TEI by example, initial report. [TBE-R001]

Ron Van den Branden, 2006/06/09

Report details	Ron Van den Branden, *TEI by example, initial report. [TBE-R001]*, Gent, 2006/06/09.
Author details	Ron Van den Branden Centre for Scholarly Editing and Document Studies Royal Academy of Dutch Language and Literature, Belgium Koningstraat 18, B-9000 Gent, Belgium ron.vandenbranden@kantl.be
Project title	TEI by example. Creating a toolkit and modules for teaching text encoding.
Project leaders	Melissa Terras, Edward Vanhoutte
Project website	http://www.teibyexample.org/
Project email	teibyexample@kantl.be

Table of contents

1. Introduction: *TEI by example*

This project aims to address an existing need for introductory training materials to text encoding within the Text Encoding Initiative (TEI) framework. Although an intitial objective of the TEI Training SIG[1], active development of tutorial materials has met with a lack of community contribution and consequently been abolished. Nevertheless, if the community wants to promote the TEI markup framework seriously, there is an urgent need for an on-line TEI course by example that introduces novice users to text encoding within the TEI framework and can serve as an introductory teaching package for instructors. Likewise, the availability of a software toolkit for teaching text encoding would support interested trainers to take up the challenge to teach TEI on several occasions. Therefore, this project will develop and document following deliverables:

- on-line tutorials *TEI by example*
- on-line software toolkit for text encoding

This initial report will identify points of attention and problems to be anticipated in section 2. These will be incorporated in a general workflow, presented in section 3. Specific needs for hosting of the project will be addressed in section 4.

2. Points of attention

2.1 What is TEI (nowadays)?

A major point of attention at the start of the project must be the status of the TEI model. Since early 2002, the TEI Consortium has been engaged in a major (backward-incompatible) revision of the TEI specification, migrating it from version P4 (2002) to P5 (200?). Featuring more than just changes in the markup model and the content of the guidelines, P5 entails an overhaul of the complete production process of the standard. A new dimension was added to the production process with the transformation of P5 to a sourceforge project[2] early 2004, introducing publicly accessible CVS repositories holding the source files for the P5 guidelines, the XSLT stylesheet suite, the Roma source for deriving TEI customisations, and a few other CVS modules (extensions, I18N, TEI Open Office auxiliary files, tei-emacs customisation). Apart from this 'real-time' dynamic release mode, every 6 months[3] snapshot releases are issued, containing the source for the guidelines and all derived documents (guidelines, DTD / Schema fragments). Up to now, 5 snapshots have been released, showing substantial differences[4]:

- 0.1.1 (2005-01-18)
- 0.1.2 (2005-02-08)

[1] TEI Training SIG: <http://www.tei-c.org/Activities/SIG/Training/>
[2] TEI sourcefoge project: <http://sourceforge.net/projects/tei/>
[3] Report of the TEI Council activities to the Members Meeting 2005, <http://www.tei-c.org/Council/Tcr03.xml>
[4] Posting on TEI-List <http://listserv.brown.edu/archives/tei-l.html> by Sebastian Rahtz, 2005/10/20 [TEI P5 release 0.2.1]

- 0.1.10 (2005-07-15)
- 0.2.1 (2005-10-20)
- 0.3.1 (2006-01-30)

Recent postings on the TEI-List[5] suggest that a) a final release of P5 will not be envisaged by the editors; yet b) stabilisation can be expected from the next snapshot onwards (planned April / May 2006).

Apart from the innovations regarding the content of the TEI markup scheme, adoption of P5 involves coping with peripheral technical innovations. The TEI Pizza Chef software for deriving P4 TEI DTDs has been superseded by the Roma system, allowing users to derive TEI customisations in a number of formal expressions, from the (innate) Relax NG Scheme to DTDs or W3C XML Schemas.

