*TEI Simple: towards an amenable TEI*

Sebastian Rahtz, Martin Mueller, Brian Pytlik

Zillig

1. Summary 2

2. Background 2

3. Objectives 4

3.1. The highly constrained and prescriptive element subset of TEI Simple 4

3.2. The processing model 6

3.3. Formal mapping of the TEI elements used by Simple to the CIDOC CRM 8

3.4. TEI-Performance Indicators 8

3.5. Integration of TEI Simple into the TEI infrastructure 8

4. The audiences of TEI Simple 9

5. Program of work 9

5.1. Project participants 10

5.2. Advisory Committee 10

5.3. Work program 11

6. Budget Narrative 12

7. Appendix A: Short biographies of participants 14

# Summary

TEI Simple aims to define

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

The project for creating TEI Simple will be directed by Sebastian Rahtz (University of Oxford), Brian Pytlik Zillig (University of Nebraska-Lincoln), and Martin Mueller (Northwestern University). Work on the project will start August 1, 2014 and end by July 31, 2015. It has a budget of $122,000, of which $72,000 have been secured through firm commitments of money or time from the TEI Consortium, the Centre for Digital Research in the Humanities at the University of Nebraska-Lincoln, the University of Oxford, Northwestern University, and the Berlin-Brandenburg Academy of Sciences. We request a matching grant of $49,800 from the Mellon Foundation. The grant would be managed by Northwestern University, with Nebraska and Oxford as subcontractors.

# Background

The Text Encoding Initiative (TEI) has developed over 20 years into a key technology in text-centric humanities disciplines, with applications ranging widely 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. Limited interoperability has been the cost of this flexibility. In our view there is a distinct set of uses (primarily in the area of digitized ‘post-Gutenberg' books) that would benefit from a *prescriptive* recipe for digital texts; this would 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 era, 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.

In earlier discussions this project was called "TEI Nudge " because it was offered in the spirit of Richard Thaler's and Cass Sunstein's *Nudge* with its persuasive argument that in many walks of life people will make better choices if they are offered well-designed default solutions, as long as they are free to opt out of them. The TEI architecture facilitates customizations of many kinds; TEI Simple aims to produce an 'out of the box' customization that 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’ is 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’, or ‘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-1)

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 them in 2020, creating a 70,000 volume and public domain archive of consistently encoded versions of just about every 'different book' published in England or North America before 1700, including works of literature, philosophy, politics, religion, geography, science and 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 texts, 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. Between 2011 and 2013 extensive work was done at Nebraska, 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. Given the extraordinary degree of internal diversity in the EEBO source files, we believe that a project starting 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 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. Developing and implementing processing rules for TEI Simple and creating a notation (as an extension to TEI's ODD metalanguage) 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 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” indicating to a user what they can expect to use the text for.
5. Full integration of TEI Simple into the TEI Guidelines and infrastructure with ongoing maintenance by the TEI Technical council

## The highly constrained and prescriptive element subset of TEI Simple

### The sources for TEI Simple

The first step in this project will be the definition of the TEI Simple dtd or schema. We envisage this as a quite practical exercise that draws on the experience of projects that have attempted to "herd" large numbers of diverse texts into confines of a single schema. The experience of the MONK project was particularly instructive. There we (Mueller and Pytlik Zillig) constructed a schema (TEI-Analytics) to accommodate various archives that were encoded by following the "recommendations for Level 4 of the TEI in Libraries Guidelines." The archives included EEBO and ECCO TCP texts, the Wright American Fiction Archive, Documenting the American South, and Chadwyck-Healey's 19th century fiction collection. Among them, these projects used about 80 elements (not counting the bibliographical elements in the teiHeader)—less than 20% of the more than 500 TEI elements.

It was both easy and difficult to construct a common schema and get all texts to validate under it. It was easy because harmonizing the encoding practices never forced us to sacrifice expressiveness for the sake of homogeneity. But it was also quite difficult because it took time and a lot of ingenuity to harmonize the different and sometimes quirky ways in which projects handled this or that accidental. The major lesson of TEI-Analytics was this: as long as encoding stays within the range of phenomena about which there is little or no disagreement, nothing is lost and much is gained by doing the same thing in the same way. The range of phenomena to which this rule applies is wider than many people think.

