Journal of the Text Encoding Initiative

Issue 2 (February 2012)
Selected Papers from the 2010 TEI Conference

Gerhard Budin, Heinrich Kabas and Karlheinz Mörth

Towards Finer Granularity in Metadata

Analysing the Contents of Digitised Periodicals

Electronic reference
Gerhard Budin, Heinrich Kabas and Karlheinz Mörth, « Towards Finer Granularity in Metadata », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 19 January 2015. URL : http://jtei.revues.org/416 ; DOI : 10.4000/jtei.416

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/416
Document automatically generated on 19 January 2015.

Gerhard Budin, Heinrich Kabas and Karlheinz Mörth

Towards Finer Granularity in Metadata

Analysing the Contents of Digitised Periodicals

1. Background

1 The ICLTT Metadata Initiative is an experimental project aiming to uncover hitherto undocumented parts of the Austrian Academy Corpus (AAC). This corpus, which has its roots in a number of smaller projects that originated in the 1990s and eventually grew into something like a full-blown national corpus project, was, unlike most corpus projects, driven by scholars of literature and history, not by linguists.

2 The vast bulk of the corpus texts dates from the first half of the twentieth century. Currently, the corpus consists of about 500 million tokens, and most of the texts contained in the collections do not strictly belong to the sphere of what traditionally would be described as *belles lettres*. As the texts were collected according to both a literary and a lexicographic perspective, the corpus also contains a considerable number of functional and informational texts. Roughly half of the data is made up of periodicals, not large-size daily newspapers but rather medium- and small-size weekly and monthly publications. There are many collective publications such as yearbooks, readers, commemorative volumes, almanacs, and anthologies covering a wide range of writers, topics, types of texts, and genres. While at a first glance the collection might appear heterogeneous, it actually represents a unique collection of historical German texts, many of which cannot be found elsewhere in digital form.

3 The texts that make up the collection have undergone various and diverging steps of manual and automatic annotation. This is due to the fact that the corpus is a compilation of data from various projects. In spite of this diversity in both content and form, the corpus is held together by a number of common principles: among these is, most importantly, the integrity of the digital objects. The units to be digitised were never single articles from a magazine or a journal or parts of novels but always the complete physical item.

4 Technically speaking, the corpus is made up of XML encoded texts, and the schemas in use are more TEI-inspired than TEI-conformant. The markup applied to these texts is more format-oriented than that which advocates of descriptive markup would usually approve. The rationale behind this was the belief that a modicum of presentational markup is indispensible to conserve the fundamental semantics of historical texts and that mapping formal phenomena of texts onto semantic categories often requires considerable additional expenditure of time and resources (Mörth 2002). However, a major effort to convert the whole corpus to TEI-conformant structures is currently under way and expected to be completed by the end of 2011.

5 The format-oriented attitude of the scholars working on the corpus working group motivated them to choose to preserve images of all digitised source documents (even though this greatly increased the size of the data). Moreover, all digital objects in the corpus have been organised on a one-page-one-document basis, which was motivated largely by considerations concerning the various workflows.

6 The metadata policy of the group of involved researchers has been based on TEI headers since the very beginning. Each digital object, representing a physical item such as a book or a bound volume, was provided with a TEI header. Editing of this data was done by means of a relational database: the existing corpus infrastructure provides tools to create well-formed TEI headers that can be exported and added to the relevant documents. The ICLTT also contributes to the *European Demo Case* experiments of the CLARIN project (Work Package 2, task 7), which is closely related to the CLARIN Metadata Initiative. One of the goals in this endeavour is the integration of TEI conformant metadata into the overall scheme.

7 Given the fact that more than half of the collection is made up of periodicals and other collective publications which only provides metadata at the level of the physical item, the percentage of identifiable texts is very low. Our knowledge of writers, genres, topics, etc. is restricted to those parts of the corpus where metadata could be attributed to the highest-level

digital object (the above-mentioned book and volume levels). To put it another way: the corpus contains a comparatively large amount of textual data and a comparatively scarce amount of metadata. Even though we are in possession of a unique and very large collection of texts, even our specialists in literary and history studies have only a very fuzzy picture of what this treasure trove really contains. At present, we are unable to provide our users with comprehensive lists of authors whose works appear in the journals, and we are unable to compare the oeuvre of any one writer to another within the corpus.

2. A Very Short Wish List

8 At the outset of the metadata project, we wished to create both (a) more fine-grained metadata that would provide detailed information about the contents of our holdings and (b) a tool for creating this metadata.

