performance Indicators” indicating to a user what they can expect to use the text for.

The aim is to lower the access barriers to working with TEI-encoded texts in various web environments. Programmers familiar with a particular web environment, whether Django, Drupal, eXist, Ruby on Rails, or others will be able to integrate TEI Simple-based projects into their environment with moderate effort and with no more than their usual tools and skills.

The project will adhere to the following principles:

* As little overlap as possible, and as much compatibility as possible, with existing repository projects
* At least as prescriptive as level 3 of the *Best Practices for TEI in Libraries*
* Encompassing I18N principles at all times
* Useable implementations of all features

Outcomes from TEI Simple, consisting of a documented definition in ODD of the TEI subset, a set of processing rules, and extensions to the TEI ODD language to record processing expectations, will be fully integrated into the TEI infrastructure with ongoing maintenance by the TEI Technical Council.

TEI Simple is intended to be *complementary* to community projects like the TAPAS project, and to the established work of TextGrid, the German Text Archive (the DTA ‘base format’, which shares many of the goals of TEI Simple) and other national projects.

# The processing model

The current plan for a processing model for Simple is that it will define three stages:

1. In the first stage, every TEI Simple element is assigned to a category, according to a taxonomy whose first draft is below. This allows a processor to know whether to handle the element or not, and broadly speaking how to display or otherwise process it. Note how XPath may be used used to refine a name.
2. In the second stage, the categories are mapped to a presentation format, using HTML and CSS concepts where possible.
3. In the third stage, a normalized set of property values is created for each element, where the combination of the *rend*, *rendition* and *style* attributes, and the <rendition> element, are interpreted to map to the names and allowed values for CSS. This all allows a processor to work out how to present that element in the current context.

|  |  |  |
| --- | --- | --- |
| **Category** | **Meaning** | **Example** |
| 1 | metadata header | fileDesc |
| 2 | section heading | head[parent::div] |
| 3 | title of object (figure, table etc) | head[not(parent::div)] |
| 4 | structural division | div |
| 5 | uncategorized block level object | quotation |
| 6 | semantic block level object | person |
| 7 | uncategorized inline object | hi |
| 8 | semantic inline object | persName |
| 9 | list | list |
| 10 | list item | item |
| 11 | table | table |
| 12 | cell | cell |
| 13 | row | row |
| 14 | out of line note |  |
| 15 | figure | figure |
| 16 | pointer | ptr |
| 17 | Janus element (alternate children) | choice |
| 18 | modern commentary element | desc |

1. A preliminary processing categorisation

Processors can now create a variety of outputs without having to maintain specific rule sets for TEI Simple. The rules for stage 1 and stage 2 are maintained as part of the ODD schema for Simple, so a processor will read both the source document and the corresponding ODD file (or some compiled version of if), and have access to all the information it needs.

The processing model assertions will be modelled in ODD using a notation similar to Schematron constraints, allowing for multiple categories for a given element, depending on an XPath filter.

There are two important differences between this setup and the "TEI to XXX" stylesheets (eg <https://github.com/TEIC/Stylesheets>) which have been developed many times:

1. The initial explicit assertion of a category gives a place for the designer of Simple to commit to a style of output, separately from the details of the rendering
2. The rules are tied to the schema, allowing the actual (eg) TEI to LaTeX or TEI to Word transformations to be completely agnostic.

We may contrast this with an example from the TEI Stylesheets, where an XSL function which determines whether to render something as bold mixes up the context with the attributes, and is very hard to tailor for a constrained subset like Simple:

<**xsl:function** **name**="tei:render-bold" **as**="xs:boolean"  
   xmlns:xsl="http://www.w3.org/1999/XSL/Transform">  
 <**xsl:param** **name**="element"/>  
 <**xsl:for-each** **select**="$element">  
  <**xsl:choose**>  
   <**xsl:when** **test**="@rend='odd\_label'">true</**xsl:when**>  
   <**xsl:when**  
     **test**="parent::tei:hi[starts-with(@rend,'specList-')]">true</**xsl:when**>  
   <**xsl:when** **test**="self::tei:docAuthor">true</**xsl:when**>  
   <**xsl:when**  
     **test**="self::tei:label[following-sibling::tei:item]">true</**xsl:when**>  
   <**xsl:when** **test**="starts-with(@rend,'specList-')">true</**xsl:when**>  
   <**xsl:when**  
     **test**="starts-with(parent::tei:hi/@rend,'specList-')">true</**xsl:when**>  
   <**xsl:when** **test**="@rend='label'">true</**xsl:when**>  
   <**xsl:when** **test**="@rend='wovenodd'">true</**xsl:when**>  
   <**xsl:when** **test**="@rend='important'">true</**xsl:when**>  
   <**xsl:when** **test**="@rend='specChildModule'">true</**xsl:when**>  
   <**xsl:when**  
     **test**="ancestor-or-self::tei:cell[@rend='wovenodd-col1']">true</**xsl:when**>  
   <**xsl:when**  
     **test**="ancestor-or-self::tei:cell[@role='label']">true</**xsl:when**>  
   <**xsl:when**  
     **test**="ancestor-or-self::\*[@rend][contains(@rend,'bold')]">true</**xsl:when**>  
   <**xsl:when**  
     **test**="parent::tei:hi[starts-with(@rend,'specList-')]">true</**xsl:when**>  
   <**xsl:when**  
     **test**="self::tei:cell and parent::tei:row[@role='label']">true</**xsl:when**>  
   <**xsl:when**  
     **test**="self::tei:label[following-sibling::tei:item]">true</**xsl:when**>  
   <**xsl:otherwise**>false</**xsl:otherwise**>  
  </**xsl:choose**>  
 </**xsl:for-each**>  
</**xsl:function**>