To summarize, it seems that the timing of this *TEI by example* project coincides with a turning point in the transition of TEI P4 to P5. A choice must be made, however, for one version of the standard to be used in the project. Following tables may provide evaluation criteria:

Pros and cons for adoption of P4 in the project:

+	-
[TEI-external]: guaranteed adherence to well-established text encoding practice	Risk of quick obsolescence

Pros and cons for adoption of P5 in the project:

+	-
[TEI-external]: incentive for P5 adoption	Intrinsic instability
[TEI-internal]: thorough evaluation of practice + contents of selected chapters	Steep technical learning curve for novice text encoders (CVS / multiple expression formats / still maturing Schema software support)
Technical flexibility (Schema flavours but also DTD) and the ODD abstraction layer equally allow to concentrate on content (if this is an argument for novices)	

All in all, the advantages of P5 adoption for this project seem to outweigh the disadvantages of P4. The most recent snapshot indeed suggests that stability is at hand, which is reflected in a more comprehensively updated documentation and preparations for a P5 TEI Lite revision. These may inform following workflow:

- adopt a current TEI snapshot at the start of the project
- “freeze” the didactic ODD + Relax NG / W3C XML Schema / DTD

[5] Postings on TEI-List <http://listserv.brown.edu/archives/tei-l.html> by Sebastian Rahtz, 2005/11/08 [Re: final release of P5?], and Lou Burnard, 2006/01/11 [Re: Minor P5 issues].

- make that choice explicit (e.g. by using the `TEI/@version` attribute throughout the examples)
- monitor changes to TEI during the project
- reserve a testing / debugging / updating phase before delivery

2.2 Didactic approach

TEI is a complex and comprehensive system of provisions for scholarly text encoding. The last printed edition of the Guidelines documents 362 distinct XML elements in 1067 pages spread over 2 volumes. To illustrate the use(ability) of this markup system, a specific subset of TEI (TEI Lite) was developed. TEI Lite aimed to address three purposes[6]:

1. to provide a TEI subset which would be adequate to most common needs
2. to provide a subset of TEI rich enough to support an authoring environment for the production of online documentation
3. to demonstrate the customization facilities of the TEI scheme

Although the main didactic locus of TEI Lite was situated more in a demonstration of the use of TEI through its customisation facilities[7], its reduced scope, wide applicability and comprehensible documentation made TEI Lite a very popular introduction to the TEI scheme. However, the first two goals still account for the fairly generic character of the TEI Lite documentation, with more features of a trimmed-down reference manual than of a proper tutorial. On the other hand, the guides to local practice featuring in the TEI Tutorials webpage[8] are quite specific (and often not very recent).

This project aims to bridge this gap by developing step-by-step practical introductions to thematic use cases for TEI encoding. A didactic approach could be modeled to that of qualitative web development tutorials, like:

- [TEI-specific]:

  TEI documentation: <http://www.tei-c.org/>, TEI Talks: <http://www.tei-c.org/Talks/>

- [generic web development tutorials]:

  W3 Schools Web Tutorials: <http://www.w3schools.com>, Web Monkey: <http://www.webmonkey.com>, Code Walkers: <http://www.codewalkers.com>, IBM developerWorks <http://www-128.ibm.com/developerworks/>

These suggest some valuable principles that can guide the structure of the tutorials in this project:

- meta-introduction: objectives, prerequisites, system requirements

[6] See TEI ED W85: *TEI Lite from P4 to P5: Continuity and Change*, <http://www.tei-c.org/Drafts/edw85.xml>

[7] TEI Lite Prefatory Note: "TEI Lite was the name adopted for what the TEI editors originally conceived of as a simple demonstration of how the TEI encoding scheme might be adopted to meet 90% of the needs of 90% of the TEI user community."

