Originally published in: Tomasi, Francesca/Vitali, Fabio (Eds.): Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities (DH-Case 2013), September 10, 2013, Florence, Italy. - New York: ACM, 2013.
https://doi.org/10.1145/2517978.2517988

Extending the possibilities for collaborative work with TEI/XML through the usage of a wiki system

Bastian Entrup
Justus-Liebig-Universität Gießen
Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D
35394 Giessen, Germany
bastian.entrup@germanistik.uni-giessen.de

Frank Binder
Justus-Liebig-Universität Gießen
Center for Media and Interactivity
Ludwigstrasse 34
35394 Giessen, Germany
frank.binder@zmi.uni-giessen.de

Henning Lobin
Justus-Liebig-Universität Gießen
Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D
35394 Giessen, Germany
henning.lobin@uni-giessen.de

ABSTRACT

This paper presents and discusses an integrated project-specific working environment for editing TEI/XML-files and linking entities of interest to a dedicated wiki system. This working environment has been specifically tailored to the workflow in our interdisciplinary digital humanities project GeoBib. It addresses some challenges that arose while working with person-related data and geographical references in a growing collection of TEI/XML-files. While our current solution provides some essential benefits, we also discuss several critical issues and challenges that remain.

Categories and Subject Descriptors

H.5.3 [**Computer-supported cooperative work**]; H.4.1 [**Workflow management**]; I.7.1 [**Document management**]

1. INTRODUCTION

The GeoBib project[1] is creating an annotated and georeferenced online-bibliography of the early German and Polish Holocaust and concentration camp literature [3]. Unfortunately, most of the early texts on the Holocaust published between 1933 and 1949 were soon forgotten or suppressed [5]. GeoBib will provide an innovative research platform, comprehensively covering the domain of early Holocaust literature with a bilingual scope on German and Polish texts. The resulting online bibliography will be based on annotation documents that contain detailed information and metadata for these Holocaust texts, but will not include the whole texts themselves. The annotation documents - a collection of TEI/XML-files - contain short abstracts, keywords, biographical information on the authors, reviews, information on persons, places, and time periods. The project's goal is to make these resources publicly accessible and searchable for a broad and interdisciplinary audience of researchers and the interested public.

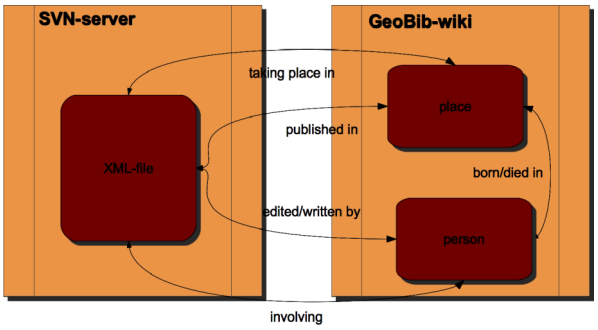

Figure 1: Data storage in the GeoBib project.

2. TWO PROBLEMS IN COLLABORATIVE EDITING

A classical approach to describing and adding annotations to texts is the usage of TEI/XML. TEI offers a broad variety of elements to precisely describe properties of texts. Within the project we use an adaptation of TEI fitting our requirements. Since we do not annotate the whole texts, but rather collect meta information, we do not use the `<text>` element at all. Hence, our schema makes all information available in the `teiHeader`.

The number of XML-documents to be created is estimated at around one thousand, each XML-file representing one early Holocaust text. Some of these include up to a few hundred persons mentioned by name, and can include just as many references to locations.

Our first attempt, the sole usage of TEI to store information on the text as well as information on persons and places, lead to two closely connected problems: data inconsistency and redundancy.

[1]Official project title: *Early Holocaust and concentration camp literature in German and Polish language (1933-1949) - an annotated and geo-referenced online bibliography for the research on narratives of remembrance*. GeoBib is funded by the German Federal Ministry of Education and Research (FKZ: 01UG1238A-B).

