Final Report on the NSF/DFG-funded Project "Middle High German Text Archive"

Electronic Text Center University of Virginia, Charlottesville, and University of Trier

DFG-Az.: LIS 1 – 554 941(1) Uv; Bewilligung vom 03.01.2001

1. *Participants:* Who has been involved?

What people have worked on the project?

Senior personnel at the University of Virginia funded by the NSF: Matthew Gibson; Jennifer E. McCarthy; Susan Munson; Cynthia F. Speer

Senior personnel at the University of Trier funded by the DFG: Ute Recker-Hamm M.A.; Dr. Rodrigo Readi-Nasser (from April to December 2001)

Graduate students at the University of Virginia funded by the NSF: Andrea Dickens, Lesley Pleasant, Jayme Schwartzberg

Graduate students at the University of Trier funded by the DFG: Niels Bohnert M.A., Claudia Bick, Cordula Michael, Shao-Ji Yao M.A.

What organizations have been involved as partners?

Electronic Text Center, University of Virginia (Charlottesville USA): University of Trier (Germany): Collaborative Research

The project team at the Electronic Text Center has been funded through the National Science Foundation (NSF).

The project team at the University of Trier has been funded through the Deutsche Forschungsgemeinschaft (DFG), the central public funding organization for academic research in Germany, and the University of Trier.

Have you had other collaborators or contacts?

Prof. Dr. Kurt Gärtner (PI), University of Trier; Dr. Thomas Burch (Center for Electronic Retrieval and Publication in the Humanities), University of Trier; Dr. Johannes Fournier (Center for Electronic Retrieval and Publication in the Humanities, through August 2003), University of Trier; Dr. Ralf Plate (Middle High German Dictionary, Project of the Academy of Sciences and Literature Mainz), University of Trier; Dr. Andrea Rapp, University of Trier, since January 2003 University of Göttingen

2. *Activities and Findings:* What have you done? What have you learned?

What were your major research and education activities?

The Etext Center and the German Department of the University of Trier have been collaborating to digitize a corpus of key Middle High German (MHG) texts, their related glossaries, and a series of Middle High German dictionaries containing references that, when used in conjunction with these primary texts, open wide avenues of study and comprehension to students of Middle High German. The corpus of MHG texts – the so-called *Findebuch*-corpus – consists of a great number of critical editions published after 1878; therefore lexicographically they are not covered by the dictionaries of Benecke/ Müller/ Zarncke (1854-1878) and Lexer (1872-1878). The *Findebuch* (1992) itself is a compilation of all the headwords listed in the glossaries of the corpus. Thus the digitization of both texts and glossaries and their interlinking on the one hand offers direct reading helps to each text of the *Findebuch*-corpus, and, on the other hand, through the interlinking of the *Findebuch* to the older dictionaries students get a maximum amount of help understanding the medieval German language. By combining structurally disparate XML-encoded (eXtensible Markup Language) resources – a large corpus of lexical resources and a another of primary text resources – "Digital Middle High German Interlinked" pushes the bar of past work in text archive creation. With complex textual structures and with complex requirements for how those texts should be presented and enriched with a maximum of reading helps, this project not only looks to bring different resources together, but also seeks to make them play and interlink with one another in an instructive, predictable, and relatively intuitive manner. The workflow and goals for the project have been:

1. Digitization of the texts and glossaries serving as sources for the *Findebuch*
2. Encoding of the texts and glossaries serving as sources for the *Findebuch*
3. Encoding of the *Findebuch* and dictionaries specifically for this project's purposes
4. Interlinking of texts, glossaries, *Findebuch* and the dictionaries
5. Development of the graphical user interface (GUI)
6. Project analysis of the technology and standards used (TEI, XML, Unicode, XSLT, etc.) and the relative benefits and drawbacks of each team's method of aggregating and delivering the text collection

What are your major findings from these activities?

