The PROBADO Project - Approach and Lessons Learned in Building a Digital Library System for Heterogeneous Non-textual Documents

R. Berndt[1], I. Blümel[2], M. Clausen[3], D. Damm[3], J. Diet[4], D. Fellner[5,6], C. Fremerey[3], R. Klein[3], F. Krahl[4], M. Scherer[5], T. Schreck[5], I. Sens[2], V. Thomas[3], and R. Wessel[3]

[1] Graz University of Technology, Austria
[2] German National Library of Science and Technology Hannover, Germany
[3] University of Bonn, Germany
[4] Bavarian State Library, Munich, Germany
[5] Technische Universität Darmstadt, Germany
[6] Fraunhofer Institute for Computer Graphics, Darmstadt, Germany

Abstract. The PROBADO project is a research effort to develop and operate advanced Digital Library support for non-textual documents. The main goal is to contribute to all parts of the Digital Library work flow from content acquisition over indexing to search and presentation. While not limited in terms of supported document types, reference support is developed for classical digital music and 3D architectural models. In this paper, we review the overall goals, approaches taken, and lessons learned so far in a highly integrated effort of university researchers and library experts. We address the problem of technology transfer, aspects of repository compilation, and the problem of inter-domain retrieval. The experiences are relevant for other project efforts in the non-textual Digital Library domain.

1 Introduction

Digital Library technology offers many effective ways to handle document content. Access and delivery of documents becomes more and more digital and decentralized, and new user groups can benefit from library services. This is true for textual documents. However, technological and scientific progress contribute to increasing availability of non-textual documents, which are worthy of library-oriented treatment. Examples include digitization efforts in Cultural Heritage, production of scientific film, recording of orchestral performances, as well as masses of primary research data produced in the natural sciences. All of these non-textual documents, while being potentially relevant for library-oriented service, are more difficult to accommodate in a Digital Library system than their textual counterparts. Main challenges in supporting non-textual documents include questions of document representation, indexing and content-based accessing, and document presentation. Specifically, content-based access in non-textual

documents is a difficult problem as appropriate methods usually are application dependent and nontrivial to implement.

From the field of multimedia databases and multimedia visualization, many promising approaches have been proposed. But even if relevant document domains, use cases, and accommodation strategies have been identified, the problem of deploying such approaches within the operational context of a library operator needs to be solved. PROBADO aims at designing, developing and deploying Digital Library functionality for non-textual documents for a selection of use cases. At the same time, the project aims to propose a general reference architecture and protocol for consolidation of distributed non-textual document repositories of heterogeneous document types.

In this paper, we report on the approach taken and the experiences made during the first three and a half years of the PROBADO project. We systematically discuss the challenges that arose so far during the project, and sketch our solutions for them. The contribution of this paper is to offer a joint conceptual and practical perspective on a substantial Digital Library research and deployment effort.

2 Related Work

We briefly recall related work on Digital Library systems and Multimedia Retrieval. Additional related work specific to the domains discussed throughout this paper is recalled in the corresponding paper sections.

Existing Digital Library systems include Fedora[12], Greenstone[20], DLib[4] and Variations[8]. Fedora, Greenstone, and DLib support building Digital Libraries for textual documents; support for multimedia documents relies on metadata annotations according to specific standards such as MPEG-7. In PROBADO, the goal is to index and access non-textual documents specifically by *content-based* approaches. Therefore, the aforementioned systems are not directly applicable to our approach.

In multimedia retrieval, commercial systems and research prototypes exist. Examples include Google's *Similar Images* and *3D Warehouse*, both of which allow for content-based search. VICTORY[6] is a research project developing content-based retrieval of 3D data using a peer-to-peer architecture. Multimedia retrieval systems such as these employ the same basic approach as PROBADO for supporting content-based search. Given a multimedia query (e.g. an example document), the system computes a mathematically tractable representation (descriptor) for this query and compares this to a database of descriptors of the indexed content. Details for search approaches in 3D and music retrieval as used in PROBADO are given in Sections 3.2 and 3.3.

3 The PROBADO Approach

PROBADO is a distributed multimedia Digital Library system developed jointly by university researchers and scientific library experts. PROBADO supports

metadata-based and content-based retrieval of *3D architectural models* and *classical music*. We give a concise review of the system components and the development and technology transfer approach.

3.1 Overview of the PROBADO System Architecture

The PROBADO framework is designed to integrate heterogeneous multimedia documents from distributed, specialized document repositories by means of a three layer architecture. User interface, middleware, and repository layers communicate by a SOAP-based web-service.