9 While the usual protocol in digital document management is to have one metadata record assigned to one document, we were aiming for metadata describing parts of one or more documents.[1] Furthermore, we wanted to create TEI P5 headers in order to ensure seamless integration with existing workflows and the corpus infrastructure.

3. The Wider Landscape

10 Initiatives for streamlining metadata creation are legion, not only in the library community but in all areas where digital data is being created. It is well known that a number of stages of the digitisation process can be automated with often quite satisfactory results, yet producing quality metadata remains an economic issue since human intervention often cannot be avoided. Publishers, libraries, and specialised service providers like OCLC maintain millions of metadata records and create new ones at an ever increasing speed, yet it seems that adding value and keeping costs down do not go hand in hand.

11 There are a number of software applications with functionality similar (in part) to what has been attempted in the project under discussion. Most of these are located in the sphere of large-scale digitisation projects, usually based at national libraries. Tools such as *docWORKS[e]* (by Content Conversion Specialists in Hamburg) and *C-3* (by ImageWare Components in Bonn) are often incorporated into large workflow engines. There are also free products such as *Goobi*, developed by the Center for Retrospective Digitization, Göttingen (GDZ), which incorporates the metadata creation tool *RusDML*, and the *Archivists' Toolkit*. Many steps of the digitisation process can be automated, and some projects even aim at advanced features such as automatic metadata creation (for example, Basic Technologies, which is part of the German *THESEUS* project).

12 However, for a number of reasons, fixed workflow chains designed for the wholesale digitisation of large libraries do not appear to offer practicable solutions for smaller projects or for projects building on existing textual data. In our experience, the main issue in putting existing software solutions to work often relates to modularity: software is available only as a package and only functions well as long as users adhere to default procedures. When deviating from the path of predefined workflow steps, handling often gets tricky.

13 In short, our search for an alternative was largely motivated not only by the limited support for modularisation of existing tools but also the considerable set-up overhead and the need for an interface suitable for the structure of the corpus at hand. While libraries often work towards standards such as METS, MODS, and ALTO, our standard is the TEI header, which is an integral component of the ICLTT's corpus infrastructure.

4. ICLTT Metadata Initiative

14 The ICLTT Metadata Initiative was launched in early 2010, when we started to selectively create metadata for a medium-sized (<100 million tokens) collection of digital periodicals. This project was designed to generate a more detailed view of the contents contained in the AAC by selectively describing lower-level digital objects that did not already have metadata.

15 The project was not intended to exhaustively assign such metadata to all the journals and magazines contained in corpus. The goal of these experiments was rather to create structures

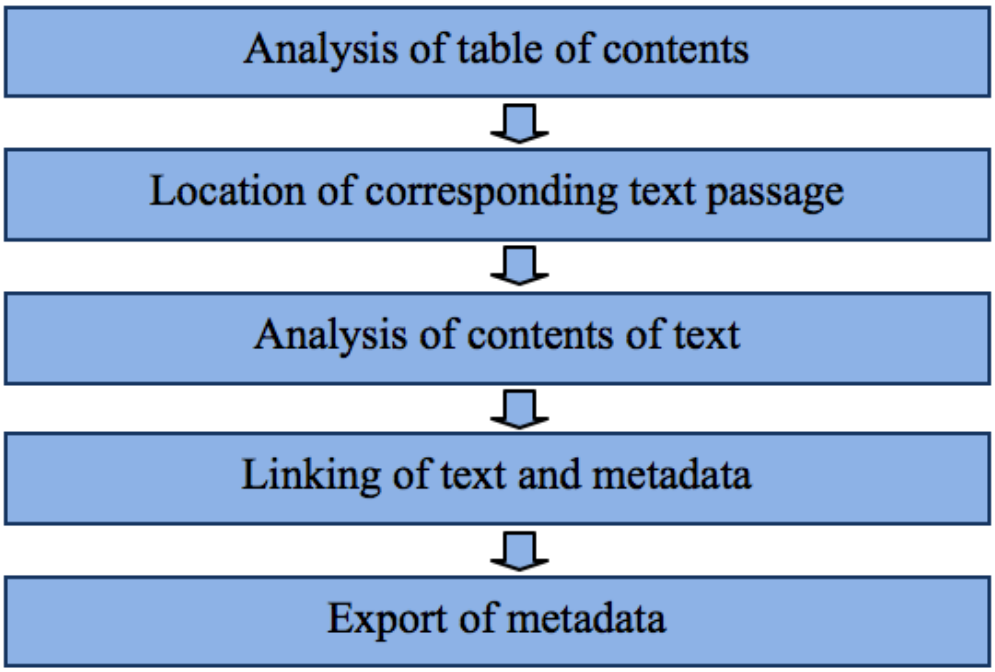
that allow the deliberate extraction of topically organised sub-corpora and to define a workable workflow for adding metadata in the future.