1. **Digitization of the texts and glossaries serving as sources for the *Findebuch*:**

   The digitization work of the project has been out-sourced to an experienced Chinese keyboarding company, which is familiar with the digitization of medieval German texts and sophisticated scholarly editions. However, due to old and almost illegible editions, there were instances where letters (such as *ae*- and *oe*-ligatures) or combined diacritical marks were obfuscated by stains or foxing and could not be distinguished without intimate knowledge of MHG. These problems could only be tackled by checking each single instance manually, which is a time consuming and esoteric issue to say the least. Therefore the major finding of the digitization process is that the final quality assurance which is essential for scholarly use of an electronic text archive has to be accomplished by the content experts, i.e. the Trier team in this project.

2. **Encoding of the texts and glossaries serving as sources for the *Findebuch*:**

Encoding the German texts and glossaries has been a challenging process spanning the length of the project. When work was begun in Virginia, before the Trier team had finalized their digitization process, staff in Virginia investigated TEI-conformant XML markup that would most appropriately describe glossaries and dictionaries. During this time the Virginia team also wrote PERL/SED routines that would be necessary to convert the TUSTEP encoding in the files from Trier.

The texts arrived from Germany in two basic stages of markup: (1) texts fully encoded in TUSTEP, the electronic database program used for all the German source files; (2) texts lightly encoded in TUSTEP, which may be more robustly tagged after their conversion to XML.

The work of encoding the text files in XML was largely undertaken in the first phase of this project; however, tagging issues were being revisited until very recently due to the complex nature of the Middle High German texts' structure, the content, and the methods of display (goal #5). One problem encountered, for example, was the encoding of texts which contain multiple parallel versions in synoptical columns. For most effective scholarly use, encoding had to be developed which would allow the textual versions to appear side-by-side on the screen in addition to allowing the text to be searchable and interlinked with lexical resources such as glossaries, dictionaries and the *Findebuch* (goal #4). In dealing with texts where the variants had no convenient way to be aligned (due to different line numbering schemas of manuscript and edited text, manuscripts which were damaged and thus contained gaps in the text, etc.) and a coding system traditionally used to mark textual content but not on-screen rendering, the process of solving this problem was a dual process of XML encoding and PERL filtering to HTML.

Encoding the glossaries of the texts was also a problematic process, due to Etext staff's lack of Middle High German language expertise and the inconsistent representation of lemma entries in the glossaries. Potentially rich TEI-compliant XML encoding of the glossaries was first imagined, with the ability to mark not only lemma forms, but also grammatical information, definitions, senses, and links to citations of the lemmas in the texts. A number of difficulties were encountered:

- There is no consistent representation of different markup categories across texts – while grammatical information is recorded in italics in some texts, in others it may be rendered differently and italics in those contexts might mean something else.
- Whereas italics may be used to mark grammatical information in a text, not all text in italics is used to mark grammatical information. This posed two problems:

  (1) Conversion to XML cannot be done automatically;

(2) Without the ability to read Middle High German, Etext staff cannot make an educated distinction between what is grammatical and what is other information.

In order to address such problems, it was necessary for the Etext staff of XML experts to become more familiar with some frequently-used German grammatical terms, and to add steps to the quality assurance phase of this encoding to allow our partners in Trier – the content experts – to verify the encoding. In addition, the Trier and Etext groups decided that the encoding of different senses of a lemma, definitions, citational examples, etc. was problematic and much too time-consuming for this phase of the project; thus a different and intermediary encoding standard was adopted which allows for more flexibility and speedier workflow. One of the major findings of the project was thus the discovery that technological innovation cannot take place when it is completely separated from content expertise.

After the first round of file conversions from TUSTEP to XML, work was begun on the development of project XSL style sheets, which were designed to display our XML documents on the web.

3. Preparing the encoding of the *Findebuch* and dictionaries specifically for this project's purposes:

The University of Trier team successfully interlinked the lexical resources (the *Findebuch*, *Mittelhochdeutsches Wörterbuch*, and *Mittelhochdeutsches Handwörterbuch*) in a previous project funded by the DFG (http://www.MWV.uni-trier.de) with an array of anchors and pointers (IDs and IDREFs) leading from one dictionary lemma to another. Using the same data typing strategy, interlinking between the texts and their correspondent glossaries by line number was easily automated by the Etext partners. The primary question to resolve was how the text/glossary grouping would enter into conversation with the lexical resource grouping.