The MONK project led indirectly to a two-year effort to harmonize the EEBO dtd (still expressed in P3 SGML) with P5, the latest and natively XML version of the TEI standard. Over the years the EEBO dtd had diverged from standard TEI, sometimes quirkily, but mostly in response to situations where the TEI standard was too rigid to cope with what encoders found on the varied terrain of Early Modern print culture. In those cases TEI P5 bowed to the experience of 25,000 TCP encoded texts and in the process became a more flexible tool. Both Pytlik Zillig and Rahtz have produced conversion tools for creating TEI P5 texts from EEBO-TCP texts.

The German Text Archive (DTA), a project of the Berlin-Brandenburg Academy of Sciences, has produced a pure subset of TEI P5 designed to support the compilation of an interoperable corpus of German writing from 1600 to 1900. This is a very carefully thought-out and documented schema that has been adopted by Clarin-D as a "base format" for other encoding projects. There are many similarities between the DTA schema, the EEBO schema, and the Level 4 Guidelines schema. This is not surprising, since all these schemas were designed to deal with Western books from Lydgate's Troy book (1473) to Joyce's *Ulysses* (1922), and the conventions of print culture across the languages of Europe and the Americas are very similar.

We will begin the task of defining TEI Simple by creating a union set of all the elements used in the schemas so far mentioned, see what can be subtracted, and make necessary additions. Whatever else it does, TEI-Simple must allow for the easy inclusion of TCP texts, the DTA archive, the Level 4 Guidelines archives of American research universities. We will follow the DTA design principle that TEI Simple must be a pure subset of TEI-all. A document that validates under TEI-Simple must validate under TEI-All. While believe that TEI-Simple with its approximately 100 elements will cover a great many cases, it will not cover everything, and where it does not, it should be a good platform for extension. It should also be possible to reduce a more richly encoded text to a sharable base-level encoding. Both these scenarios require that the base level is a pure subset of the complete schema.

The final subset of TEI Simple will include mandatory vocabularies for *@type* attributes, and more constrained content models for some elements, aiming to reduce the number of ways an encoder can record the semantics of the digital text.

### A note on manuscripts

TEI Simple is a schema for books from the first 500 years of Western print culture. The transcription of manuscripts is out of scope for this project, but it is something we want to keep an eye on. Robert Binkley in an often-cited *Yale Review* essay of 1935 (a virtual contemporary of Walter Benjamin's more famous essay about the loss of 'aura') saw the technological innovations of his time as countering the pressures of mass production and as " capable of working the other way—as implements for a more decentralized and less professionalized culture, a culture of local literature and amateur scholarship." One might say the same of Internet-based tools that support the collaborative transcription and annotation of manuscripts, whether an ambassador's letter in the Venetian state archives or the diary of a nineteenth-century farmer in Henry County, Indiana. In the digital world manuscript transcription remains as fundamental a philological activity as it has been for centuries, and the humble act of transcription may still be the best introduction to experiencing the joys and challenges of working with archives. There is a large role for the TEI in in making transcribed manuscripts sharable and interoperable documents. Projects like Will Thomas' [History Harvest](http://historyharvest.unl.edu) could benefit from a "TEI MS Simple," provided it can substantially lower the real or perceived obstacles to transcription. This is a project beyond TEI Simple, but we want to design TEI Simple in such a way that a manuscript-based project can build on or re-use parts of it.

## The processing model

A "cradle to grave" processing model is at the heart of this project. For many scholarly users the encoding of a text is the easy part. Its principles are readily grasped because the problems faced in it are closely related to problems of making sense of a text. Encoding is a way of making sense of a text. What to do with the encoded text is another question. It very quickly takes you into a world of acronyms and technologies that are not easy to learn, especially if you do not really want to learn in the first place. A very similar problem exists on the other side of the divide: for clever Web programmers html, json, jquery, php, and mysql are second nature but they may be put off by the perceived complexity of TEI, and they do not want to go there.

The processing model offers a bridge across that divide. 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. Scholars or other individuals who primarily care about getting the encoding right need not concern themselves with subsequent stages. If they can do their encoding within the limits of TEI Simple, they can be confident that the fruits of their encoding are available to users. Users, in turn, can have greater confidence that the query potential of encoded texts is in fact made available to them through appropriate query routines. This is conspicuously not the case with most TEI-encoded texts now: it is very difficult to write or train a search engine if you cannot make reasonably stable assumptions about the structure of the documents to be queried.