The problem of redundancy exists on two different levels: data storage and data acquisition. For the time consuming task of data acquisition, i.e. researching biographical information on persons to make them recognizable over different texts, redundant effort must be avoided. This aspect is especially important for collaborative work where many people are working with the same entities. Secondly, information must not be stored redundantly, since that would lead to possible inconsistencies: How could information on entities be managed efficiently, when this information is spread over a number of documents? Some kind of data base or data storage is needed that meets the following requirements:

- Each entity, i.e. each person and each location, needs a fixed ID/URI that can be used over all annotation documents.
- Collecting biographical information should be possible in a way that combines both running text and a structured form that can be filled with information such as birth dates and other structured data as well as spelling variants[2]. The annotator must be able to find an entity when searching for one of these variants.
- The system has to support the collaborative workflow of the project.
- The last but essential requirement, is the usability for the ordinary computer user.

While the first three requirements are easily met by a database system, its usability for the classical researcher is limited. Experience with other large data collections based on collaborative work, most prominently the Wikipedia[3], shows that there are systems that can easily be learned and used without much prior knowledge or training.
We chose to use a MediaWiki system[4] which comes with some disadvantages on the information processing side[5], but offers an intuitive user-interface. Our experiences in the GeoBib project is that the wiki system was easily understood and enthusiastically picked up, whereas working with the less user-friendly XML-files was disregarded by the classical humanist.
We approach the problems of data inconsistency and redundancy of data and work as follows: By linking entities in the annotation documents to pages in our wiki, we can make sure that entities, even though referenced by different text and different colleagues, are identified using the same URL, i.e. ID. This separation of concerns is shown in Fig 1.[6]

[2]The spelling of names and locations in these texts is not normalized. The names are often written based on hearing. Authors from different linguistic backgrounds, e.g. Polish or German native speakers as well as Jiddish speakers, write names of different language areas, resulting in a wide variety of orthographic variants for both names and places. A simple string lookup or matching would not suffice here.
[3]http://www.wikipedia.org/
[4]http://www.mediawiki.org/
[5]The information in the wiki is not as strictly structured as it would be in an XML file. Still, using templates we can ensure that the information will be structured and can be processed automatically.
[6]The information currently collected in the wiki will later be added and combined with information from other sources in one database and thus be accessible through the planned web platform. The current wiki itself is only for project internal use and not publicly available.

```
<particDesc>
    <listPerson>
        <person xml:id="FilipFriedman" role="author">
            <ref target="http://wiki.geobib.info/index.php/Filip_Friedman">Filip Friedman </ref>
            <note>Autor des Vorwortes </note>
        </person>
        <person xml:id="GerszonTaffet" role="author">
            <ref target="http://wiki.geobib.info/index.php/Gerszon_Taffet">Gerszon Taffet </ref>
            <note>Autor der Einführung </note>
        </person>
        <person xml:id="WaltervonBrauchitsch" role="undef">
            <ref target="http://wiki.geobib.info/index.php/Walter_von_Brauchitsch">Walter  von
Brauchitsch</ref>
            <note>Generalfeldmarschall</note>
        </person>
```

Figure 2: Referencing the wiki entries from within XML.

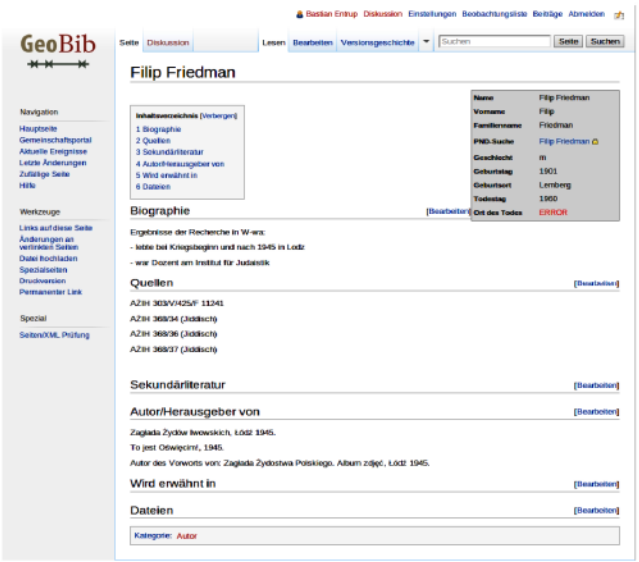

Figure 3: Example entry in the wiki.

