Towards a Digital Archive for Handwritten Paper Slips with Ethnological Contents

A.C. Schering[1], I. Bruder[2], C. Schmitt[3], H. Meyer[1], and A. Heuer[1]

[1] Database Research Group, Dept. of CS, University of Rostock, Germany
[2] IT Science Center Ruegen, Putbus, Germany
[3] Wossidlo Archive, University of Rostock, Germany

Abstract. Contemporary digital libraries and archives of ethnological information focus mainly on document based storage and access methods for their data. However, our archive is designed to manage smallest pieces of information and can enable ethnologists not only to easily store and access their material, but also to derive new knowledge by combining existing data. In this paper, we present the first steps in building a digital archive for paper slips with ethnological contents from the 19th and the beginning of the 20th century. Along with the architectural and accessibility aspects of the *WossiDiA* system, we describe enhancements for efficient retrieval and for supporting modifications to access structures.

1 Introduction

The Wossidlo Archive embraces one of the most diverse collections of highly interconnected paper slips documenting regional folk culture, worldwide. It was founded by Richard Wossidlo, an ethnologist and ethnographer who studied and collected terms and definitions in the Low German language related to ancient customs with location-specific usage in the region of Mecklenburg, Germany, between 1883 and 1939. With the help of several hundreds of contributors, he labeled millions of slips of paper with cognitions from his field research. These slips are organized in small bundles which are physically represented by envelopes, stored in wooden boxes which are arranged in huge shelves.

This paper represents the latest results from the *WossiDiA (Wossidlo Digital Archive)* project, developing a digital archive for a collection of about two million slips of paper, about 30,000 envelopes, more than 1,200 boxes, and additional material such as correspondences and literature excerpts. Among the scanned images of the paper slips and documents we store a substantial number of corresponding metadata such as bibliographic and classification data, reflecting both the diverse aspects on the contents of slips and documents as well as their correlation and interdependence. The information encoded in the paper slips, such as the cultural fact, description, date and place of origin as well as contributor is of great interest to cultural scientists. To retrieve this information we employ full-text and hierarchical retrieval techniques, as used in most existing digital document archives. Wossidlo defined complex access mechanisms using detailed finding aids for different topics and a number of indexes and thesauri. The interesting feature of his slip collection is that he had not interlinked the papers

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 61–64, 2007.

with just the access mechanisms, but even the papers among each other as well. Comprehensive information cannot be found in a single paper slip. It has to be compiled from slips and documents all across the archive. Therefore the true value of *WossiDiA* lies in the structures and relationships between facts and contents of the papers in conjunction with the comprehensive finding aids (e.g., shepherd is specified in the categories labour, rites, and magic (rites→shepherd rites→shepherd (labour)←shepherd magic←magic). The challenges we discuss particularly in this paper are modeling and implementation of the digital archive as well as providing means for efficient storage, complex retrieval, and structural modifications supported by data mining techniques within this framework.

2 Model, User Scenarios and Architecture

The first step in designing the digital archive as proposed by [2] was to identify the material to be included, such as scanned image data (slips of paper and auxiliary documents), corresponding metadata, and detailed finding aids including indexes and thesauri. We have devised a conceptual data model to represent materials, digital data and metadata, as well as access structures on three levels. It defines (1) concrete entities, such as slips of paper, envelopes and boxes on the physical level, (2) digital images and their metadata, such as bibliographic data, descriptions, place and time of origin on content level, and (3) finding aids, taxonomies, indexes and thesauri on the access level. The metadata in the digital archive are stored in METS format [3]. This is especially useful for bibliographic metadata and can be further utilized to hold descriptive metadata recorded in formats such as DC, MODS, EAD, etc. Multimedia objects are modeled by using the multimedia metadata standard MPEG-7. Such multimedia objects are image data (slips) and image parts (special regions providing signature, keywords, names, dates, etc.) as well as audio data (voice recordings for particular slips). MPEG-7 fragments can be easily integrated into METS. Both data model and formatting issues are thoroughly described in [5]. The system design (Fig. 1), is based upon the data model described above. *WossiDiA* is implemented upon the Oracle 10g database system, which provides support to efficiently store and manage data and XML metadata. Furthermore, it features a data mining option as well as spatial extensions to store the geospatial data and to write mapping modules. The digital archive provides interfaces for three basic usage scenarios:

Manipulation comprises two tasks: (1) data and metadata input performed by the data entry staff using entry forms provided by the Protégé-Frames editor, specially enhanced to enter *WossiDiA*-specific data, and (2) structural modification of the metadata model collaboratively performed by professional researchers (see Sect. 3). Both user groups are supervised by the domain director who is in charge of the digital archive and the metadata model in particular.