# Programme of work and budget

We propose a total budget of $100,000, divided into labour costs ($36,000), travel and meeting costs ($59,000), and a contingency fund ($5000).

The work will be undertaken in three ways:

1. Contributed time and expertise by members of the advisory panel, and other members of the TEI community, in consultation events.
2. Technical development sprints during which the core programming tasks will be completed, in some cases as a group activity at the same location.
3. Outputs commissioned at an agreed cost from partner organisations or individuals, under supervision of one of the PIs.

The aim is to ensure that payments from the project are tied to explicit deliverables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Task | Activity | Method | Participants (leader in bold) | Date |
| 1 | Management | Overseeing the progress of the project | Two face to face meetings of the PIs and advisory group; monthly video-conference management meetings | **Rahtz**, Mueller, Pytlig Zillig |
| 2 | Simple subset of TEI | Definition of the prescriptive subset of the TEI, expressed as an ODD customization | Workshop sprint (3 days) | **Mueller**, Burnard, Cummings, Geyken, Glorieux, Burnard, Pytlig Zillig, Rahtz |
| 3 | Model notation | Extension to the TEI ODD language to support processing expectations | Workshop (3 days) | **Rahtz**, Cummings, Glorieux, Jolivet, Turska |
| 4 | Model for Simple | Definition of processing expectations for TEI Simple, using the model notation | Workshop sprint | **Rahtz**, Cummings, Jolivet, Pytlig Zillig, Turska |
| 5 | Model Implementation | Implementation of processing expectations, as a rewrite of the TEI XSLT stylesheets | Workshop sprint (5 days) | **Pytlig Zillig**, Rahtz, Turska |
| 6 | Documentation | Maintainable open access documentation for TEI Simple | Continuous refinement | **Mueller**, contractor TBD |
| 7 | Mapping to RDF | Mapping TEI Simple to other ontologies, eg CIDOC CRM, expressing the mapping using the TEI ODD language | Continuous refinement | **Rahtz**, contractor TBD |
| 8 | Profile markup | Development of a markup system for recording the formal profile of a text | Workshop sprint (3 days) | **Pytlig Zillig**, Geyken, Glorieux, Mueller, Turska |
| 9 | Profile implementation | Implementation of profile reporting | Continuous refinement | **Pytlig Zillig**, Geyken, contractor TBD |
| 10 | Integration into TEI | Integration of TEI Simple into the Guidelines and other TEI Consortium infrastructure | **Cummings**, Burnard, Rahtz |
| 11 | User validation | Explanation and demonstration of TEI Simple to check that user requirements are being met | Open workshop at midpoint of project (2 days, local expenses only) |

1. Project tasks

The following people will work on TEI Simple over the course of the project lifetime:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Institution** | **Project role** | **Expertise** |
| Sebastian Rahtz | University of Oxford | PI | TEI processing, ODD language |
| Martin Mueller | University of Nebrasks | PI | Re-useable digital text, analysis |
| Brian Pytlik Zillig | Northwestern University | PI | Normalizing text corpora |
| Lou Burnard | independent consultant | contractor | TEI design and semantics |
| James Cummings | University of Oxford | contractor | TEI re-use and processing |
| Frédéric Glorieux | Université Paris-Sorbonne | TEI processing |
| Vincent Jolivet | Université Paris-Sorbonne |  | TEI processing |
| Alexander Geyken | Deutsches Textarchiv (DTA) | Normalizing text corpora |
| Magdalena Turska | University of Oxford | Developer | TEI processing |

