The *e-Laborate* Project and the Usability of Another Textual Paradigm

Joris van Zundert
(joris.van.zundert@niwi.knaw.nl)

Dept. Dutch Linguistics and Literary Studies

Karina van Dalen-Oskam
(karina.van.dalen@niwi.knaw.nl)

Dept. Dutch Linguistics and Literary Studies

In 2003 we embarked upon the project *e-Laborate: a digital platform for partnerships in the humanities and social sciences*. The web application (at <http://www.e-laborate.nl/>) resulting from this project is intended as a virtual workplace for researchers in the humanities and social sciences.The *e-Laborate* collaboratory contains text collections, collections of statistical data and basic content management tools for sharing and working on text material and datasets. The project allows individual researchers as well as research groups to explore the potential of the collaboratory and to generate feedback. The tools enable users to expand the collection of material continuously and to improve its quality. In our paper we will present *e-Laborate* as an on line research collaboratory and as a web enabled tool for editing and analysing textual content. We will also show how *e-Laborate* provided a research environment in which we can explore the usefulness and usability of a specific text paradigm.

The text material we used in our project issued from the historical cultural journal "Vaderlandsche Letteroefeningen". The title means "National Literary Exercises" and in academic writing is usually shorthanded as "VLO". Published between 1761 and 1876, the "VLO" is of great importance for every research discipline concerned with the study of culture in the Netherlands during that period. There has long been a widely held desire to see a complete set of the journal available in digital form. However, because of its huge size and the enormous costs that digitisation would entail this has not been possible before now. The approach we have chosen differs fundamentally from the way in which textual material has usually been digitised and published in the past. The "VLO" component of the *e-Laborate* project uses a bottom up collaborative approach, drawing upon the assistance of researchers, to produce a continuous developing and evolving digital version of the publication. Using this approach NIWI will now be able to publish facsimiles (scans) of the first 50.000 pages of "VLO" editions by the first quarter of 2005.

We will describe the development process used in building *e-Laborate*.The *eXtreme Programming* protocol (XP) was closely followed. Researchers' demands concerning the texts were closely monitored during the project and used to drive the development of the electronic tools for joint working on text and textual material. Every two weeks new elements were delivered, tested and approved of or commented on. Also critique and additional wishes were communicated with the developers. In this way we made sure that the tools would really be what researchers collaboratively working on text wanted and needed. The participating researchers are enthusiastic about this development approach and about the tools delivered up till now. Formal evaluation and retrospection showed especially appreciation for the pragmatically forward looking vision of the project (i.e. building the collaboratory brick by brick, feature by feature).

The paper will provide a functional and architectural overview of *e-Laborate* as a collaborative tool for supporting the production of digital editions. At the core of *e-Laborate* is the *transcription object*. The transcription object is a container object holding the scanned image of a page from an original publication and a transcription field. Each transcription object's authorisation may be tailored by its creator / owner. Depending on the user's authorisation the transcription field of a transcription object is depicted either as a text edit box or as rendered text. Arbitrary additional metadata may be added. In the case of the "VLO" a standard id field is added to hold the number of the year, volume and page the scan shows. Standard content management utilities available within the *e-Laborate* platform allow for the arbitrary placing and grouping of individual transcription objects into a page or folder hierarchy. Any transcription object is automatically indexed so an authorised user or editor can search through the text base and present the search results in a comprehensive way. A fuzzy matching algorithm amends search input as well as the indexed material for spelling variants. In the future tools to further process or statistically analyse those results may be added. The addition of modules, tools, or components to *e-Laborate* is easily facilitated by its plain plugin architecture and open source nature. Current additions under development are the inclusion of an open source OCR engine to facilitate text recognition on demand for uploaded scans.

Current work in the project is focused on the development of a flexible annotation tool. This tool will empower researchers to create annotations to every part of a scan or the transcription text of a transcription object, simply by pointing to and highlighting the image part or text range they desire to annotate. Researchers will also have the possibility to react to annotations by annotating the annotation (*ad infinitum*). A researcher may choose to categorize his or her annotation using a standard or personalised typology of annotations. Standard annotation typologies that will be provided concern a.o. basic formatting (italic, bold, capitalization etc.), ranges of interpretation (word, part of the text etc.) and information type (back ground historical information, biographical etc.). Any annotation may be categorised in multiple typologies.