Our approach allows a well-structured collaborative workflow on the data collection and facilitates data acquisition, storage, and reuse. Furthermore, the created XML-files are stored and interchanged using `subversion` (SVN), which allows the users to keep track of changes made to the files they work on, while at the same time allows interchanging the files between the members of the team. Especially when it comes to making corrections to the TEI-schema files or the oXygen[7] Author Mode (GUI), this automatic file exchange is very practical: new versions are automatically installed on the users system without any need to intervene or make adjustments.
Related problems of annotating entities from texts arise in other projects within the Digital Humanities as well. In [1] a similar problem of annotating persons in literary texts is described. Their solution is to use an XML-file to collect information on persons and share this file among all participants. New entries can be added from within oXygen. The XML-files are saved in an eXist database[8] and thus collaboratively available. We chose a different solution for our repository. For one, an XML file containing biographical information for a few hundred entries and in total containing

[7]http://www.oxygenxml.com/
[8]http://exist-db.org/exist/apps/homepage/index.html

a few thousand entries[9], can get very confusing. From our experience in other projects, the usage of eXist is not applicable when it comes to large singular files or a collection of many files. Response times increase dramatically and it seems virtually impossible to set up the system correctly.
Directly using an existing repository, such as the *Name Authority File* (*Personennamendatei* or short *PND*) [9], was no option, since only a very small percentage of the necessary entries already exist in the PND. Nonetheless, the GeoBib project uses templates in the wiki to link entities to the PND where possible. Furthermore, the project aims at contributing to the PND, e.g. by reporting synonymous entries and by adding missing entries.[10]

3. COLLABORATIVE DATA EDITING: XML AND WIKI

In Fig. 2 the linking between wiki and XML-file is shown. A corresponding entry in the wiki system can be seen in Fig. 3. All information regarding the single entity rather than the text itself is outsourced to the wiki page. Only information belonging specifically to the text is stored in the XML-file.[11]

3.1 Working with the oXygen Author mode

Working with an XML editor simplifies typing considerably. Still the work can be cumbersome for scholars from classical humanities. The oXygen tool offers the possibility to build a custom-tailored graphical interface to work with XML, which creates a more fluent and intuitive work process.
Making use of *cascading stylesheets* (CSS), the XML content can be represented in a more user-friendly way. Even though this already facilitates working with XML, another substantial advantage comes with the declaration and implementation of self defined actions, buttons, and functions. They can make use of the full set of possibilities offered by either the oXygen Java-API or Java in general.
Besides general functions, such as inserting a new paragraph at the cursor position, project specific functions have been implemented for our environment. Of special interest are those actions that establish a connection to the wiki.
Fig. 4 shows a screenshot of the GUI. The XML is represented using different colors, boxes, and tables according to the underlying CSS. The red box highlights the project-specific toolbar. After clicking the button to add a new person reference, a dialogue window presents the user with a list of all available wiki entries within the respective namespace . As the user types letters, the list will automatically be filtered according to the user's input, allowing to easily find and select the appropriate wiki page. Some further attributes can be set, and after these selections have been made, the corresponding XML-code, as shown in Fig. 2, is

[9]Within the first five months of using the wiki 3417 entries for persons and 615 place entities were created. Only for the authors of texts biographical information is collected systematically. For other persons just enough information is collected to make them distinguishable.
[10]This will be performed by one of the project partners, the Herder Institute, who has editing privileges for the PND.
[11]Besides this project-internal linking between different data, the entities described in our project are also intended to be linked to external repositories or to authorities files, e.g. the previously mentioned PND. Furthermore, data sets collected during the project will be made available to relevant (library) catalogues.