In this context it may be worth drawing attention to [BlackLab](https://github.com/INL/BlackLab), the Lucene based search engine developed by the Institute for Dutch Lexicology, which is rapidly acquiring limited but quite powerful XML-aware search capabilities. The combination of a TEI-simple encoded corpus with a search engine like BlackLab is a very powerful philological resource and greatly increases the chances that what was encoded is in fact available for decoding by the end user.

TEI Simple will develop a notation by which a TEI profile expresses how documents meeting that profile can be published or made accessible. TEI elements are generally *descriptive* of the interpreted semantics of the source text (‘this is a title’, ‘this is a quotation’), where the interpretation is often based on a human interpretation of layout. The *@rend* attribute is sometimes, but not necessarily, used to describe that layout. The resulting encoded text is very amenable to analysis, but for the common case of re-presenting the text on the web, the TEI semantic model is incomplete.

There are three stages to the processing model for TEI Simple.

1. In the first stage, every TEI Simple element is assigned to a category. This allows a processor to know whether to handle the element or not, and broadly speaking how to display or otherwise process it. The table below is a first stab at such a categorization, and the final list may be a little longer or a little shorter. Note how XPath may be used to refine a name.

|  |  |  |
| --- | --- | --- |
| **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. In the second stage, the categories are mapped to a presentation format, using HTML and CSS concepts where possible.
2. 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 its current "context," i.e. its place in the hierarchy of "parents", "ancestors," "siblings", and "children", to use the family metaphors that have become the technical jargon for defining the place of a given element in the "tree structure" or hierarchy of an XML document.

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 TEI Simple, so a processor will read both the source document and the corresponding ODD file (or some compiled version of it), 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.

How does this processing model differ from the various "TEI to XXX " stylesheets, which have been developed many times and of which [Rahtz' implementations](https://github/com/TEIC/Stylesheets) are probably the most widely used example? A stylesheet is just a bundle of rules that turn an input like "<supplied>foo</supplied>" into an output like "[foo]." A very simple business, but it gets complicated fast, if there are many rules and they get in each other's way. From one perspective, the processing model for TEI Simple is just a stylesheet for a tighter schema, and much of Rahtz's work on stylesheets will directly or indirectly find its way into TEI Simple. But the rules are more tightly integrated into the schema, and they are part of the same document (ODD) that defines the schema. They are also formulated at a higher level of abstraction: instead of stylesheets, there is a framework that is used to compile an ad hoc style from and ODD for a given domain.

## Formal mapping of the TEI elements used by Simple to the CIDOC CRM

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 CIDOC Conceptual Reference Model (CRM) are parts of an evolving ecosystem of metadata standards for cultural heritage documentation across the many languages and cultures of Europe. Several investigations and proofs of concept have demonstrated a close relationship between much of the TEI and the CIDOC CRM ontology. The TEI ODD already has a notation (the <equiv> element) for expressing the relationship between TEI and RDF, and this will be used (possibly with proposed extensions to the ODD language) to secure map all the elements of Simple to the CRM and FRBRoo. The implementation of this will follow that of the processing model, and be embedded in the same transformation toolset.

## TEI-Performance Indicators

Although TEI Simple is designed to be very constrained, it will still allow for many choices by 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? This will affect the *query potential* of a corpus of texts, and cannot be done simply by analyzing the markup.[[2]](#footnote-2) The TEI already has extensive provision in the metadata header for describing the encoding decisions which have been made, but this is largely targeted at storing human-readable notes, and is thus not machine readable. This objective is to develop and implement an extra level of notation aimed at automatically profiling a text. This notation will be stored as machine-readable data in the teiHeader. If mediated through a good visualizing tool, such data can alert an end user to the fact that texts in a corpus do or not include linguistic annotation, named entity identification, etc. They are, if you will, metadata about metadata. TEI Simple will not develop that visualizing tool, but it will lay the groundwork for data structures that routine visualizing tools can easily access and mediate—provided the data are there in the first place and in a machine-readable form.