4. **Interlinking of texts, glossaries, *Findebuch* and the dictionaries:**

   Because there are no direct "hooks" (a systematic and predictable series of corresponding IDs and IDREFs) between the lexicons and the primary texts and glossaries – and because there is no method to automate this process due to the incredible range of variation for any given lemma – this project links the text and glossary lemmas to the *Findebuch* by running a full-text search against the *Findebuch* for its match (if that match indeed exists). Being a project described as an "XML Archive," project staff had originally planned to perform this interlinking and deliver the content with XSLT on the fly (see goal #2). However, primarily due to the need for greater speed and because the project had to deliver large texts or large text pieces at any one time, the team decided to accomplish its objectives with speedier PERL filters.

5. **Development of the graphical user interface (GUI):**

   Even before the first encoding pass had been finished for texts in the archive at Etext, the project partners needed to visualize a GUI design that would meet the archive's complex functionality requirements – this due primarily to the fact that, in the end, every textual piece must connect to another textual piece. In other words, a user should be able to link from a text to its glossary and go right back to the text from the glossary. A user should be able to link from the glossary to the *Findebuch* and then go from the *Findebuch* to every dictionary as well as back to a the glossary of the text. While the collaborators didn't want the screen to be consumed by an over-proliferation of browser windows, they also wanted to try and limit any erasure or obscuring of the scholar's path of discovery and comparison.

   This predicament raised two important issues, one inevitably practical and the other largely theoretical. First, how would a scholar and researcher use this electronic archive and what enhancements would digitization enable for that individual? Second, if XML theoretically describes structure and content divorced from its form, what might be said of this project specifically and of XML in general if, in this instance, the presentational form is on equal if not more important par than the content?

   As an answer to the first question, using the archive is an exercise of multidirectional discovery. Just as a scholar might sit at a physical desk surrounded by stacks of lexical and primary texts, the digital archive must maintain access to and availability of these materials in a single screen for comparative and discovery exercises dealing with, for instance, etymological or lexicographical research. The desire of the investigators of the project was to simulate what the scholar might do with the physical texts in front of him or her but also to utilize the digital efficiencies of searching and interlinking that the physical texts do not offer.

   Regarding the second and more theoretical issue – what might be said of the use of XML in the project if the end form dictated aspects of the encoding – the Middle High German project stands to argue that form and meaning cannot be separated. In many ways, encoding for structural meaning is useful in the practical world only if that encoding is done with an eye towards how the product will *look* and *function* – particularly when the information one is dealing with includes multiple volumes of dictionaries and primary texts where the potential for information overload is very acute.

6. **Project analysis of the technology and standards used (TEI, XML, Unicode, XSLT, etc.) and the relative benefits and drawbacks of each team's method of aggregating and delivering the text collection:**

While XML has been used throughout the collaboration, Etext and Trier have utilized very different open-source methods to both aggregate and disseminate the project. Likewise each group has learned a great deal about the pros and cons of both aggregation and delivery methodologies.

At first, both teams explored XSLT as a delivery mechanism for this dynamic transformation between native XML to web-viewable HTML. However, XSLT proved, at this point in its existence, to be much too slow for usability. Because the XSL transformation engines are Document Object Model (DOM)-based where the entire document must be held in memory to create a tree of nodes for the stylesheet to then process, and because the texts in the archive itself (in particular the dictionaries) are quite large, XSLT proved to be less effective than had first been imagined for this project.

Trier uses a MySQL relational database and pre-processes each text to place certain encoded parts of the texts in different relational tables and fields. To deliver the archive, Trier uses PHP. While there is more pre-processing for each single text before it is placed into the database, such pre-processing provides, out the back end, less processing time to deliver the primary content and deeper and more specific functionality within individual texts; in the Trier GUI, for instance, one can search for a line number or lemma form within the same window of a particular text without invoking other processes outside of that environment.