[8] TEI Tutorials webpage: <http://www.tei-c.org/Tutorials/>

- thematic introduction: high-level overview
- step-by-step introduction of new markup: short text excerpts and examples
- summary
- download area: sample code
- resources: links to background information and reference material

2.3 Didactic setup

Related to the choice of a TEI flavour is the question for practical scope of the didactic setup. Since the project explicitly aims at introducing novice users to digital text encoding with TEI, it could make sense to include a choice for a text editor in the didactic setup. Without willing to restrict the exemplified use of TEI to just one software configuration, consequent use of a text editor throughout the tutorial chapters might alleviate the technical learning curve. Besides didactic issues (as suggested by sometimes vehement discussions on the TEI-List[9]), adopting TEI P5 furthermore adds technical requirements to the choice for an XML editor. A good starting point for identification of useful criteria for this choice is provided by the AHDS Information Paper *Choosing an XML editor* (Van den Broeck, 2004). For this project, following features could be proposed:

- code-based XML editing
- validation (against DTD, Schema (varieties))
- XML-aware editing assistance (tag completion, context-sensitive tag suggestion)
- free
- support for different operating systems
- activity status: i.e. active maintenance and updates of the software

A number of good basic XML-aware open source text editors exist, as included in the project software toolkit, but at the moment of writing, Schema support (W3C XML Schema, Relax NG) is rather infrequent in non-commercial XML editors. Following editors are potentially interesting higlights from the evaluation in the AHDS study mentioned and some excellent online XML software listings[10]:

[9] A search containing the terms 'XML' and 'editor' on the TEI-List (<http://listserv.brown.edu/archives/tei-l.html>) will produce a long list of postings proving the sensitivity of the TEI community to this topic.

[10] Useful online XML software listings:
- Linda Van den Brink's *XML Software* page <http://www.xmlsoftware.com/>
- Lars Marius Garshol's *Free XML Tools* page <http://www.garshol.priv.no/download/xmltools/>

	Schema	Auto-completion	Validation	Platform	License	Cost
TEI-emacs[11] (21.3)	DTD, Relax NG	+	on-the-fly	Windows, Mac, Linux, *nix	open source (GNU GPL)[12]	free
Jedit[13] (4.2)	DTD, W3C	+	on-the-fly	Java	open source (GNU GPL)	free
Eclipse[14] + Web Standards Toolkit[15] (1.0.1)	DTD, W3C	+	at prompt	Java	open source (EPL)[16]	free
Architag XRay[17] (2.0)	DTD, W3C	-	on-the-fly	Windows XP	restricted	free
Butterfly XML editor[18] (1.0 beta)	DTD, W3C	+	on-the-fly	Java	open source (GNU GPL)	free
OXygen[19] (7.0)	DTD, W3C, Relax NG	+	user prompted	Java	commercial[20]	$180
Exchanger[21] (3.2)	DTD, W3C, Relax NG	+	user prompted	Java	academic, Lite	free
					academic, Pro	$48

Far from pretending exhaustivity, this list presents an overview of different potentially interesting options. One end of the spectrum features non-dedicated text editors with XML add-ons like jEdit, eclipse and emacs. These are often rooted in software development communities and conceived as advanced text editors. Through extension mechanisms they can be enriched with powerful XML-aware editing features. Still, they remain quite technical, require some degree of configuration to set up and sometimes impose considerable system requirements. In the TEI community, emacs has been considerably supported through TEI-enabled emacs packages. Although it can be considered a very performant TEI promoted editor-of-choice, it is generally deemed rather technical for introductory purposes. Another category are dedicated small-scale XML-editors like XRay and Butterfly. Although rather limited, they can implement very qualitative XML editing features. Often the result of small software development projects, they remain rather experimental and sometimes buggy. Stability and performance are optimalised in industrial-strength dedicated XML editors. Of the few (nearly) free pieces of software, oXygen and Exchanger Lite deserve some comments.

[11] TEI-Emacs: <http://www.tei-c.org/Software/tei-emacs/>
[12] GNU GPL (GNU General Public License): <http://www.gnu.org/copyleft/gpl.html>
[13] jEdit Programmer's Text Editor: <http://www.jedit.org>
[14] eclipse: <http://www.eclipse.org/>
[15] eclipse Web Standards Toolkit: <http://www.eclipse.org/webtools/>
[16] EPL (Eclipse Public License): <http://www.eclipse.org/org/documents/epl-v10.php>
[17] Architag XRay XML editor: <http://architag.com/xray/>
[18] Butterfly XML editor: <http://www.butterflyxml.org/>
[19] oXygen XML editor: <http://www.oxygenxml.com/>
[20] oXygen End User License Agreement: <http://www.oxygenxml.com/eula.html>
[21] Cladonia Exchanger Lite XML editor: <http://www.freexmleditor.com/>