Users formulate content-based queries using document-dependent search interfaces provided by the repository layers. These queries are forwarded to the middleware. Any user interface needs to implement at least one of the search functions provided by the middleware. These query interfaces support either the search for textual metadata, the search for content-specific data or multi-modal search for both content and metadata [5]. The middleware layer forwards *content-based* queries to all connected repositories supporting the addressed search functions. *Metadata* queries are evaluated directly in the middleware, which hosts a consolidated index of metadata of all repositories. A synchronization mechanism keeps this metadata index up to date with the repositories. The repositories process the content-based queries. Result lists are returned to the middleware for aggregation and presentation to the user.

3.2 PROBADO 3D Repository

The PROBADO 3D Repository supports content-based indexing and retrieval in 3D architectural model data. It aims to support the architectural design process by searching in a Digital Library of architectural model data for re-usage, comparison and inspiration purposes. Useful content ranges from small furnishing objects to environmental elements up to building units and whole buildings.

Current approaches to 3D shape retrieval mainly focus on search for models that are geometrically similar to a query object. These methods are usually based on global or local shape descriptors. Additionally, view-based algorithms as well as graph-based approaches have been proposed. A detailed overview of state of the art methods in this area can be found in [15].

Data Preprocessing. During preprocessing, low-level technical metadata of the 3D model are extracted, previews are generated and for subsequent topological indexing, 3D building models are oriented and scaled consistently [3].

Content-based Indexing & Metadata. Content-based indexing allows searching in a query-by-example scenario and enables high-level metadata generation. For each model, a global shape descriptor is computed. Additionally, local shape descriptors are computed providing a high-quality object description, serving as a starting point for high-level metadata generation, eventually producing a

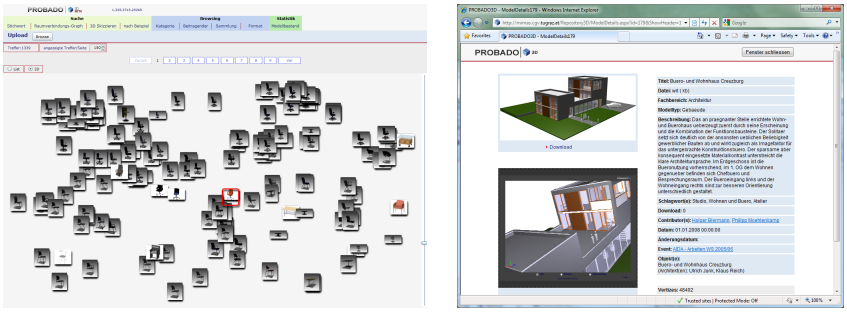

Fig. 1. (left) 2D result visualization. (right) Model details with integrated 3D preview.

Room Connectivity Graph (RCG) [19] characterizing their topology. From the RCG extraction phase, also high-level metadata like height of building models, the number of floors, doors, windows etc is obtained and stored for user access. Based on a supervised learning framework [18] using a preclassified 3D architecture benchmark [17], the model category is predicted and stored as well.

The 3D repository additionally stores metadata provided by the model creators including title, description, contributor information etc. These metadata together with the extracted semantic metadata can be queried for by means of simple and extended search forms.

Query-by-example. We currently provide four ways to formulate a query-by-example based on complete 3D models: (1) upload of example model; (2) a 3D sketch interface based on GML[2]; (3) a plug-in for the Google^{TM}Sketchup modeling tool; and (4) using a previous query result as a query key. (2) is tailored to building models and based on searching the extracted RCGs for certain spatial arrangements of rooms and floors. We provide visual-interactive interfaces for all content-based search modalities as described in [1].

Result Visualization. Apart from traditional sequential result lists, the 3D layer currently provides a 2D visualization for results based on global object similarity, which is realized using multidimensional scaling. The details page for a selected result contains also a 3D preview based on PDF (see Fig. 1).

3.3 PROBADO Music Repository

The PROBADO Music Repository supports content-based indexing and retrieval of digital classical music documents. This document notion includes different document types representing different aspects of a piece of music (e.g., sheet music, compact discs, and libretti). At the Bavarian State Library (Bayerische Staatsbibliothek, BSB) a digital collection of western classical music has been established. The collection currently contains approx. 96,000 pages of sheet music

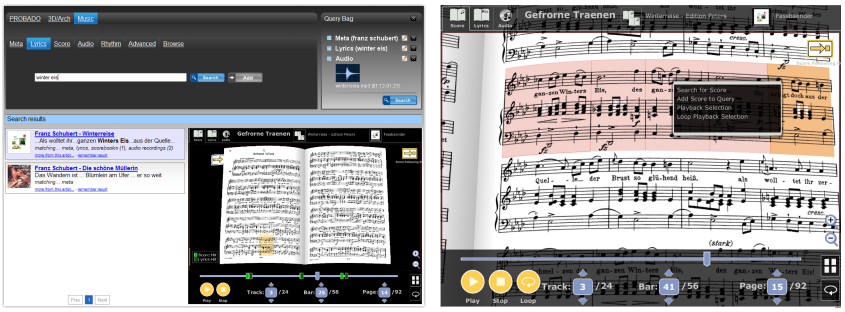

Fig. 2. (left) The PROBADO music frontend with integrated Score Audio Player. (right) The sheet music visualization can be used to perform content-based retrieval.

and corresponding audio recordings from compact disks. Facing such large multimodal digital document collections, systems to manage, process, browse, and access this data are required. Within PROBADO, those requirements are being implemented. In addition, the well-established metadata search is expanded by offering content-based search functionalities.