Etext, on the other hand, maintains texts in their fully-encoded document form. While some pre-processing is done to the texts to enable searching and consistency across the archive, Virginia relies more on event-driven post-processing to deliver dynamic web-content from the original XML source. Because of its rapid event-driven processing, Etext therefore decided to use a series of PERL and CGI routines that enable both more rapid delivery of the primary content as well as utilize the XML encoding to produce the complex dynamic links between the different texts. Thus, while the Trier interface and delivery gives the scholar tools to explore a particular text with depth, the Etext method gives the scholar tools to explore the entire archive with breadth. Within the Etext environment, one can, for instance, open Heinrich von Hesler's *Apokalypse*, read the text if he or she wants to, click on line numbers that invoke another process that interlinks with lemmas in the text's glossary and then either go to other lines in the text where that particular lemma occurs or go from the glossary lemma to the *Findebuch*. Once the scholar has been brought to the *Findebuch*, he or she can then go to the other dictionaries and back to the text again.

The limitations of using a PERL and event-driven system to deliver such a deeply-encoded archive are quickly apparent. For one, more encoding exceptions have to be accounted for in the code to allow for different combinations of XML elements that might be present in the archive. This results in much longer scripts and filters that are difficult to maintain and can easily break. Because of this, once DOM-based XSLT engines become faster and more robust, they would be an obvious choice to deliver the archive since processing would be based on nodal context in an object "tree" rather than based on the linear processing of any combination of elements whose hierarchical order is not necessarily important or considered in that processing. But given the two limitations – speed versus cleaner modularity and maintenance – Etext has chosen speed.

Finally, a major obstacle both teams have run up against is in the use of Unicode standards for which there are no fonts, as of yet, to display certain characters. One of the most problematic of these issues is for combined characters that have, for instance, a small "e" above a large "o". Because this character does not exist in Unicode as a single entity and because fonts that exist cannot incorporate them using "combined character" ranges, Etext has attempted to solve this with Cascading StyleSheet (CSS) workarounds where the smaller letter is superscripted and moved backwards a certain number of pixels to give the "appearance" of a single character entity. However, if a user has a certain font-size chosen to view the archive, this workaround becomes an apparent rough solution. Therefore, much of what we have attempted to do in this project has met the walls of what is technically possible in the present environment of the World Wide Web.

What opportunities for training and development has the project helped provide?

Due to the complexity of the textual material and the need for materials with inherently different structures to be aggregated into a single collection, Etext has done more exploration of XML in general and the Text Encoding Initiative (TEI) specifically to enable the valid encoding of such texts. In the collaboration, Etext and Trier have also learned a great deal about different aggregation and delivery methodologies and the pros and cons of these methodologies (see above #6).

Since our last annual report, project staff at Etext and at Trier were able to convene for a final meeting in Trier in January 2004. Such meetings between staff at Etext and Trier have taken place up to several times annually throughout the funding period and have been crucial to the success of the project, as they enable staff at Etext to better understand the programming needs posed by the German texts, and staff in Trier to better understand the difficulties and occasional limitations their texts and needs impose on XML encoding.

What outreach activities have you undertaken?

Matthew Gibson, project PI in Virginia, and Ute Recker-Hamm, project manager in Trier, made four presentations of the research questions and accomplishments of Middle High German Interlinked: the first to an audience of computing scholars at the University of Trier in December 2002; the second at the Joint International Conference of the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH) in May 2003; the third to the University of Virginia community in November 2003; and the fourth at the University of Trier in January 2004.

3. *Products:* What has the project produced?

What have you published as a result of this work?

Major Journal Publications

Thomas Burch und Kurt Gärtner: Arbeiten des Kompetenzzentrums „Neue Publikationsformen und Erschließungsverfahren für geisteswissenschaftliche Grundlagenwerke“ an der Universität Trier, in: *Akademie der Wissenschaften und der Literatur Mainz. Jahrbuch 2002, 52. Jahrgang.* Stuttgart 2002, S. 258-268, here 261-263. (Reports about the project also in the *Jahrbuch der Akademie 2000 and 2001*).