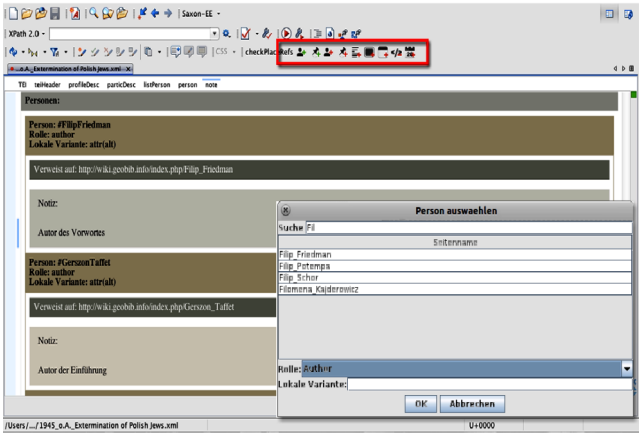

Figure 4: The GUI: CSS formatted XML, customized toolbar, and wiki-connection dialog.

generated and inserted into the XML-file.[12] The automatically generated code is fail-save regarding the XML-syntax, the code's validity, and the existence of the linked entity in the wiki.

3.2 Quality management

Some further steps are necessary to ensure the separation of concerns between XML-files and wiki entries, and to successfully overcome the problem of inconsistencies. When new pages are added to the wiki, they can immediately be referenced from XML-files. In cases where wiki page needs to be moved or renamed, existing links could lose their validity. Therefore, we added the possibility to show all XML-files that currently point to a given wiki page (see Fig. 5). A script to be run before (re)moving a wiki page takes the page title as input and checks every reference in every XML -file for that URL. This procedure is possible since all XML-files are simultaneously available on the server and to all users thanks to the central SVN repository used for data exchange.
Other scripts in oXygen check for false cross references, i.e. ID/IDREFS that do not point to a valid XML-node or ID. Furthermore the URLs contained in an XML file can be checked for validity from within oXygen, e.g. to make sure a corresponding wiki page exists.

4. DISCUSSION

The decision to provide a specifically tailored virtual back office for our project team shall not be left uncommented, since it raises questions regarding the necessity and future perspective of this effort, especially in the light of ongoing general infrastructure developments targeting the realm of the digital humanities.
The distinction between service and research in the digital humanities [2] has been much debated and more recently perceived as constructed rather than natural [7]. In our case, the distinction of roles between service and research is neither easily applicable nor stable over time. The actual

[12]Images, such as scans of covers or illustrations, can be added to the document in a similar fashion.

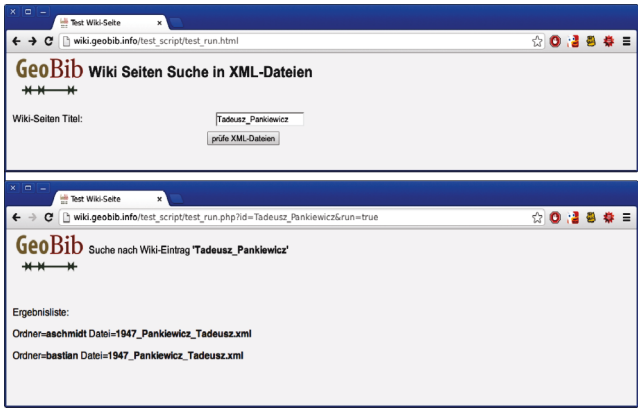

Figure 5: Querying XML-files for referenced wiki-page. Top: Start query for a string (page title). Bottom: Result: Creator and name of files containing search-string.

relations are more complex and dynamic. While the 'technical' team is in service position when offering a working environment for the humanities researcher to enable them to 'do their work' within the project, these roles are being switched in the next stage. During data processing and quality assurance, the human annotators, i.e. the 'researchers', feed the data into the technical process. They enable the 'technical team' to do their job. The resulting research platform is finally expected to enable humanities researchers to shed light on issues previously unknown or unsolved - the service role has switched again. In a chain of scenarios like that, e-humanities is a cascade of alternating role-taking in enabling each other.