Music information retrieval (MIR), amongst others, comprises the fields of content-based music retrieval and music alignment. The aim of content-based retrieval is to search for all occurrences of a query (e.g., melody, excerpt of a score, audio fragment) or slight variations thereof in a collection of music documents [10, 14]. In the field of music alignment, different representations of the same piece of music are linked with each other, such that given a position within one document, the position within the other document describing the same musical position can be obtained [9, 13, 11]. For further literature on these and similar topics we refere to the proceedings of the annual ISMIR conference.

Applied MIR Techniques. In PROBADO we apply MIR techniques to preprocess a music document collection, to enable content-based retrieval, and to offer a holistic, attractive access to music documents. The developed preprocessing workflow provides a user interface for classical library tasks like metadata annotation. Moreover, automated MIR tasks are included (e.g., segmentation of scores, calculation of alignments between different music representations) [16].

Content-based Search Functionalities. For music documents, query engines are available, which process the following query formulations: (1) metadata; (2) lyrics; (3) audio fragments; (4) sheet music extracts; (5) a virtual piano to enter a music query.

Presentation. Presentation of music documents is realized by the Score Audio Player applet [5, 16] (SAP, Figure 2). Its goal is an integrated presentation of

all music documents representing the same piece of music. Due to the alignment information, synchronized playback of an audio recording while highlighting the corresponding bar (measure) within the sheet music is supported, allowing score based navigation. Also, the user can switch between different recordings while maintaining the musical position. Using the sheet music visualization, a number of bars can be selected and directly be used as content-based query. The SAP also provides a detailed view on the matching regions within the piece of music.

4 Lessons Learned

4.1 System Architecture

The architecture of a distributed Digital Library system faces several challenges including metadata abstraction, relevance feedback, and inter-domain retrieval. To integrate heterogeneous document types a consolidated meta data abstraction is crucial. The trade-off between few but generic metadata fields and more, possibly specialized fields has to be regarded. In PROBADO a decision in favor of a compact DC-oriented metadata set was taken, securing extensibility to new domains. Specific metadata queries are still possible by directly querying in the individual repositores, but a joint metadata search over all repositories is evaluated using the unified DC scheme.

Relevance feedback (RF) techniques are important to support effective retrieval in multimedia data, but are difficult to apply in a heterogeneous and distributed environment. Results to be given feedback about may originate from different repositories. But since a given repository usually does not have information about the content of other repositories, it cannot solely apply the RF optimization mechanisms. Therefore RF-techniques are not employed within PROBADO.

Searching for multimedia data across domain boundaries is an open research question. To formulate a query-by-example, which is to be evaluated in combinations of domains, a compatible query syntax is necessary. E.g., for a combined content-based query in 3D models and 2D image data, a common syntax could be based on 2D images, as any 3D model can be projected to a 2D image. For other domain combinations like 3D model data and classical music, no such projection exists. Nonetheless, textual annotations and query-by-text can support inter-domain retrieval. Automatically generating semantically meaningful textual annotations from multimedia content is another research challenge. Inter-domain retrieval by textual queries is possible in PROBADO, restricted to manually obtained textual metadata.

4.2 Two Alternative Approaches to Repository Compilation

Our project includes the compilation of document repositories for each domain for three main purposes: (1) a reference collections for development and testing; (2) serve for demonstration purposes, raising interest; (3) obtain experience with digitizing and obtaining of documents from external providers.

The music reference collection is a large-scale digitization effort carried out in-house with the BSB library. For PROBADO purposes, this digitization workflow was augmented by an OMR-process ("optical music recognition"). The metadata model within the PROBADO music repository uses a work-centric data model that is based on the Functional Requirements of Bibliographic Records (FRBR)[7]. This *institution-oriented* approach is a highly structured process providing full control over the repository w.r.t. content, quality, and metadata.

The 3D repository comprises about 8,000 indexed models including buildings, construction units, furnishing etc. Providers include architectural component manufacturers, web portals for 3D content, and architecture faculties of universities. File formats, level of detail, content, quality, and the existence of metadata vary substantially. This *provider-oriented* approach is characterized by heterogeneity of the documents. Focus, format, resolution, and level of detail varies between documents.