Kurt Gärtner: Comprehensive Digital Text Archives: A Digital Middle High German Text Archive and its Perspectives, in: *First EU/NSF Digital Libraries All Projects Meeting, Rome March 25-26, 2002.*

- http://delos-noe.iei.pi.cnr.it/activities/internationalforum/All-Projects/RomeSlides/DTArchives.pdf

Ute Recker: Digital Middle High German Text Archive. In: *Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001.* Hrsg. von Thomas Burch, Johannes Fournier, Kurt Gärtner und Andrea Rapp (Akademie der Wissenschaften und der Literatur Mainz. Abhandlungen der Geistes- und Sozialwissenschaftlichen Klasse; Einzelveröffentlichung Nr. 9). Stuttgart 2002, p. 307-309.

Books and other one-time publications

Annual reports concerning the projects of the Trier University Center for Electronic Retrieval and Publication in the Humanities:

- http://germa83.uni-trier.de/KoZe/docs/ab.pdf (pp.19f.)
- http://germa83.uni-trier.de/KoZe/docs/ab2002.pdf (pp.14-16)

What Web site(s) or other Internet site(s) reflect this project?

The Middle High German Interlinked project resides at
http://etext.lib.virginia.edu/german/mhg/browse/ .

The Trier Middle High German web interface is at http://mhgta.uni-trier.de/Demo/ .

What other specific products (databases, physical collections, educational aids, software, instruments, or the like) have you developed?

The project consists of a database and collection of 4 large lexical resources that Etext has converted to XML from their original SGML encoding and presently well over 40 full primary texts and their ancillary glossaries.

4. *Contributions:* How has the project contributed…

…to the development of the principal discipline(s) of the project?

Middle High German Interlinked is a study in using XML to bring together different types of digital content. The project will serve as a searchable independent resource for answering research questions posed by historians of language and literature, especially those investigating the study of Middle High German.

…to other disciplines of science or engineering?

The project's investigation and use of XML and other open source technologies such as PERL, PHP and XSLT to aggregate information of disparate structures provides a potential model and solution for other projects engaged in questions of knowledge management and knowledge dissemination. The complexity of managing the information in this particular project, however, poses further questions particularly in regards to issues of the user's cognitive experience with the presentation of digital information.

As stated before, one of the primary aims of this project was both to replicate the scholar's experience of sitting at a desk with access to any number of physical texts, glossaries, and dictionaries before her or him and to enhance that experience with tools that the promises of digitization might offer. The architects of the project were faced with a daunting problem: while an over-proliferation of information on the screen becomes confusing and might actually impede research, maintaining a quickly recoverable if not simultaneously visible history of the scholar's path of discovery and comparison was imperative. The project collaborators have learned that the rapid proliferation of information in different windows has very concrete repercussions upon the user's ability to digest the information on the screen and, perhaps of greater importance, upon the user's directional orientation. There were and continue to be many moments in the project when, in using the archive, a user has lost track of exactly where they have gone; are they in a primary text? A glossary? A dictionary? And if so, which one? One of the main questions the collaborators would like to investigate is the processor of the human being as he or she is faced with more and more options and directions in a world of quick-click information and hyper-active interlinking. Context is a big key here… and how is context both defined and maintained in this hyperactive world of digital media? And as the onus of aggregating, delivering, and organizing digital information is, more and more, being undertaken by libraries, what types of investigations on usability have actually occurred to know that we are heading in the right direction?

As libraries become more and more "digital" in order to make information retrieval for the user more seamless and accessible, are they potentially going to encumber the user, holding her or him more responsible to bring tools of complex analysis to the information itself? How does one translate – or perhaps transcend – the experience and present model of print research to the web environment? While the computer processor may expedite searching, the discovery of relationships, and quantitative analysis, how is the human processor ingesting and responding to this digital experience and, primarily based on the smaller real-estate of the computer screen, how might design and interface function to solve, or perhaps ameliorate, those issues?

… to physical, institutional, and information resources that form the infrastructure for research and education?

Collaboration between the University of Virginia and the University of Trier continues to explore the possibilities and advantages offered by the training of an international digital library staff devoted to the development of a common text archive.

…to the public welfare beyond science and engineering? (commercial technology, the economy, solutions to social problems)

N/A.