While the use of open source or reasonably priced standard software makes it inviting to assemble a particular working environment, out-of-the box solutions like TextGrid [6] or the eHumanities Desktop [4] must also be considered. From our point of view it was necessary to find the right balance between DIY-efforts and flexibility on the one side and mid-term dependencies on large scale humanities research infrastructures on the other side. While we would not necessarily consider the latter as "a dead end for digital humanities" [8], we decided that a certain level of DIY leaves us with reasonable control and flexibility regarding our team's working environment.

5. SUMMARY

The data management within the GeoBib project raised some challenges: How to collect data and information on such different objects like literary texts, persons, and places within a coherent and highly collaborative workflow that involves contributors from such different backgrounds as literary studies, history and geography? Our initial attempt relying solely on TEI/XML was error-prone. Problems of inconsistency and redundancy arose.

As a consequence we currently apply a strategy that separates data concerning the Holocaust texts one the one side from data describing persons and places on the other side. We use TEI to do what it does best: collecting information regarding texts. But in addition we employ a wiki system, which combines machine-readability and processability while still being easy-to-use for our collaborators. Information on persons and places that are shared among various TEI/XML-files are now being collected within the wiki system, which provides a fixed URI for each place and person. To integrate XML/TEI and the wiki, we use the oXygen XML Editor along with a few Java-based extensions. These extensions assist in regular data management tasks and facilitate the linking of entities occurring in the XML-files to their corresponding wiki pages.

Setting up such a particular working environment for our project was both an obvious albeit disputable endeavor. Nevertheless, we chose this approach in order to maintain a reasonable level of control and flexibility regarding our team's working environment.

6. REFERENCES

[1] S. Dumont and M. Fechner. Digitale Arbeitsumgebung für das Editionsvorhaben "Schleiermacher in Berlin 1808–1834", 2012. `http://digiversity.net/2012/digitale-arbeitsumgebung-fur-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834/`, visited 2013-08-14.

[2] E.C. (European Commission). Riding the wave: How europe can gain from the rising tide of scientific data., 2010. `http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf`, visited 2013-08-14.

[3] B. Entrup, M. Bärenfänger, F. Binder, and H. Lobin. Introducing GeoBib: An Annotated and Geo-referenced Online Bibliography of Early German and Polish Holocaust and Camp Literature (1933–1949). In *Digital Humanities 2013. Conference Abstracts.*, 2013.

[4] R. Gleim, P. Warner, and A. Mehler. ehumanities desktop - an architecture for flexible annotation in iconographic research. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10), April 7-10, 2010, Valencia*, 2010.

[5] K. Hickethier. Biographie, Autobiographie, Memoirenliteratur. In L. Fischer, editor, *Literatur in der Bundesrepublik bis 1967*, pages 574–584. München, 1986.

[6] H. Neuroth, F. Lohmeier, and K. M. Smith. Textgrid - virtual research environment for the humanities. *IJDC*, 6(2):222–231, 2011.

[7] S. Palfner. E-Science-Interfaces – ein Forschungsentwurf. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann, editors, *Beiträge der Tagung "Digitale Wissenschaft - Stand und Entwicklung digital vernetzter Forschung in Deutschland"*, pages 123–129, 2010.

[8] J. van Zundert. If you build it, will we come? large scale digital infrastructures as a dead end for digital humanities. *Historical Social Research*, 37(3):165–186, 2012.

[9] N.-O. Walkowski. Das Konzept einer polysemischen Datenbank und seine Konkretisierung im Personendaten-Repositorium der BBAW. *Jahrbuch für Computerphilologie - online*, 2011.