4.1. Tools

16 We endeavoured not to invent anything from scratch and decided at an early stage to make use of the ICLTT's standard encoding tool. This tool, which is called *corpedUni* ("corpus editor + Unicode"), was chosen because it is well adjusted to the local corpus infrastructure and has a number of reusable functionalities that could be used for this project. *CorpedUni* is an XML editor that allows the user to work with a large number of documents and to define workflows. In employing this tool, which has been in continuous development at the ICLTT (and its predecessor the Austrian Academy Corpus) for more than a decade, we could keep overhead comparatively low.

4.2. The Workflow

The workflow for creation of metadata for lower-level digital objects can be described as a five step process:



4.3. Analysing *Tables of Contents*

17 The basic input for the metadata creation process was gathered from various sources. In earlier experiments, metadata creators or editors simply worked their way through the digital texts, jumping from one title to the next to create digital tables of contents. This approach worked in some cases but not in others. Since most of the corpus data in this project was produced automatically without manual intervention, titles of particular texts (an article in a journal, for instance) were not encoded consistently. Even for the texts that had been manually encoded, the decision as to which titles were relevant and which were not made creating a digital table of contents rather cumbersome.

18 In more recent experiments, we tried to approach the problem from a completely different angle. Instead of analyzing huge amounts of text by sifting through all the pages and extracting relevant titles, we tried to exploit the table of contents found in the print versions of the digitised issues. Not all of them had such navigational aids, but most of the journals and magazines contained either a table of contents or a comparable form of access structures, and sometimes these were available in secondary sources. Refactoring this data in advance proved to be good way of both speeding up the process and improving data quality.

19 As a first step, the human editor analysed the structure of the table of contents to identify the author's name, the title of article, and the location within the issue. In most cases, location is

simply represented by a page number, but it can also be a combination of page number and issue number. Though this task may appear simple, a closer look will reveal that it is a process that has a great number of stumbling blocks. It is not only the order of the basic constituents of items of a table of contents that displays a considerable degree of variation, but also the separators between them differ and the number of constituents is not fixed.

20 In this phase, the human editor first has to determine the overall structure of the table of contents at hand. CorpedUni offers the user a list of genres and can even suggest a particular genre. In the end, the human editor has to choose the appropriate one and to determine the basic structure of the items: i.e. the sequence of constituents and type of separators that stand between these constituents. At this point, one might expect the program to be able to proceed automatically. Experience from practical work on the contents items has shown that the number of possible deviations from what could be expected is very large—everything from printing errors in the source documents to OCR errors. In the end, all relevant data are transformed into a simple two-dimensional structure: a data grid which serves as a starting point for the next step.

4.4. Finding and Isolating the Text Passages

21 For basic cataloguing purposes, a digital table of contents could be the endpoint of the whole process. However, to allow NLP software to do a little bit more with the data, the process needs to be taken a step further: identifying the beginning and end of the relevant text passages. To achieve this, the software first attempts to find the titles of each article in the journal issue (which, as previously mentioned, were rarely tagged correctly to begin with). Having identified these titles, the software can then determine the beginning and end of each article.

22 The task of finding the titles is achieved by means of simple string-matching algorithms. The results depend to a great extent on the quality of the digital text: although the system is constantly being improved, it still produces plenty of errors. After the automatic insertion of initial and final division tags, a senior encoder checks the plausibility of the automatic tagging and applies manual corrections wherever necessary. The experiments have shown that, in spite of some errors, most of the text passages can be identified in a first run. Figures at this stage would be hardly conclusive as the results for historical material hinges on a number of factors that are difficult to assess.

23 Having inserted the division elements, the software creates pointers from the digital table of contents to the beginning and the end of the texts and assigns unique identifiers to the respective titles. The next step is the automatic application of unique identifiers and links that ensure navigability within the corpus. These identifiers are ultimately attached to idno elements in the TEI header.

4.5. Manual Override: Categorizing Contents

24 While corpus projects working on similar sources often settle for very general labels (such as *journalistic prose*) for type of text or domain, this project requires a more specific taxonomy. In working towards this end, we have been motivated by a vision of smaller corpora tailored to particular needs: sub-corpora that can be created from the overall collection at the push of a button.

