Int J Digit Libr (2012) 12:53–71
DOI 10.1007/s00799-012-0087-y

A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction

David Damm · Christian Fremerey · Verena Thomas · Michael Clausen · Frank Kurth · Meinard Müller

Published online: 20 June 2012

Abstract In this paper, we present a digital library system for managing heterogeneous music collections. The heterogeneity refers to various document types and formats as well as to different modalities, e. g., CD-audio recordings, scanned sheet music, and lyrics. The system offers a full-fledged, widely automated document processing chain: digitization, indexing, annotation, access, and presentation. Our system is implemented as a generic and modular music repository based on a service-oriented software architecture. As a particular strength of our approach, the various documents representing aspects of a piece of music are jointly considered in *all* stages of the document processing chain. Our user interfaces allow for a multimodal and synchronized presentation of documents (WYSIWYH: what you see is what you hear), a score- or lyrics-based navigation in audio, as well as a cross- and multimodal retrieval. Hence, our music repository may be called a truly *cross-modal* library system. In our paper, we describe the system components, outline the techniques of the document processing chain, and illustrate the implemented functionalities for user interaction. We describe how the system is put into practice at the Bavarian State Library (BSB) Munich as a part of the German PROBADO Digital Library Initiative (PDLI).

Keywords Music digital library system · Multimodality · Cross-modal navigation · Content-based retrieval · Music synchronization · Music information retrieval

D. Damm (✉) · C. Fremerey · V. Thomas · M. Clausen
Department of Computer Science III, University of Bonn,
Römerstr. 164, 53117 Bonn, Germany
e-mail: damm@iai.uni-bonn.de

C. Fremerey
e-mail: fremerey@iai.uni-bonn.de

V. Thomas
e-mail: thomas@iai.uni-bonn.de

M. Clausen
e-mail: clausen@iai.uni-bonn.de

F. Kurth
Fraunhofer Institute for Communication,
Information Processing and Ergonomics (FKIE),
Neuenahrer Str. 20, 53343 Wachtberg, Germany
e-mail: frank.kurth@fkie.fraunhofer.de

M. Müller
Saarland University and Max-Planck-Institut für Informatik,
Campus E1 4, 66123 Saarbrücken, Germany
e-mail: meinard@mpi-inf.mpg.de

1 Introduction

Recently, significant digitization efforts have been carried out for large collections of multimedia content, including books, newspapers, images, music documents, and videos. This naturally leads to the need of powerful tools that automatically process, analyze, and annotate the scanned documents, providing the basis for efficient and effective content-based searching, navigation, and browsing in the digitized data. The mere digital capture of documents, i. e., digitization and storage, is not a problem—scanned documents can mostly be created and stored automatically. Especially for libraries holding a vast amount of musical content that is steadily increasing due to ongoing digitization, there is a high demand for automatisms to cope with the large number of documents. Two key challenges are how a growing collection can be organized automatically and how users can be enabled to access the documents intuitively.

In the case of scanned text documents, various solutions for automated document processing have been proposed, which typically contain a component for Optical Character Recognition (OCR) to extract the textual content from the

images, as well as a component for fault-tolerant full-text indexing and retrieval. The general idea is to suitably combine the strengths of both types of data representations, i. e., scan and text, for a convenient navigation and search in the scanned text documents. A well-known example for this is the Google Book Search project [26], where one can search and navigate in whole books.

In spite of these advances in the textual domain, there is still a significant lack of corresponding solutions for handling general digitized non-textual documents including musical data such as audio recordings and sheets of music or graphical data such as images, videos, and 3D data. In particular, tools are needed to automatically extract semantically meaningful entities or regions of interest from the scanned documents and to create links between related entities.

In this paper, we consider the domain of music. Increasing digitization of musical data of all kinds—comprising various documents of diverse types expressing musical content at different semantic levels and addressing different modalities—led to extensive and often unstructured music collections. In real-life application scenarios, these stocks are in general heterogeneous and contain visual, auditory, and textual content of different formats. CD-audio recordings and scanned sheets of music play a major role in ongoing automatic digitization efforts and constitute the main types of music representations considered in this paper. In addition, there exist other related documents such as libretti (lyrics), album cover arts and score-like documents in digital formats such as MIDI [31] or MusicXML [25]. We refer to [57] for a survey of symbolic music formats.

