

# THE PROBADO MUSIC REPOSITORY AT THE BAVARIAN STATE LIBRARY

**Jürgen Diet**

Bavarian State Library  
Ludwigstraße 16  
80539 Munich, Germany  
juergen.diet@bsb-muenchen.de

**Frank Kurth**

University of Bonn  
Römerstraße 164  
53117 Bonn, Germany  
frank@iai.uni-bonn.de

## ABSTRACT

In this paper, we describe the Probado music repository which is currently set up at the Bavarian State Library, Munich, as part of the larger German Probado digital library initiative. Based on the FRBR approach, we propose a novel work-centric metadata model for organizing the document collection. The primary data contained in the repository currently consists of scanned sheet music and digitized audio recordings. The repository can be searched using both classical and content-based retrieval mechanisms. To this end, we propose a workflow for automated content-based document analysis and indexing.

## 1. INTRODUCTION

The vast majority of content in today's digital libraries that are in productive use consist of textual documents. Even though a lot of research has been done on how to manage, retrieve, and present multimedia documents (the research presented at the ISMIR conferences shows the progress in case of music documents), there is still the need for integrating multimedia documents in existing library work-flows. User-friendly tools must be developed so that the management of multimedia documents for librarians and the user access to these documents (both content-based and in the conventional way of searching the metadata) become possible.

In this paper, as a concrete step in this direction, we describe an ongoing project initiative that aims at setting up a digital music repository at the Bavarian State Library ("Bayerische Staatsbibliothek", BSB), Munich. In a cooperative approach with MIR researchers, the project aims at developing an integrated work-flow for both document handling and cataloging according to the classical librarian workflow *and* content-based document processing, i.e., making the collection accessible through content-based retrieval, the latter involving automatic content-based document analysis and indexing. As key contributions, we

- propose a generic structure of a music repository for storing relevant data of real-world digital music libraries that is suitable for both conventional and content-based access to music documents,
- propose a general metadata scheme for music repositories based on the FRBR model [1],

- describe a concrete design of a corresponding music repository that is currently set up at BSB and in particular describe a prototypical workflow for automatic content-based document analysis, annotation, and access,
- sketch how the proposed repository will be integrated into the larger-scale Probado-framework that facilitates access to general document types stored in digital libraries.

After summarizing relevant related work in Section 2, Section 3 describes the Probado digital library project. Sections 4 and 5 are devoted to the Probado music repository currently set up at BSB.

## 2. RELATED WORK

The Variations2 project at Indiana University in Bloomington [2] provides a user-friendly access to music in form of scanned sheet music and audio. Its metadata model is also based on the FRBR model. Intended users are students in music and in musicology. Two research projects in the area of digital music libraries and archives have been started recently and are funded by the European Union: The EASAIER project [3] which is coordinated by Queen Mary University of London will create new and innovative methods of access to sound archives. In the DISMARC project [4], a consortium of 10 project partners, will make audio collections searchable, discoverable and, where the content owner provides access to audio, listenable.

## 3. THE PROBADO-PROJECT

Probado is a cooperative German digital library project funded by Deutsche Forschungsgemeinschaft (DFG). The project started in February 2006 and has a tentative duration of 5 years<sup>1</sup>.

The main goal of Probado is to integrate general (in particular non-textual) multimedia documents into the work-flow of existing libraries. Important subtasks of the project are to develop and implement (a) methods to support automatic processing of general documents in the library processing chain of document acquisition, annotation, search & delivery, and storage, (b) local Probado repositories which are located at particular libraries and in which documents and annotation data for single subject areas (e.g., music) can be organized, and (c) a common Probado platform serving as a web-

based access point for searching and accessing documents stored in the connected repositories. Rather than being a pure research project, it is a special focus of Probado to achieve long-term usage of the developed systems and work-flows at the cooperating libraries. To achieve the above goals, project partners from different universities, each having expertise in distinct areas of multimedia document analysis and retrieval, are cooperating with partners of two large German libraries (TIB, Hannover, and BSB, Munich).

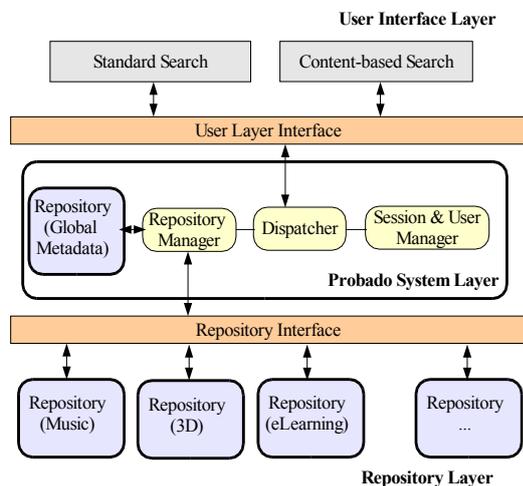


Figure 1. Overall system architecture of the Probado framework.