## Integration of TEI Simple into the TEI infrastructure

TEI Simple will be fully integrated into the TEI infrastructure once all its elements are in place, consisting of

1. a documented definition in ODD of the TEI subset,
2. a set of processing rules,
3. and extensions to the TEI ODD language to record processing expectations,

The TEI Council will be responsible for its maintenance and further development. TEI Simple will be very fully documented, and the documentation will take great care to address the need of novice users. Like the TEI Guidelines, the maintenance of TEI Simple will be a continuing responsibility of the TEI Council. Whether the documentation will be absorbed into the Guidelines or remains as a separate document is a practical question. Perhaps the two documents will relate to each other like Luther's Large and Small Catechism. The design of the documentation will make it easy to move from one to the other.

The TEI standard and all documents related to the maintenance of that standard are fully in the public domain. So will TEI Simple and all documentation relating to it.

# The audiences of TEI Simple

### Individuals

TEI Simple should appeal to different individual audiences across different ancient and modern languages, including Greek, Hebrew, Latin, and the Germanic, Romance, and Slavic languages of Europe and America. Individuals will find it complementary to the TEI Archiving Publishing and Access Service (TAPAS) that will debut later this year. Broadly speaking, TAPAS encourages diverse encoding practices on a shared infrastructure, while TEI Simple encourages uniform encoding that will work on different infrastructures. There is room for both in the world, and by temperament, experience, or circumstance, project leaders will tend towards one or the other.

### Libraries

We hope that libraries will take an interest in TEI Simple, and in whatever marketing we end up doing we will emphasize what TEI Simple can do for libraries. At the most basic level, TEI Simple will be compatible with the third edition of "[Best Practices for TEI in Libraries: a guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI.](http://www.tei-c.org/SIG/Libraries/teiinlibraries/) At a higher level, TEI Simple's association of a schema with a processing model solves many problems of what to with the texts. Many "Best Practices" texts will validate under TEI-Simple. It will be trivial task to convert others through programs like Oxford's OxGarage or Nebraska's Abbot. If libraries find it convenient to use TEI Simple, there is likely to be some increase in cross-institutional interoperability. Incremental improvements in interoperability are a good thing.

# Program of work

Much of the work on this project will be done at Oxford and by people at Oxford. With one exception, face–to-face meetings will be scheduled in Oxford because that will minimize travel expenses both for the direct participants and for the European members of an advisory committee that we would like to recruit.

## Project participants

The principal investigators of the project are Sebastian Rahtz, Brian Pytlik Zillig, and Martin Mueller.

Sebastian Rahtz will be in charge of the overall direction of the project. Rahtz is Director (Research) of Academic IT in the University of Oxford. He has served on the Technical Council of the TEI since 2001 and is widely known in the TEI community for his work on anything that involve the transformation of a TEI document into something else.

Brian Pytlik Zillig is professor and librarian at the University of Nebraska-Lincoln. He has special expertise in XSLT programming. In the MONK project he developed routines for harmonizing diversely encoded TEI documents, which were later perfected in [Abbot](http://abbot.unl.edu/cocoon/vicar/Signin.html), a tool designed to convert dissimilar collections of XML texts into a common interoperable format. Pytlik Zillig will be responsible for much of the implementation of the processing model at the core of TEI Simple.

Martin Mueller is Professor emeritus of Classics and English at Northwestern University. In the MONK project he focused on increasing interoperability among TEI encoded texts from different collections. He will act as coordinator in the project and be responsible for its documentation.

Other participants in the project include Lou Burnard, James Cummings, Magdalena Turska, and one or two developers to be appointed.

Lou Burnard was the co-editor of the original TEI Guidelines , founded the Oxford Text Archive, was a key player in the design of the British National Corpus, and is one of the leading authorities on the TEI standard.

James Cummings is a Senior Digital Research Specialist for the [IT Services](http://www.it.ox.ac.uk/) of the [University of Oxford](http://www.ox.ac.uk/). He is the Oxford principal investigator for the [Marie Curie ITN 'DiXiT'](http://dixit.uni-koeln.de/) on digital scholarly editions. He has served on the TEI Technical Council since 2005 and has been its chair since 2011 .

Magdalena Turska is a researcher at Oxford on a fellowship from Marie Curie Dixit Marie Curie DiXiT (Digital Scholarly Editions Initial Training Network) project.. Her role in this project will be to explore bridges from TEI Simple to the more complex structures of full-fledged digital editions.