25 When we started the project, we were looking for a system onto which the avowedly particular needs of the scholars in our working group could be easily mapped. Having eschewed the idea of creating anything from scratch, we decided to turn towards existing descriptive thesauri. We were in need of a scheme that would not only identify types of texts, but also provide a sophisticated taxonomy for our literary projects ready to be applied.

26 As UDC (*Universal Decimal Classification*) appears to have lost ground, we had a closer look at other projects such as *Library of Congress Subject Headers*, the German *Schlagwortnormdatei* (SWD) and *Dewey Decimal Classification* (DDC). In the end, DDC (version 22, German) was chosen for our current trial runs. There were a number of arguments in favour of DDC: it covers much–though not all–of what was needed to classify the texts and contents at hand, it has been translated into over 30 languages and can thus be easily mapped

from one language onto another, it has an ever-growing international community and there are a number of projects working on DDC interfaces with other systems (such as CrissCross[2]).

27 Our first experiments applied a dual system: at the core of the classification was DDC22. Wherever necessary, additional keywords for which there is no appropriate DDC22 class have been added, which are linked in the TEI header to a separate *aacmetakeys* scheme.

28 Another important issue is assessing intellectual property rights, which is carried out at this stage. Having identified the author of a text, the next step is to determine biographical data in order to know whether the text is still protected by copyright. In the medium term, the department is planning an interface between the editing application and a prosopographic database to serve the purpose of finding and validating biographical data.

4.6. Further Automatic Steps: Linking, Tag Usage, Word Count

29 The *corpedUni* software can do automatic analysis of tag usage and perform statistical analysis such as word counts. If the data has been furnished with POS or lemma data, this can also be analysed and added to the metadata records. A host of further procedures is conceivable.

30 Given the fields of interest of the ICLTT, the resulting portions of text chunks will offer an ideal ground for experiments in text categorisation and term extraction, the results of which could lead to more efficient methods of metadata creation, especially in those parts of the process that are currently being performed by human editors.

4.7. Exporting the Data

31 The last step in the workflow is the creation of TEI headers. The AAC's metadata has been organised in a twofold manner: production data is stored in separate documents alongside the digital objects in the file system, whereas descriptions of the digitised sources have been kept in a relational database, the fields of which correlate with TEI header elements. The database has primarily served the purpose of easing the process of editing and validating the input. TEI headers created for the top-level digital objects have not been edited manually. Nevertheless, more recent *corpedUni* versions include a simple built-in editor for TEI headers, which allows viewing and editing such data in a more comfortable way than what general XML editors offer.

32 During the final export process, all metadata records created in *corpedUni's* metadata tool are saved as TEI headers. Actually, two datasets are merged here: part of the data comes from the hierarchically superior TEI header and the other part comes from the data stored in the data grid described above. All this data is integrated into one TEI header which is saved into a separate document. The routine simply fills in a template.

33 The headers produced are to the best of our knowledge conformant with TEI P5. There is only one customisation to be mentioned here: the routine adds an icltt:filename element which allows for straightforward linking with particular files in the corpus. A second equally important reason for this unorthodox way of linking to an outside file is the fact that the indexing software we have been using, Dialing/DWDS-Concordancer

34 DDC (Dialing/DWDS-Concordancer), a search engine optimised for linguistic purposes. It is published under a GNU licence (http://sourceforge.net/projects/ddc-concordance/).

35 See, for example, all the partners of the *Korpus C4* project (http://www.dwds.ch/index.php?option=com_content&task=view&id=36&Itemid=66) or the CroCo project (Klinger, Vela et al. 2006).

5. Sharing Our Tools

36 The workflow described in this paper has been tested for almost a year now. Needless to say, the quantities of data being produced as part of this experimental project are not comparable in any way to what is being produced in wholesale digitisation projects at large libraries. So far, we have created 8,000 metadata records. However, our first projections suggest that the creation of about 100 records per day and editor is a realistic assumption.

37 The tools used in this project are all freely available via the ICLTT Web site. The most recent version of *corpedUni* can be accessed at http://corpus3.aac.ac.at/showcase and can be freely used for all non-commercial purposes. A more detailed description of the usage of the software in the form of a tutorial is currently being created. We plan to package this with some trial data

that allows other scholars to work along similar lines. All of this should be made available in 2011.

Bibliography

Betram, Jutta. 2005. *Einführung in die inhaltliche Erschließung: Grundlagen – Methoden – Instrumente*. Würzburg: Ergon.