Retrieval is the main usage scenario performed by public users and professional researchers. The public user is only interested in searching and browsing the archive whereas the professional researcher also intends to alter structures

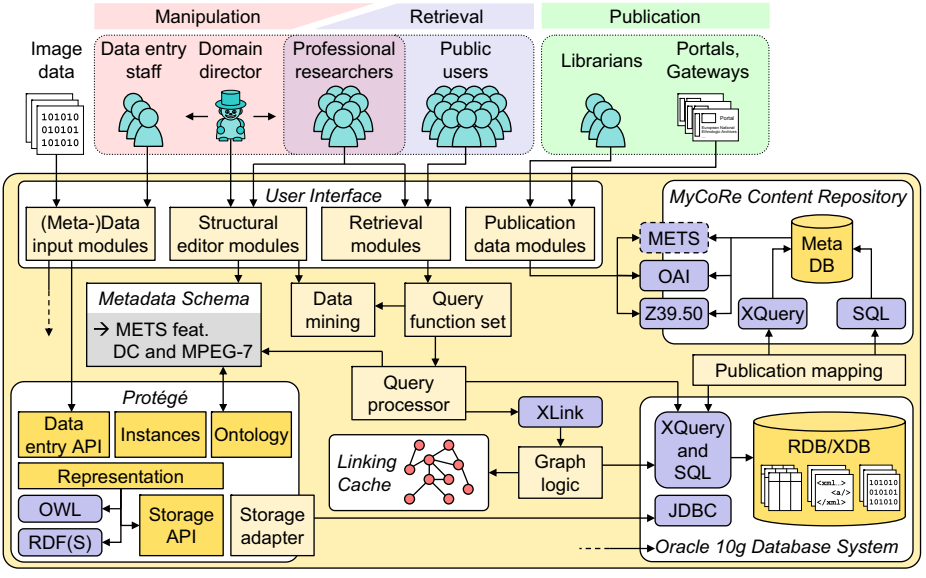

Fig. 1. *WossiDiA* Digital Archive architecture

(see above). The retrieval scenario requires the most efforts in *WossiDiA*, since it is responsible for putting together the various tiny pieces of information mentioned in the introduction. Full-text and hierarchical search mechanisms are used to provide the user with information about a particular topic. Using the various access structures he can obtain more information, e.g., about adjacent topics, from the same contributor, in the same time frame, or within a particular place.

Publication is the scenario used by librarians and archivists to retrieve the information of the archive in a combined manner, such as articles about particular topics. In contrast to the retrieval scenario, publication focuses on content generation according to specific standards and formats, including, but not limited to METS, DC, OAI, and Z39.50. Thus, the *WossiDiA* material can easily be integrated into libraries as well as portals and gateways in the field of cultural sciences. We use a MyCoRe content repository as publication subsystem to provide the formats mentioned above and to enable *WossiDiA* to be integrated into the web portal of the Library of the University of Rostock.

3 Enhancing the Architecture

Structure editing enhancements: The main problem in the course of creating an adequate system design is to give the professional researcher the opportunity to efficiently alter finding aids and taxonomies. For this purpose we have introduced the structural editor module to the system architecture. It enables the user to conduct modifications of the structures manually. Since the number of data and metadata records is extraordinarily high and the amount of relationships between

the paper slips and their access mechanisms is quite substantial the professional researcher needs an additional help provided by data mining techniques [1]. Data mining algorithms are utilized to derive implicit, previously unknown, cultural information hidden in the interior of Wossidlo's collection all across the archive on every single slip of paper. To help the user to develop new and to adjust existing Wossidlo-specific finding aids we provide solutions to employ automatic data mining algorithms in the context of the structural editor module. In the course of that we use clustering analysis to identify homogeneous groups of information units upon which terms used in the respective finding aids are to be built. Another important data mining scenario is the identification of groupings in the data for retrieval purposes, e.g., to find contributors writing about a certain fact limited regionally and chronologically. This requires the integration of data mining aspects into the retrieval module as well. For that scenario, we use in particular the clustering algorithms k-means (partitional) and O-Cluster (hierarchical) [4], both featured by the Oracle database system.

Retrieval enhancements: WossiDiA includes a query processor for a domain specific query language based on XQuery. Some extensions, such as XLink and graph logic, are necessary because XQuery is primarily based on a hierarchical data model and thus cannot handle diverse graph structures efficiently, especially inter-slip relationships. A function set provides methods to query and to navigate our XML-based model. A linking cache is introduced to efficiently manage relationship information condensed to a very small size in order to keep as much links in main memory as possible.

4 Conclusions

The digital archive concept presented in this paper allows data storage, access, analysis, retrieval and presentation of this unique cultural material. Dealing with a huge number of smallest pieces of information which are highly interconnected, the implementation of *WossiDiA* requires unusual techniques in terms of conventional digital archives. Furthermore, the system will provide intelligent means enabling knowledge deduction and reclassification of finding aids.

References

1. Han, J., Kamber, M.: Data Mining: Concepts&Techniques. Morgan Kaufmann, San Francisco (2006)
2. Witten, I., Bainbridge, D.: How to Build a Digital Library. Morgan Kaufmann, San Francisco (2003)
3. The Library of Congress. Metadata Encoding and Transmission Standard: Primer and Reference Manual (2007), `http://www.loc.gov/standards/mets/`
4. Milenova, B.L., Campos, M.M.: O-Cluster: Scalable Clustering of Large High Dimensional Data Sets. In: ICDM, pp. 290–297 (2002)
5. Schering, A.-C., Bruder, I., Meyer, H.: Management of Highly Interconnected Information Units in the Digital Wossidlo Archive. In: Proceedings of the 19th GI-Workshop on Foundations of Databases, Germany (2007)