## Advisory Committee

We anticipate an advisory committee with representation from the TEI, and significant TEI projects in France and Germany. James Cummings will represent the TEI as chair of its Technical Council. Alexander Geyken will represent the German Text Archive. We have not yet settled on a French project.

## Work program

The work will proceed in five phases from August 2014 through June 2015. Much of the critical work will be done in in "work sprints," where the participants can give uninterrupted attention to it for periods of three or five days. Pytlik Zillig will visit Oxford for three such periods, and work during those visits will focus on the design and implementation of the processing model as the central feature of TEI Simple. We attach actual and percentage costs to each phase The percentage figures, which add up to 91% do not include a contingency fund of $6,000 and the costs of a meeting in Chicago. They are rough estimates subject to change, but they may still be useful in communicating a sense of the relative importance of the different parts of the project.

### August –September 204: defining the TEI Simple schema

Work on the design of the TEI Simple schema will begin in August and culminate in a work sprint in mid-September, when Mueller and Pytlik Zillig will spend a week in Oxford. The first meeting of the Advisory Committee will take place during this phase.

Total costs for Phase 1 are $25,600 or 21% of the total budget.

### October – December 2014: Designing and implementing the Processing Model

Work on the processing model will begin in October 2014 and continue through December. A work sprint in October will involve Rahtz and Cummings and focus on the development of a model for notating the processing in ODD. Pytlik Zillig and Rahtz will continue to work on the processing separately and complete this phrase in a five-day work sprint in December during Pytlik Zillig's second visit to Oxford.

Mueller, Pytlik Zillig and Rahtz will meet in Chicago in late October, following the TEI 2014 meeting to review the progress of the project

Total costs for this phase are $46,600 or 38% of the total budget.

### January –June 2015: Documentation

Documentation of the project will begin in January 2015, when the TEI Simple schema and processing model are complete. The documentation will be written by Mueller with help from Burnard and Cummings.

Total costs for this phase are $4,800 or 4%. This figure is deceptive since Mueller's time as an emeritus is not on the budget. But he expects to spend at least a month writing and re-writing documentation.

### March 2015: Performance indicators

The design and implementation of the "TEI Performance" indicator will be the work of Rahtz, and Pytlik Zillig, who will go to Oxford for his third and final work sprint. Other project staff will assist.

The costs for Phase 4 are $17,300 or 14% of the total budget.

### April 2015: Mapping TEI Simple to RDF and other ontologies

The mapping of TEI Simple to RDF , CIDOC-CRM and other ontologies of the Europeana Data Model will be done by project staff during April 2015.

The costs for Phase 5 are $3,600 or 3% of the total budget.

### May 2015: Integration of TEI Simple into TEI infrastructure

Lou Burnard and James Cummings will be responsible for the integration of TEI Simple into the TEI infrastructure.

The costs for Phase 6 are $3,600 or 3% of the total budget.

### User work shops and dissemination

At some point of the project we will conduct workshops with users in Lincoln and Oxford to test TEI Simple. We will also commit some resources to disseminating information about the progress of the project.

The costs for this part of the part of the project are $9,600 or 8% of the total budget.

# Appendix A: Short biographies of participants

**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.

**Brian L. Pytlik Zillig** is a Professor, University Libraries, at the University of Nebraska-Lincoln. He is the Digital Initiatives Librarian and participates in the project design and management of digital scholarship projects. His digital projects include Token-X, an XSLT based text analysis tool, and he was the chief architect of [Abbot](http://abbot.unl.edu/cocoon/vicar/Signin.html), a tool designed to convert dissimilar collections of XML texts into a common interoperable form. His publications include:

Pytlik Zillig, B. L. (2013) Logging the Abbot: Reflection‐Oriented XSLT Programming for Corpora Conversion and Verification. *Journal of the Text Encoding Initiative*, (March).

Lorang, E., Pytlik Zillig, B. L. (2012). Electronic Text Analysis and Nineteenth-Century Newspapers: TokenX and the Richmond Daily Dispatch. *Texas Studies in Literature and Language, 54*(3), 303-323.

Pytlik Zillig, B. L. (2011). 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

Bernholz, C., Pytlik Zillig, B. L. (2011). Comparing nearly identical treaty texts: a note on the Treaty of Fort Laramie with Sioux, etc., 1851 and Levenshtein’s edit distance metric. *Literary and Linguistic Computing*, *26*(1), 5-16.