“Politically”, the commercial oXygen XML editor is the odd one out of this list. Its inclusion is motivated from a TEI-internal point of view. First, it comes packaged with TEI DTDs, Schemas and stylesheets that facilitate TEI editing. Second, the TEI consortium has negotiated a discount on the oXygen XML editor for TEI members and subscribers. This discount of 20% (reducing the price for an individual academic license + maintenance pack to 51.20 USD)[22], might add an extra motivation for TEI membership for the target audience of this project. The Exchanger Lite XML editor is a basic version of a commercial professional XML editor. Yet, it has surprisingly complete XML editing capabilities and is completely free for non-commercial academic purposes.

Of course the choice for a specific piece of software is always debatable, especially in a didactic context introducing people to text encoding with an explicitly application-independent technology as XML. Nevertheless, with necessary caution, explication and pointing to alternatives (for which can be referred to the software toolkit), the didactic benefits can justify the choice for one editing environment for the tutorials. The discussion above seems to suggest the Exchanger Lite XML editor as a satisfactory didactic tool to this end.

2.4 Text examples

Ideally, the examples illustrating the use of TEI markup would be based on existing digital texts. However, as the motivation for this project suggests, the amount of digitally available TEI marked up text materials is not overwhelming. Even within the TEI user community, explicit initiatives for sharing TEI texts have met with low enthousiasm for actually doing so[23]. Even the specific TEI Wiki for text samples currently lists 2 entries: one of which features collection of P5 texts, the other consisting of texts illustrating the transition from SGML P3 to XML P4[24]. Possibly the richest source of quality texts is the Oxford Text Archive (OTA), hosted by the Arts and Humanities Data Service. Although not all materials are available in TEI, mere availability would already help in assembling the examples for the tutorials. The OTA responded fairly positive on our inquiries, provided that the texts should not be used for commercial purposes, and should not be used integrally. Perhaps another call on the TEI mailing list would be a good approach. This implies a good view on the kind of materials this project aims at. Of course, the genre stratification of the different tutorial chapters requires texts illustrating TEI use for marking up prose, poetry, drama, manuscript transcription and scholarly editing. On the other hand, the amount of sample materials might be expected to influence the practical organisation of the tutorial chapters as well (e.g. progressive in-depth illustration of few examples vs. a larger number of different fragments).

An important point of attention in this regard is a copyright arrangement. To guarantee the rights of copyright holders as well as the liberal use of the deliverables of this project, a formal copyright notice provides the best solution. The modular system of licensing types developed by the

[22] Posting on TEI-List <http://listserv.brown.edu/archives/tei-l.html> by Syd Bauman, 2005/07/18 [new member benefit -- discount on oXygen with P5]

[23] See several calls on the TEI-List throughout the years, and intentions for a formal attempt by the TEI Presentation Special Interest Group, TEI-List <http://listserv.brown.edu/archives/tei-l.html> by Matt Zimmerman, 2005/01/31 [Presentation Tools SIG report - Baltimore 2004].

[24] TEI Wiki Samples page: <http://www.tei-c.org/wiki/index.php/Samples>

Creative Commons Initiative[25] can provide a good basis for this aim. Of the six Creative Commons licence types, following three non-commercial types could be considered:

- Attribution Non-commercial No Derivatives (by-nc-nd)

  allows unmodified redistribution under identical terms, for non-commercial purposes

- Attribution Non-commercial Share Alike (by-nc-sa)

  allows modified redistribution under identical terms, for non-commercial purposes

- Attribution Non-commercial (by-nc)

  allows modified redistribution under different (yet non-commercial) terms

3. Development of deliverables: workflow

Although preparatory actions can be undertaken earlier, content development of the project deliverables will start from July 2006, when Edward Vanhoutte will be able to devote 20% of his time to daily management of the project.