Burnard, Lou, and Syd Bauman, eds. 2010. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlotteville, Nancy. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Klinger, Annette, Mihaela Vela, and Silvia Hansen-Schirra. 2006. "Kodierung von Metainformation". http://fr46.uni-saarland.de/croco/corpus_meta.pdf.

Knull-Schlomann, Kristina, ed. 2008. *New Perspectives on Subject Indexing and Classification. Essays in Honour of Magda Heiner-Freiling.* Leipzig, Frankfurt am Main, Berlin: Deutsche Nationalbibliothek.

Mörth, Karlheinz. 2002. "The representation of literary texts by means of XML: some experiences of doing markup in historical magazines." In *Digital Evidence. Selected papers from DRH 2000, Digital Resources for the Humanities Conference*, edited by Michael Fraser, Nigel Williamson and Marilyn Deegan, 17-32. London: Office for Humanities Communication.

Reiner, Ulrike. 2010. "Automatische DDC-Klassifizierung bibliografischer Titeldatensätze der Deutschen Nationalbibliografie". *Dialog mit Bibliotheken* 22 (1): 23-29.

Wang, Jun. 2009. "An Extensive Study on Automated Dewey Decimal Classification". In *Journal of the American Society for Information Science and Technology* 60/(11): 2269–2286.

Notes

1 It is interesting to see that large libraries often pursue a policy of excluding smaller, heterogeneous objects or works of 'elusive' literary character altogether from being assigned detailed metadata in their digitisation efforts (http://www.d-nb.de/wir/pdf/nichterschliessen_gesamt.pdf).
2 This is a project mapping the headings of the *Schlagwortnormdatei* (SWD) to classes in the German version of DDC (http://linux2.fbi.fh-koeln.de/crisscross/index_en.html).

Cite this article

Electronic reference

Gerhard Budin, Heinrich Kabas and Karlheinz Mörth, « Towards Finer Granularity in Metadata », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 19 January 2015. URL : http://jtei.revues.org/416 ; DOI : 10.4000/jtei.416

Authors

Gerhard Budin
Full professor at the University of Vienna, chair of terminology studies and translation technologies, faculty dean of the Centre of Translation Studies at the University of Vienna and director of the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences.
Heinrich Kabas
Senior researcher at the Institute for Corpus Linguistics and Text Technology (Austrian Academy of Sciences).
Karlheinz Mörth
Senior researcher and project leader at the Institute for Corpus Linguistics and Text Technology (Austrian Academy of Sciences) and lecturer at the University of Vienna.

Copyright

Abstract

In early 2010, the Austrian Academy of Sciences' ICLTT instituted an experiment in selective metadata creation for a medium-sized collection (<100 million tokens) of digitised periodicals. The project has two main objectives: (a) assigning basic structures to previously digitised texts, so-called *divisions* in TEI nomenclature, thus creating a set of new digital objects, and (b) the subsequent categorisation of these texts with the purpose of being able to create thematically organised sub-corpora. An additional objective was to have metadata stored as TEI headers.
Attempts at streamlining metadata creation are legion, in particular in the library community. Tools to do the job are often incorporated into workflow engines which consist of commercial products (such as *docWORKS[e]* and *C-3*) as well as free products such as *Goobi*, which incorporates the metadata creation tool *RusDML*, and the *Archivists' Toolkit*™.
The experimental workflow being tested at the ICLTT is an attempt to capture detailed metadata for a comparatively large collection of digitised periodicals and other collective publications such as yearbooks, readers, commemorative publications, almanacs, and anthologies. While all higher-level digital objects in the corpus were furnished with metadata from the beginning of the digitisation process, the current experiment is designed to enrich this data to more fully describe the contents of the material at hand. To achieve this end, the department's standard tools were adapted, which had the added benefit of keeping software production costs at a minimum.
While in earlier experiments of our group of researchers (metadata creators) created the TEI header for each text division manually, we have been trying to approach the problem by exploiting the contents section of the digitised issues and/or other secondary sources, which has resulted in a tangible acceleration of the process. Together with collecting basic data such as *author*, *title*, *publication date*, and *creation date*, the project classifies each division with a type of texts and topics, the latter using the standard *Dewey Decimal Classification* (version 22, German) with supplementary keywords.
This paper discusses a number of issues concerning the quality and type of resulting data. It also touches upon the issue of automation and at what points in the process human intervention is indispensible. Particular attention is directed at the software module for creating TEI headers.

Index terms

Keywords : corpus annotation, metadata, metadata creation, TEI headers, tools