Beyond the mere recording and digitization of musical data, the key challenge in a real-life library application scenario is the automated processing of and the subsequent access to the musical data. On the one hand, methods for automatic annotation, linking, and indexing of different music documents are required. On the other hand, tools and user interfaces for a unified and adequate presentation of as well as navigation and search in documents that explicitly consider multiple available modalities have to be devised.

In our application scenario, where sheets of music as well as audio recordings belonging to the same piece of music are given, each of the documents represents the same musical content, however, using different modalities. Here, our strategy is to exploit the availability of those multimodal data to implement the particular tasks of (a) automatic annotation via cross-modal alignment, i. e., the spatio-temporal alignment or *synchronization* of scanned sheet music images and audio recordings, (b) cross-modal navigation, i. e., using the spatio-temporal alignment to synchronously navigate and present specific parts in both representations, (c) cross-modal search and retrieval, where a query is formulated in one modality while results may be of another modality, as well as (d) multimodal search, where queries may be composed using information from multiple modalities, like for example lyrics and score fragments.

Over the last decade, research efforts in the field of Music Information Retrieval (MIR) have produced various methods for automatic content-based analysis, synchronization, indexing, and retrieval of and navigation in music documents, see Sect. 2. In contrast, to this date there is a lack of suitable software frameworks that provide *integrated* solutions for use in real-world music digital libraries. Such frameworks should systematically exploit and combine available MIR techniques with appropriate tools and user interfaces for multimodal music access, including multimodal playback, cross-modal navigation, and cross- and multimodal search as well.

In this paper, we propose such an integrated approach to a music repository framework. As a main goal, we aim at processing digitized music documents *automatically* to a large extent. Particularly, we address the following tasks:

1. Digitization of music documents and creation of digital versions.
2. Indexing, annotation, and synchronization of digitized documents.
3. Design and development of a digital library software system considering specific library requirements.
4. Design and development of administrative user interfaces for data maintenance and quality assurance.
5. Design and development of user interfaces for document presentation, navigation, and searching.

Our proposed framework incorporates methods for automatically organizing large collections of digital music documents and exploits modern MIR techniques. It comprises a preprocessing workflow consisting of feature extraction, audio indexing, and music synchronization. A long-term goal was to develop a generalized workflow that allows libraries, specifically its personnel, to efficiently handle collections of digital music documents while minimizing the required administration effort for managing the documents. Considering user interaction, we offer novel and easily operated user interfaces for multimodal music presentation, i. e., audio-visual playback, cross-modal navigation, and cross-modal content-based search and retrieval. In the design of the user interfaces, we made sure that the access to as well as the interaction with music documents is as natural as possible, i. e., as one is accustomed to from their respective physical pendants. With this, we aim at bridging parts of the gap between the real and the digital world and to give users an intuitive way to handle and explore music documents. In the PROBADO Digital Library Initiative (PDLI) [36], both the preprocessing framework as well as the user interfaces are to be integrated into the library service system of the Bavarian State Library (BSB) Munich, that holds a vast amount of

music documents, particularly collections of scanned sheet music and CD-audio recordings with associated metadata.

The main contributions of our paper are summarized as follows:

- We propose a complete music digital library system comprising (a) a software system architecture, (b) a data model for organizing both heterogeneous document collections and associated metadata, as well as (c) a full-fledged workflow for content-based document processing.
- For the design of the proposed system, we identified several practice-relevant requirements and worked out practicable software solutions, guided by a real-world application scenario. (The system is currently installed at the BSB Munich as part of the PDLI.)
- As a fundamental paradigm of the proposed system, cross-modal document processing is exploited in all stages of the document processing chain.
- To facilitate cross- and multimodal retrieval, we propose an enhanced retrieval strategy offering composite queries which can be constructed by a weighted linear combination of a number of sub-queries, each of which may independently address an individual modality.