3.1 Creation of a software toolkit for text encoding

3.1.1 Aims

Based on the toolkit Edward Vanhoutte, Melissa Terras and Ron Van den Branden have compiled for several XML, XSL and digitization courses, this toolkit will present a reviewed set of freely available tools for the creation and digitization of texts. The toolkit brings together information under the following headings:

- Editors
- XML/XSLT processors
- XQuery tools
- Browsers
- XML Publication Systems
- Viewers
- Validation Services
- Miscellaneous Utilities
- DTDs
- Miscellaneous Files
- TEI
- DALF
- Specifications
- Guides to Good Practice
- Journals & Series

[25] Creative Commons: <http://www.creativecommons.org>

- Useful URLs

The toolkit will be presented as an annotated XHTML list, containing title, version number, description, installation instructions and homepage for each piece of software. It will be mirrored as a downloadable CD-ROM image file containing this index and local installers of the software. A CD-label in PDF, mentioning the Association's label, the title of the CD-ROM and the logos of the project's funders (King's College, CTB, ALLC, UCL, TEI) would also be offered.

3.1.2 Workflow

1. Send current list sent for review to co-proposers and TEI Education SIG in the first phase of the project
2. Checking status of tools (rights) and asking permission to include tools in toolkit before the start of the project
3. XML-izing current list
4. Updating current list (versions, reviews, installation instructions)
5. Reviewing new tools
6. Authoring introductory paragraphs
7. Wrapping up, transforming to XHTML, delivery
8. Creating CD-image
9. Maintenance (ongoing)

Timing: week 1

3.2 Creation of on-line tutorials for TEI

3.2.1 Aims

The on-line tutorials will introduce the novice to text encoding and the TEI by taking them step by step through the workflow of encoding a (set of) document(s). The encoding strategies demonstrated by example will be based on the version of the TEI Guidelines chosen (ideally: P5). The following table lists the chapters that are planned in the tutorial (column 1), the corresponding chapters of the TEI P5 documentation (column 2) and references to additional materials on which the chapters will be based (column 3):

	Chapters tutorial	Chapters TEI P5	Additional literature
1	Introduction to text encoding and the TEI	1, 2, 3	Vanhoutte (2004), Morrison e.a. (2000), Sperberg-McQueen & Burnard (2002b)
2	The TEI header	5, 24	Morrison e.a. (2000), Vanhoutte (2000), Vanhoutte & Van den Branden (2002)

	Chapters tutorial	Chapters TEI P5	Additional literature
3	Prose	6, 7, 14, 20	
4	Poetry	6, 7, 8, 14, 20	
5	Drama	6, 7, 9, 14, 20	
6	Manuscript Transcription	6, 7, 13, 14, 17, 18, 20	
7	Scholarly Editing	6, 7, 14, 17, 19, 20	
8	Customizing TEI, ODD, Roma	27, 28, 29, 30	Burnard (2005), Burnard & Cummings (2005)

Ideally, the tutorials will be encoded using the markup (version) described:

- minimally: TEI P4
- optimally: TEI P5, using provisions of the tagdocs module for integration of the example snippets

From the XML sources, XML, XHTML and PDF versions will be derived for delivery

3.2.2 Workflow

Each tutorial chapter can be developed in following steps:

1. Selection, analysis and markup of example document(s)
2. Analysis of appropriate TEI Guidelines chapters + additional literature
3. Setup of didactic approach
4. Step by step illustration and documentation
5. Finalizing tutorial
6. Revision
7. User testing and evaluation
8. Revision, update to latest TEI version where necessary
9. Transformation to XHTML + PDF

Timing: weeks 2-12

3.3 Other steps

3.3.1 Aims

Since the tutorial will be delivered in different formats, XSLT stylesheets will be developed for transformation of the XML sources to XHTML and PDF. Furthermore, the workflow of the project will be documented in a final project report. Finally, the tutorials will be tested by a panel of trainers and trainees. Trainers will be recruited from the members of the project committee and volunteers who reply to a call on TEI and Humanist public discussion lists. A user testing panel of trainees will be provided by students of the courses *Humanities Computing: Electronic Text* at the University of Antwerp and *Digital Resources in the Humanities* at University College London.