Pytlik Zillig, B. L. (2009). TEI Analytics: converting documents into a TEI format for cross-collection text analysis. *Literary and Linguistic Computing, 24*(2), 187-192.

Bernholz, C., Pytlik Zillig, B. L., Anderson, C.G. (2008). The Words You Brought Us: Digital Lexicon of the Recognized American Indian Treaties. *Library Collections, Acquisitions, and Technical Services, 32*, 104-106.

Bernholz, C., Pytlik Zillig, B. L., Weakly, L., Bajaber, Z. (2006). The Last Few American Indian Treaties. An Extension of the Charles J. Kappler Indian Affairs: Laws and Treaties Internet Site at the Oklahoma State University*, Library Collections, Acquisitions, and Technical Services, 30*, 47-54.

Barney, B., Ducey, M. E., Jewell, A., Price, K., Pytlik Zillig, B. L., Walter, K. L. (2005). Ordering Chaos: An Integrated Finding Aid and Online Archive of Walt Whitman's Poetry Manuscripts. *Literary and Linguistic Computing, 20*(2), 205-217.

**Martin Mueller** is Professor emeritus of English and Classics at Northwestern University. He is the author of a monograph on the Iliad (1984, 2nd edition, 2009) and of *Children of Oedipus and other essays on the imitation of Greek tragedy, 1550-1800 (*1980). His digital projects include [The Chicago Homer](http://digital.library.northwestern.edu/homer/) and [Wordhoard](http://wordhoard.northwestern.edu/userman/index.html), an application for the close reading and scholarly analysis of deeply tagged texts. Most recently he has worked with Phil Burns on [MorphAdorner](http://morphadorner.northwestern.edu), a project to provide part-of-speech tagging, lemmatization, and virtual orthographic standardization for Early Modern texts.

**Lou Burnard,** formerly assistant director at Oxford University Computing Services, 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.

[James Cummings](http://blogs.it.ox.ac.uk/jamesc/about/) is the Senior Digital Research Specialist for the [IT Services](http://www.it.ox.ac.uk/) of the [University of Oxford](http://www.ox.ac.uk/). He helps academics from the Humanities and Social Sciences plan research projects with digital aspects, and is the department's liaison for [Digital Humanities at Oxford](http://digital.humanities.ox.ac.uk/dhoxss/) activities. James is Director of the annual week-long [Digital Humanities at Oxford Summer School](http://digital.humanities.ox.ac.uk/dhoxss/). James often acts as a liaison between developers and researchers, sometimes taking on principal investigator or project management aspects of digital humanities projects. He still does some active coding and development on workpackages related to XML, markup, schemas and digital editions. He is the Oxford principal investigator for the [Marie Curie ITN 'DiXiT'](http://dixit.uni-koeln.de/) on digital scholarly editions.

Since 2005 he has been an elected member to the [Text Encoding Initiative Consortium](http://www.tei-c.org/)'s [Technical Council](http://www.tei-c.org/Activities/Council/) and is now its Chair. This technical council develops new features and fixes bugs in the TEI Guidelines for encoding digital text, and undertakes other activities related to supporting this standard. James has a great deal of experience in teaching digital publishing concepts and XML-related technologies.

James completed a [PhD in medieval drama](http://james.blushingbunny.net/phd.html) from the [University of Leeds](http://www.leeds.ac.uk/), an MA in Medieval Studies from the [University of Leeds](http://www.leeds.ac.uk/), and a BA in Medieval Studies from the [University of Toronto](http://www.utoronto.ca/). [His PhD](http://james.blushingbunny.net/phd.html) investigates the archival survival of information concerning the performance of drama in medieval culture. It involved a significant amount of archival transcription.

In addition to the archival survival of information about early performance, the relationship of medieval manuscripts to their digital surrogates is one of his interests. From 2009-2012 he was the elected director of the executive board of the [Digital Medievalist](http://www.digitalmedievalist.org/) project.

In addition to the work on TEI-Simple, James plans to use it as a starting point for a schema targeting scholarly digital editions of medieval manuscripts.

**Magdalena Turska is a** Researcher for the Digital Scholarly Editions IT Services, University of Oxford and is a 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. She was the technical editor and 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

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-1)
2. The absence of <persName> markup may be because the encoder has decided not to distinguish them, or because there are no names. [↑](#footnote-ref-2)