The paper is organized as follows. In Sect. 2 we give an overview on the different types of music representations. We then describe existing approaches in the field of music digital libraries and briefly summarize some of the relevant MIR technologies. Section 3 presents the software system architecture of the proposed music repository. As the first major part of this paper, Sect. 4 describes the steps of the document processing chain for cross-modal music processing. The second major part, Sect. 5, presents the user interfaces and functionalities for multimodal music access comprising presentation, navigation, and query-retrieval. Section 6 briefly summarizes the application scenario of our system at the BSB Munich within the PDLI. Concluding, Sect. 7 gives some prospects on future challenges and ongoing work.

2 Music documents and related work

2.1 Music representations

Music digital libraries contain textual data (e. g., libretti), symbolic data (e. g., MusicXML), visual data (e. g., scanned sheet music or CD album cover arts), audio data (e. g., audio recordings), and audiovisual data (e. g., video recordings from orchestral performances), as depicted in Fig. 1. Among these various types of information, music data pose many problems, since musical information is represented in diverse data formats. These formats, depending upon particular applications, differ fundamentally in their respective

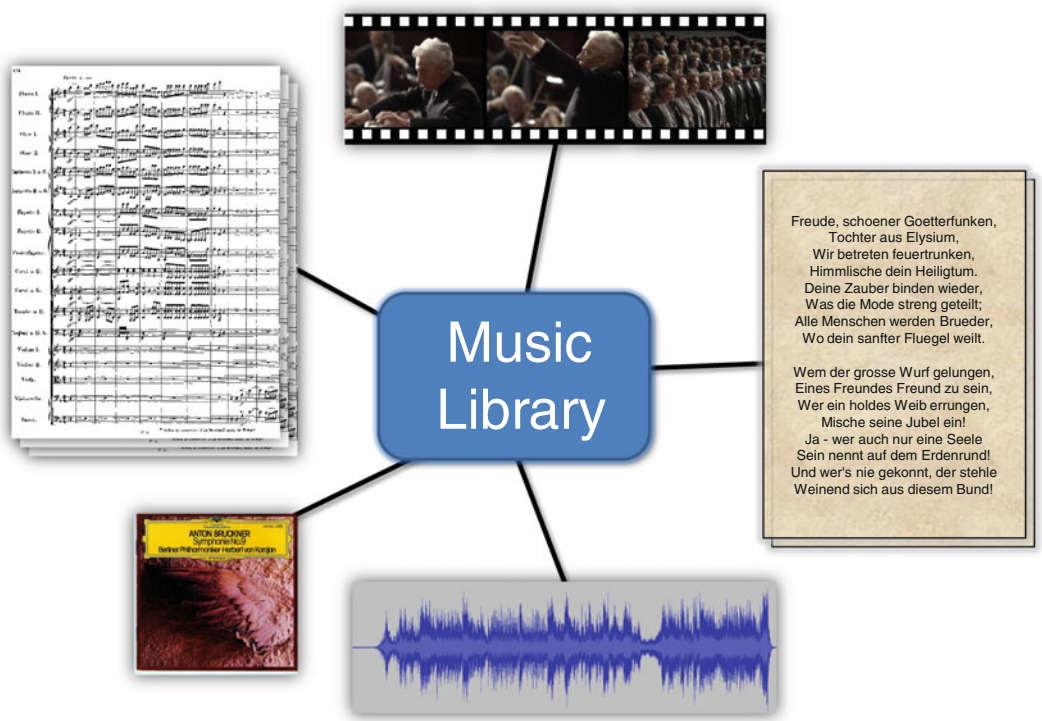

Fig. 1 Different document types typically available in a music library

structures and content. In this paper, we concentrate on three widely used formats for representing musical data: the symbolic *score format* contains information on the notes such as musical onset time, pitch, duration, and further hints concerning dynamics and agogics. The purely physical *audio format* encodes the waveform of an audio signal as used for CD-audio recordings. Finally, the *MIDI format* may be thought of as a hybrid of the last two data formats that explicitly represents content-based information such as note onsets and pitches but may also encode agogic and dynamic subtleties of some specific interpretation. While MIDI and audio data are available immediately in a digital format, symbolic score information is in most cases only indirectly available as scans of sheet music. In order to obtain this symbolic score information one uses Optical Music Recognition (OMR) software to extract the symbolic score information from the scans of sheet music. (However, this process is error-prone.)