3.3.2 Workflow

1. Stylesheets:
    - analysis of tutorial input structure
    - analysis of tutorial output structure
    - development of XSLT and XSL FO stylesheets
2. Report: project diary, management report
    - project diary
    - revision report for each step taken
    - management report
    - aggregation in final report

Timing: weeks 1 - 12

3. User testing:
    - localization and identification of user test panel
    - authoring questionnaires for trainers and trainees
    - processing questionnaire results
    - revision of tutorials

Timing: outside project time

4. Hosting of the project: needs

The project will be published at <http://www.teibyexample.org/> and <http://www.teibyexample.com>. Needs can be expressed minimally and maximally:

- minimal: web server + storage room for static delivery
- maximal: web server + storage room + database + presentation layer (Java / Php) for dynamic delivery

Initial arrangements have been made with King's College London for hosting of the project. In the meantime, both URLs point to the provisional web page that has been set up at the Centre for Scholarly Editing and Document Studies. Depending on the exact possibilities, interactivity and dynamic delivery options can be provided. Minimally, delivery will be static, consisting of pre-compiled versions of the source and derived deliverables. If possible, dynamic delivery could add opportunities for user-driven interactivity, and possibly for integration in established electronic publishing models at the hosting institution. Further arrangements are needed, however, in order to produce a practical workflow protocol.

Another point of attention for the publication phase is employing channels like the TEI Wiki pages, the TEI Projects page, and the TEI List to publicise the project.

5. Referenced literature

Burnard, L. (2005). *EDW88: The TEI P5 How To*. <http://www.tei-c.org.uk/Drafts/edw88.xml>

Burnard, L. and Cummings, J. (2005). *Customizing the TEI*. <http://www.tei-c.org.uk/Talks/OUCS/2005-02/talk-contents.pdf>

Morrison, M., Popham, M. and Wikander, K. (2000). *Creating and Documenting Electronic Texts: A Guide to Good Practice*. Oxford: OTA. <http://ota.ahds.ac.uk/documents/creating/>

Sperberg-McQueen, C.M. and Burnard, L. (2002a). 'A Gentle Introduction to XML'. In: Sperberg-McQueen, C.M. and Burnard, L. (eds.) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange (XML-compatible edition)*. Text Encoding Initiative Consortium: Oxford, Providence, Charlottesville, Bergen. <http://www.tei-c.org/P4X/SG.html >

Sperberg-McQueen, C.M. and Burnard, L. (2002b). *TEI Lite: An introduction to Text Encoding for Interchange*. <http://www.tei-c.org/Lite/teiu5_en.html>

Sperberg-McQueen, C.M. and Burnard, L. (2005). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Revised and re-edited. Text Encoding Initiative Consortium: Oxford, Providence, Charlottesville, Nancy. <http://www.tei-c.org.uk/release/doc/tei-p5-doc/html/>

Van den Broeck, T. (2004), *Choosing an XML Editor.* Arts and Humanities Data Service Information Papers. <http://www.ahds.ac.uk/creating/information-papers/xml-editors/index.htm>

Vanhoutte, E. (2000). *It's all in the Head(er). From minimal to optimal use of the TEI Header*. <http://www.kantl.be/ctb/vanhoutte/pub/2000/headerproposal.htm>

Vanhoutte, E. (2004). 'An Introduction to the TEI and the TEI Consortium.' In: Matts Dahlström, Espen S. Ore & Edward Vanhoutte (eds.), *Electronic Scholarly Editing – Some Northern European Approaches.* A Special Issue of *Literary and Linguistic Computing*, 19/1: 9-16.

Vanhoutte, E. & Van den Branden, R. (2002). *DALF guidelines for the description and encoding of modern manuscript material*. Gent: CTB, 2002. <http://www.kantl.be/ctb/project/dalf/>