The annotation tool will be as much WYSIWYG as possible. This means that a researcher wanting to add annotations will not be bothered by laborious tagging and will need no prior knowledge of any particular mark up language. This is a design choice fundamental to our view of text and textual research. We think that it's not a researcher's concern to produce or validate XML or any other marked up form of text. Knowing about mark up is not fundamental to the task of a text researcher, but making inferences about the meaning, structure and form of a text and putting such inferences into annotations is. Therefore tools for the production and enrichment of digital editions should focus on that research related task and not on mark up particulars. As a consequence the digital editing tools of *e-Laborate* will take care of the creation of valid mark up 'in the background', providing ample information about the name of the user who created the annotation, the date and time of creation, the part of the text or scan the annotation belongs to, and of course, the annotation text itself and any additional metadata provided by the user.

Elementary for our project are the leading principals behind the design choices described in the preceding paragraph. That is, the design choice not to define yet another mark up solution, but to concentrate on the researcher's interactions with the textual material, leaving the description of these interactions in the form of XML to the application. We will show that these principals define another textual paradigm, meaning *another* textual paradigm than the text paradigm implicitly emanating from the concepts of *TEI*.

At present a powerful surge of *TEI*-driven edition projects, seems to have propagated *TEI* into a de facto standard. Although undeniably useful as a means for marking up texts for editorial use, the apparent all round applicability and efficiency of *TEI* needs to be contested. We will argue that *TEI* in it's form of explicit mark up is not a very efficient means of editorial mark up. We will also argue that *TEI* is far from efficient nor very useful when computer supported textual analysis is the focus of research. We will show that the use of *TEI* forces an a priori, top down view of text onto a researcher trying to model a text using *TEI*-tagging. *TEI*'s particular use of XML and its DTD implicitly present a vision of a text being a flat hierarchy of meaningful text elements. To a researcher wanting to express and analyse overlapping interpretations, associative relations, layered narratives (to name but a few common textual constructs *TEI* has difficulty expressing) *TEI* does not provide effective or efficient solutions. We will argue that such a researcher would be better off considering the use of lightly embedded mark up solutions and layered cross tagged mark up as provided for example by the *JITT* and *LMNL* models. Although problematic in themselves, these models do address the non linear, non hierarchical nature of texts more adequately than *TEI*. We will also argue how these models can be combined to provide an intuitive way of structuring and annotating text, resulting in a dynamic layered model of text that can be represented by proper XML. We will show how within the context of *e-Laborate* a graphical user interface enables structuring and annotating texts according to this dynamic model of text representation. We are convinced that this interface enables a researcher to interact with a text on a research and interpretative level rather than a mark up level. We will also show that in such a dynamic research environment it is still possible to provide backward compatibility with *TEI* mark up using transformational languages.

Bibliography

Agosti, M., I. Ferro, I. Frommholz, and U. Thiel. "Annotations in Digital Libraries and Collaboratories." *Proceedings of the 8th European Conference, EDCL 2004. Bath, UK, September 12-17, 2004.* Ed. R. Heery and L. Lyon. Berlin: EDCL, 2004. 244- 255.

Buzzetti, D. "Digital Representation and the Text Model." *New literary History* 33 (2002): 61-88.

DARE, Digital Academic Repositories. Accessed 2004-11-20. <http://www.surf.nl/en/themas/index2.php?oid=7>

JITT. Accessed 2004-11-20. <http://www.sbl-site2.org/Extreme2002/> and <http://www.idealliance.org/papers/xml02/dx_xml02/index/title/e93017c13fc3874332dee40367.html>

LMNL. Accessed 2004-11-20. <http://lmnl.net/>

McGann, J.P. "Dialogue and interpretation at the interface of man and machine, reflections on textuality and a proposal for an experiment in machine reading." *Computers and the Humanities* 36 (2002): 95-107.

NHDA. Accessed 2004-11-20. <http://www.niwi.knaw.nl/en/geschiedenis/collecties/>

NIWI-KNAW. Accessed 2004-11-20. <http://www.niwi.knaw.nl>

SURF. Accessed 2004-11-20. <http://www.surf.nl/en/home/index.php>

TEI and TEI-Consortium. Accessed 2004-11-20. `<http://www.tei-c.org/>`

Women Writers. Accessed 2004-11-20. `<http://www.roquade.nl/womenwriters/>`

XML and the World Wide Web Consortium. Accessed 2004-11-20. `<http://www.w3c.org/XML>`

Xpast. Accessed 2004-11-20. `<http://www.e-laborate.nl/nl/new_2/toon>`

e-Laborate. Accessed 2004-11-20. `<http://www.e-laborate.nl/>`

van Dijk, S. "Introduction." *'I have heard about you'. Women's writing crossing borders.* Ed. S. van Dijk, P. Broomans, J.F. van der Meulen and W.R.D. van Oostrum. Hilversum: Verloren, 2004. [Information about the "VLO".]