5 Conclusions and Future Work

We reported on the approach and lessons learned in developing and deploying content-based Digital Library support for certain non-textual documents. While much has already been achieved in terms of functionality, selecting and transferring a suitable subset of functionality into practical operation represents organizational and technological challenges. Architectural and application implications relating to the distributed and heterogeneous system model, have been identified and were discussed. Two modes of repository compilation and two operation models were identified and compared.

Next steps involve actual transfer of functionality to the project library partners BSB and TIB, and customization of functionality for user needs. Steps in this stage include: (1) selection and consolidation of system functionality to be deployed, from the larger pool of developed functionality; (2) shaping the interfaces of the components to suit the hosting operational environment; (3) documentation and training of librarians and IT technicians; and (4) testing and usability iterations.

Due to the middleware abstraction layer, our approach does not restrict the supported document model. Consequentially, integration of additional document repositories is possible and will be aimed at. In the long run, research questions relating to the development of a document model supporting retrieval, presentation, and annotation of collections of heterogeneous non-textual documents need to be addressed.

Acknowledgments

PROBADO is a joint research project supported by the German Research Foundation DFG under the LIS program. PROBADO started in February 2006 with a tentative duration of five years. Sven Havemann, Harald Krottmaier, Frank Kurth, and Thorsten Steenweg made valuable contributions to the project effort. For further information, please visit the project website at http://www.probado.de/.

References

1. Berndt, R., Blümel, I., Krottmaier, H., Wessel, R., Schreck, T.: Demonstration of user interfaces for querying in 3d architectural content in PROBADO3D. In: 13th European Conference on Digital Libraries (2009)
2. Berndt, R., Havemann, S., Fellner, D.: 3D Modeling in a Web Browser to Formulate Content-Based 3D Queries. In: Behr, J., Walczak, K. (eds.) Proceeding of the 14th International Conference on 3D Web Technology. Eurographics Association, Darmstadt, Germany (2009), http://www.eg.org/EG/DL/PE/WEB3D09/111-118.pdf
3. Berndt, R., Blümel, I., Wessel, R.: Probado3d towards an automatic multimedia indexing workflow for architectural 3d models. To be presented at 14th International Conference on Electronic Publishing, Helsinki (Jun 2010)
4. Castelli, D., Pagano, P.: Opendlib: A dl service system. In: ECDL (2002)
5. Damm, D., Kurth, F., Fremerey, C., Clausen, M.: A concept for using combined multimodal queries in digital music libraries. In: 13th ECDL (2009)
6. Daras, P., Tzovaras, D., Dobravec, S., Trnkoczy, J., Sanna, A., Paravati, G., Traphoener, R., Franz, J., Kastrinogiannis, T., Malavazos, C., Ploskas, N., Gumz, M., Geramani, K., Wintterle, G.J.: Victory: a 3d search engine over p2p and wireless p2p networks. In: 4th International Conference on Wireless Internet (2008)
7. Diet, J., Kurth, F.: The PROBADO music repository at the Bavarian State Library. In: 8th International Conference on Music Information Retrieval (2007)
8. Dunn, J.W., Byrd, D., Notess, M., Scherle, R.: Variations2: Retrieving and using music in an academic setting. Communications of the ACM 49 (2006)
9. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2003)
10. Kurth, F., Müller, M.: Efficient index-based audio matching. IEEE Transactions on Audio, Speech, and Language Processing 16, 382–395 (2008)
11. Kurth, F., Müller, M., Fremerey, C., Chang, Y., Clausen, M.: Automated synchronization of scanned sheet music with audio recordings. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR) (2007)
12. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. Int. J. Digit. Libr. 6, 124–138 (2006)
13. Orio, N.: Alignment of performances with scores aimed at content-based music access and retrieval. In: Proceedings of the 6th ECDL (2002)
14. Suyoto, I., Uitdenbogerd, A., Scholer, F.: Searching musical audio using symbolic queries. IEEE Transactions on Audio, Speech, and Language Processing 16 (2008)
15. Tangelder, J.W., Veltkamp, R.C.: A survey of content based 3d shape retrieval methods. Multimedia Tools and Applications 39, 441–471 (2008)
16. Thomas, V., Fremerey, C., Damm, D., Clausen, M.: SLAVE: a Score-Lyrics-Audio-Video-Explorer. In: Proceedings of the 10th ISMIR (2009)
17. Wessel, R., Blümel, I., Klein, R.: A 3d shape benchmark for retrieval and automatic classification of architectural data. In: EG Workshop on 3D Object Retrieval (2009)
18. Wessel, R., Baranowski, R., Klein, R.: Learning distinctive local object characteristics for 3d shape retrieval. In: Vision, Modeling, and Visualization (2008)
19. Wessel, R., Blümel, I., Klein, R.: The room connectivity graph: Shape retrieval in the architectural domain. In: WSCG (2008)
20. Witten, I.H., Mcnab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: Proceedings of the Fifth ACM International Conference on Digital Libraries (2000)