An overview of the Probado system architecture, which is currently under development, is depicted in Figure 1. It consists of three layers. The user interface layer enables a user to perform both classical text-based queries as well-as content-based queries for the specific subject areas (e.g., query-by-humming for the music subject area).

The main component of the Probado system layer is a task dispatcher that is used to assign incoming queries to the respective query engines located at the connected repositories. As a special feature, global Dublin-Core-style metadata for all documents contained in the connected repositories are managed in a central metadata repository that is hosted at the Probado system.

The repository layer consists of an arbitrary number of Probado repositories, each holding documents and associated data for a specific subject area. In the current stage of the project, Probado repositories are set up for the particular subject areas of music, 3D graphical models, and e-learning. In this paper, we will describe the Probado repository for the subject area of music, which is currently set up at BSB in cooperation with the Multimedia Signal Processing Group at the University of Bonn [8]. Each repository offers a specific number of query engines which are registered at the central repository manager located in the system layer. Besides a mandatory query engine for descriptive metadata, a Probado repository may contain an arbitrary number of query engines for content-based queries.

#### 4. THE METADATA MODEL FOR THE PROBADO MUSIC SUBJECT AREA

The Probado-model is based on the Functional Requirements for Bibliographic Records (FRBR). FRBR is a suggestion of the International Federation of Library Associations and Institutions for an improved structure of bibliographic metadata. The FRBR document [1] has already been published in 1998, and received a lot of attention during the last years. The new cataloguing standard RDA (“Resource Description and Access”) currently developed by the Joint Steering Committee for the Revision of AACR and scheduled for release in early 2009 [5] will incorporate the FRBR concepts.

The FRBR model is very suitable for music because it is work-centered and distinguishes between a work and its expressions. Therefore, several innovative projects in the music library area are based on the FRBR model, like the Variations2 project [2] and MusicAustralia at the National Library of Australia [6].

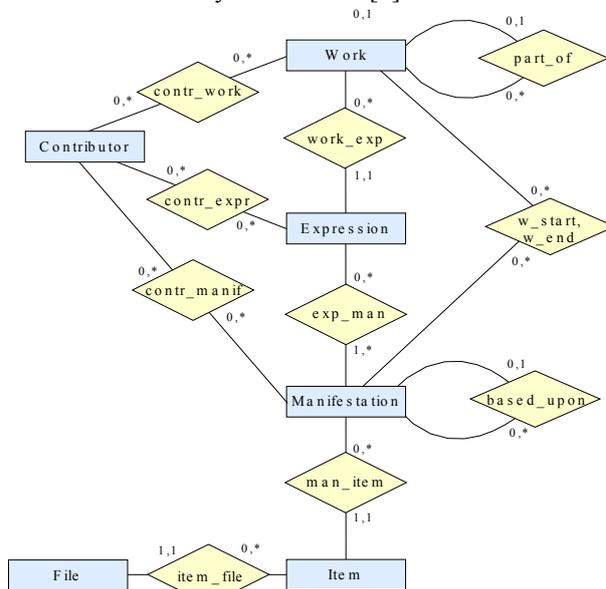


Figure 2. Metadata Model of BSB Music Repository

The FRBR document uses an entity-relationship model for describing the proposed data structure of the bibliographic records. The entities are divided into three groups, where group 1 describes intellectual and bibliographic units, group 2 describes the involved agents, and group 3 describes different subjects. In the remainder of this paper, we will only deal with the group 1 entities “work” (denoting a musical work as an abstract entity independent of any performance or recording, such as “J.S. Bach’s six suites for unaccompanied cello”), “expression” (denoting particular performances of a work as abstract entities, e.g. “the Bach suites as performed by Janos Starker and recorded in 1963 and 1965”), “manifestation” (denoting particular physical realizations of an expression such as “a recording of the Starker performance released on 33 1/3 rpm sound discs in 1965 by Mercury”) and “item”

(denoting particular locations where manifestations are available, e.g. a CD available at BSB).