Costs are calculated on the following basis

1. The time of Pytlig Zillig, Mueller, Geyken, Glorieux and Jolivet will be contributed by their organizations
2. Expert advisors and technical contractors will be paid, to a maximum of $350 a day FTE
3. Travel and living expenses will be paid for all attendees at workshops and development sprints
4. Local expenses only will be paid for the open workshop
5. Commissioned outputs managed by one of the PIs are allocated a number of days technical assistance which they can use to pay individuals or organisations

# Collaboration, management, timing

TEI Simple is a cross-Atlantic collaboration between (at least) the following partners:

1. Northwestern University, Chicago: Professor Martin Mueller
2. The University of Oxford: Sebastian Rahtz
3. The Text Encoding Initiative Consortium: Elena Pierazzo (Chair)
4. The University of Nebraska: Brian Pytlik Zillig
5. Université Paris-Sorbonne: Frédéric Glorieux
6. The Deutsches Textarchiv (DTA): Alexander Geyken
7. TextGrid

TEI Simple will commence work on 1st August 2014, and aim to complete the first stage of work in time for the TEI Annual Meeting in late October 2014. The launch at the Annual Meeting will be followed by eight more months of development and revision before final sign-off on July 1st 2015.

1. Curricula vitae
   1. Sebastian Rahtz

Sebastian Rahtz is Director (Research) of Academic IT in the University of Oxford IT Services, and leads for the department on research support, open source support, research data management, open data, and information strategy. He is a member of IT Services Senior Management team, and represents IT Services on a variety of university committees.

On the research side, Sebastian is technical director for the data store of the CLAROS project (http://www.clarosnet.org/), based in the Oxford e-Research Centre, directs its work on geolocation models and data, and maintains the input of the Lexicon of Greek Personal Names project at Oxford to CLAROS. He assists the Lexicon of Greek Personal Names in its typesetting and sustainability activities. He provides extensive editorial and development work to the Text Encoding Initiative (TEI) Consortium (http://www.tei-c.org/), for which he has served on its Technical Council since 2001.

Sebastian has led a number of successful funding bids in the last decade: the OSS Watch (http://www.oss-watch.ac.uk), open source advisory service, which he directed from 2003-2007; the OXCRI project (2006) to implement XCRI to describe Oxford courses (now being restarted in 2012); and, the Erewhon project (http://erewhon.oucs.ox.ac.uk/) (2008-2010) to investigate mobile access and geolocation of university resources. He was co-leader of Oxford participation in the EU-funded ENRICH project on manuscript encoding (http://enrich.manuscriptorium.com/). Sebastian co-directs and teaches a yearly summer school on text encoding, is a regular speaker at digital humanities and text encoding conference, and teaches TEI workshops.

Sebastian has authored and maintains a wide portfolio of open source software in the area of TEI and TeX typesetting.

* 1. Martin Mueller
  2. Brian Pytlik Zillig
  3. Lou Burnard

Formerly assistant director at Oxford University Computing Services, Lou Burnard is one of the original editors of the TEI Guidelines and has been closely involved with the TEI throughout its evolution. He is an Oxford graduate with a Masters degree in English Language and Literature who has worked at the frontier between computing and the humanities since the 1970s, when he founded the Oxford Text Archive. During the nineties and noughties he was responsible for many other key developments in what we now call the digital humanities, notably the British National Corpus, and the UK's Arts and Humanities Data Service.

* 1. James Cummings
  2. Alexander Geyken
  3. Frédéric Glorieux
  4. Vincent Jolivet
  5. Magdalena Turska

**April 2014** Researcher for the Digital Scholarly Editions IT Services, University of Oxford. Fellow of the Marie Curie DiXiT (Digital Scholarly Editions Initial Training Network) project responsible for requirements and development of a modular publication architecture targeting multiple media

**2007-2014** Laboratory for Editing Sources, Faculty „Artes Liberales”, University of Warsaw. Co-author of “Corpus of Ioannes Dantiscus’ Texts and Correspondence” (dantiscus.al.uw.edu.pl) responsible for TEI/XML encoding policies, design and implementation of software for on-line and printed scholarly editions of renaissance correspondence and other manuscript sources encoded with TEI/XML

**1999 – 2014** Faculty „Artes Liberales”, University of Warsaw, Poland. Lecturer (computer workshops for students & staff) and academic research IT support

**2003 – 2014** InviMed – developer (PHP), Poland. Software system for patients’ medical records for InviMed fertility clinics nationwide

1. In “Consumers, creators or commentators? Problems of audience and mission in the digital humanities “; Arts and Humanities in Higher Education published online 1 December 2011, http://ahh.sagepub.com/content/early/2011/11/30/1474022211428215 [↑](#footnote-ref-2)
2. The paper on “Documenter des “attentes applicatives” (processing expectations)” by Frédéric Glorieux and Vincent Jolivet at TEI Members Meeting 2013 (http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper/ ) also addresses this area. [↑](#footnote-ref-3)