Figure 2 depicts the metadata model that is used for the music subject area of the Probado project. The model is shown as an entity-relationship diagram where entities are depicted by rectangles and relationships by diamonds. The cardinalities of the relationships are included in (min,max)-notation. “min” stands for the minimal number of relationships that an entity takes part in and “max” for the maximal number (“\*” allows for an arbitrary number of relationships). E.g., the cardinality (0,\*) on the “work”-side of the relationship “work\_exp” means that one work relates to at least 0 and at most arbitrarily many expressions.

The Probado model in Figure 2 is very similar to the FRBR model. All entities in the Probado-model also appear as entities in the FRBR model except for the “file” entity. We introduced the file-entity because we need the information about the physical files that store our music documents. The information contained in the item-entity is not sufficient for modelling digital documents because an item can consist of several files. The Probado-model also extends the FRBR model by two new relationships: “part\_of” and “w\_start, w\_end”. In “part\_of” we store the information which work is part of which other work and can therefore include the structure of a musical work (e.g., the individual “Lieder” that are part of Schubert’s “Winterreise” or the movements that make up a piano sonata). Using this mechanism of a “part\_of”-relation hence allows us to capture the structure of a work by both modeling possible subdivisions into subparts (like movements) as well as allowing to group several works to be subparts of a superordinate work entity (like the Winterreise cycle being superordinate to the constituting Lieder). Therefore, from a technical point of view, the work entity is extended to not only include complete works but also parts and unions thereof. Note that, as for each such work entity we additionally store a type information (e.g., “piece of music”, “movement”, “lied cycle”), this is a proper generalization of the FRBR model. In particular, we do not lose any expressibility of the FRBR approach. The second new relationship compared to FRBR is “w\_start” and indicates the offset at which a work or a work part starts in a particular manifestation. In case of audio, “w\_start” contains the time offset, and in case of sheet music or scanned sheet music it contains the page number (the same applies to “w\_end”, indicating the offset where a work ends).

## 5. THE MUSIC REPOSITORY AT THE BSB

A prototypical repository for collections of music data is currently set up at the BSB in Munich. The BSB is a good candidate for a music repository because its music department owns a large document collection including 347,000 sheet music documents, 36,000 music manuscripts, 78,000 audio recordings, and 132,000 books and journals (on music and musicology).

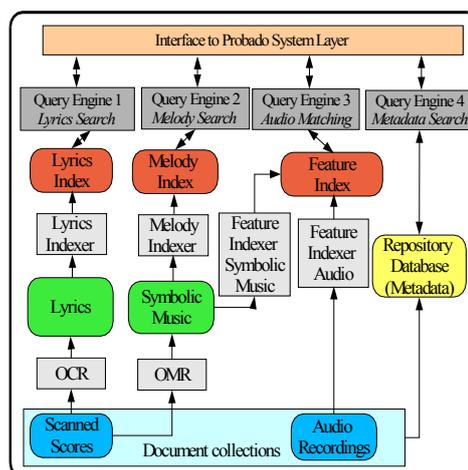


Figure 3. Components of the Probado Music Repository

### 5.1. Data Collections and Digitization

Figure 3 shows an overview of the components of the repository. The repository consists of two document collections, one containing audio recordings, the other one containing scanned images of musical scores. The underlying music material consists of classical and romantic piano sonatas (Haydn, Mozart, Beethoven, Schubert, Schumann, Chopin, Liszt, Brahms) as well as a collection of German 19th centuries piano songs. This collection amounts to approximately 6.000 images of scanned sheet music and about 1.200 musical works in total. The score material has been scanned at the BSB in high quality (600 dpi) and stored in TIFF-format with loss-less compression. For each musical work, audio recordings are currently digitized at the BSB and stored in WAV- and MP3-format within the music repository.

### 5.2. Obtaining the Metadata

A goal of the Probado project is to minimize the manual cataloging work and to automatically generate the appropriate metadata wherever possible. The main source for the Probado metadata is the BSB cataloging database. It contains metadata for the whole library collection in the MAB-format, the German standard corresponding to the Anglo-American MARC-format. The MAB-records describing the Probado collections have to be transformed into the FRBR-based Probado repository database, a process called “FRBRization” [7]. It turns out that this process can not be fully automated. Especially the extraction of the metadata on the work and expression levels needs manual intervention. Although it will be possible to do this cataloging work manually for the initial Probado collections, further automation of this process is required for the future augmentation of the Probado collections.

### 5.3. Searching the Music Repository

The user can search in the repositories connected to Probado in different ways: when no specific subject area is selected, the user enters text in a Google-like

basic search form and the search is done on global Probado metadata which is a subset of the metadata of all connected Probado repositories (similar to Dublin Core). For music-specific searches, the Probado system allows (a) searching in the metadata of Probado music repositories and (b) content-based search.

#### (a) The Probado Music OPAC

The search in the metadata of the Probado music repositories is similar to the traditional search in OPACs of large music libraries. It is a text-based search with a basic search form (with just one input field) and an advanced form. In the advanced case, the user can specify entities and attributes whose values must match the text that the user enters into the input fields.

For the presentation of the search results, it is not sufficient to provide the user with an unstructured list of documents. Instead, the FRBR-structure of the metadata should be preserved. E.g., the Probado music OPAC should differentiate the identification of a musical work from the identification of a particular recording.

#### (b) Content-based Search

For content-based search, the repository will offer three query engines (see Figure 3). Those query engines facilitate text-based search in the textual transcriptions of the piece's vocal tracks (lyrics), melody-based search (Query-by-humming) in the musical notes from the scores, and audio retrieval (Query-by-Example) based on the audio matching approach. For technical details on the corresponding retrieval problems, we refer to the corresponding technical papers [9][10][11]. For efficient retrieval, the query engines employ particular search indexes which are created from the document collections by suitable indexing applications in an offline-operation.

#### 5.4. Content-based Document Analysis and Indexing

Because of the unavailability of comprehensive (i.e., complete and consistent) lyrics- and score-material in a symbolic digital format, the scanned scores are processed by helper applications (Figure 3, light boxes) for optical character recognition (OCR) and optical music recognition (OMR). Although the extracted lyrics- and score-material deviates from the correct data because of OCR- and OMR-errors, the material is nevertheless useful for the search and retrieval process. Using suitable indexer applications, lyrics- and melody-indexes are created from the OMR- and OCR-data. The audio recordings are processed by an audio indexer that extracts chroma-based audio features representing the rough harmonic progression of a piece of audio [9]. From those features, a feature index is created. Without going into technical details, we mention that it is also possible to extract a chroma-based representation from the symbolic (OMR-) data. Adding the resulting features into the audio feature index, audio matching is possible, at the same time, on the audio recordings *and* the score material. Technical details will be reported elsewhere.

## 6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented the Probado music repository for organizing and making accessible collections of digital music documents which is currently set up at the Bavarian State Library, Munich, as a component of the larger Probado framework. The next steps in setting up the music repository will consist of finishing the meta-data acquisition for the initial Probado music collection and subsequently increasing the number of music documents considerably. Concerning content-based document analysis, the next steps consist of evaluating and completing the workflow for automatic OMR/OCR-based indexing and annotation. For content-based retrieval, existing query engines [11] have to be adapted and integrated into the repository. For presenting retrieval results to the user and for facilitating content-based browsing within the music collection, we will adopt components of the existing SyncPlayer framework [10] for multimodal browsing, retrieval, and presentation of music data. Finally, the repository will be connected to the overall Probado framework and the Probado service will be integrated into the conventional library services. For information on the Probado project, we refer to [12].

## 7. REFERENCES

- [1] IFLA Study Group on the Functional Requirements of Bibliographic Records. *Functional Requirements for Bibliographic Records; Final Report*. Saur, Munich, 1998. available at [www.ifla.org/VII/s13/frbr/frbr.pdf](http://www.ifla.org/VII/s13/frbr/frbr.pdf)
- [2] Dunn, J. W., Byrd, D., Notess, M., Riley, J., and Scherle, R. *Variations2: Retrieving and Using Music in an Academic Setting*. Comm. ACM, 49 (8), 2006, pp. 53-58.
- [3] <http://www.easaier.org/>
- [4] <http://www.dismarc.org/>
- [5] <http://www.collectionscanada.ca/jsc/rda.html>
- [6] <http://www.musicaustralia.org/>
- [7] Yee, M. M. *FRBRization: A Method for Turning Online Public Finding Lists into Online Public Catalogs*. Information Technology and Libraries, 24 (2), 2005, pp. 77-95
- [8] <http://www.mmdb.iai.uni-bonn.de/>
- [9] Müller, M., Kurth, F., Clausen, M.: Audio Matching via Chroma-based Statistical Features. In: ISMIR, London, GB., 2005.
- [10] Kurth, F., Müller, M., Damm, D., Fremerey, C., Ribbrock, A., Clausen, M.: Sync Player — An Advanced System for Multimodal Music Access. In: ISMIR, London, GB, 2005.
- [11] Clausen, M., Kurth, F.: A Unified Approach to Content-Based and Fault Tolerant Music Recognition, IEEE Trans. Multimedia 6 (5), 2004.
- [12] <http://www.probado.de/>