The audio and the visual representations are most widely employed by users for accessing music. Thus corresponding multimodal user interfaces are of high importance. It turns out that the key challenge in designing such interfaces and in suitably preprocessing the underlying music documents is to find an appropriate common representation for the various music modes in order to compare and relate the musical content. Below we will discuss reductions of the various document types to a common representation.

To organize the various types of music representation, the Functional Requirements for Bibliographic Records (FRBR) model that has been introduced in the library community [32] has been adapted within our music repository framework. FRBR is a standardized, widely used conceptual entity-relationship model developed by the International Federation of Library Associations and Institutions (IFLA) and is built upon relationships between and among entities. It represents a more holistic approach to the retrieval and the access of entities as the relationships between the entities provide links to navigate through the hierarchy of relationships. In brief, the FRBR model considers the following entities, which describe intellectual and bibliographic units: (i) a musical work (in an abstract sense), (ii) an expression (of a musical work), (iii) a manifestation (of an expression), and (iv) an item (of an manifestation as physical object).

2.2 Background on existing music digital libraries

Over the past years, several digital library systems for music documents were developed that include printed music (e. g., sheet music and musicological books), and various systems are currently available [56,63,65,66,70].

Hankinson et al. [29] evaluated several of these music digital library systems with respect to their user interfaces and identified three main drawbacks that may be observed in most of the systems. First, the systems do not keep document integrity and present the documents as a series of separate images. Second, simultaneous presentation of related music documents is often not possible. As third drawback, the metadata of the currently selected music document can not be accessed at a glance, omitting further valuable information.

Besides those shortcomings, these systems restrict the user in the possibilities of experiencing a musical work. As mentioned before, a piece of music has various representations, describing it on different semantic levels and addressing different modalities. Therefore, a music digital library system should offer the access to as many different representations as possible. An example for a multimodal digital library system is the Europeana project [20]. Europeana offers open online access to a large collection of text, audio, video, and image documents of different European cultural institutions. The collection naturally also contains large amounts of music related documents. Further examples of multimodal general digital libraries are the Internet Archive [33] and the World Digital Library [64].

A further disadvantage of the music digital library systems mentioned so far is that nearly none of them allows for comparable content-based search functionalities as library systems for textual documents. They mostly are restricted to metadata search functionalities. However, there are various MIR techniques (see Sect. 2.3) available which would enhance the functionalities of a music digital library system by, e. g., allowing for a multimodal content-based search.

Two systems, already fulfilling most of the listed requirements for a music digital library system are Variations2 [19] and EASAIER [15,40]. Variations2 is a digital music system for educational institutions to share music collections throughout the class room. Besides user interfaces for the simultaneous visualization of metadata information, audio tracks, and sheet music, the system offers tools for manual music analysis (musical structure, musical beat). To offer enhanced search functionalities, a query-by-humming system is proposed [6,7]. EASAIER enables access and simultaneous visualization of various music representations like audio, score, and images. In addition to content-based search mechanisms, several different audio analysis and processing tools are available (e. g., time stretching and source separation). However, both projects do not incorporate MIR techniques for calculating structure information or cross-modal alignments between different representations of a piece of music.

The Greenstone project [72] is another interesting project in the context of multimodal digital libraries. In contrast to the projects mentioned so far, Greenstone aims at offering tools for the creation, management, and presentation of digital document collections. Some of the main features are the support of multimodal document collections, possibilities for content-based retrieval, a plug-in mechanism to add

functionalities, and a basic, extensible user interface. Furthermore, a tool for the creation of a digital library from a given digital document collection was proposed [4].

Recently, a new standard—IEEE 1599—to encode music with XML was published [41]. The new format offers the possibility to combine all information related to a musical work (different audio interpretations, scores, lyrics, images, annotations) in a single XML file. The standard also provides for the possibility of adding synchronization information [13] as well as MIR models [54] to the XML file. Using this standard, user interfaces for a holistic presentation of musical works using all information available were proposed [3].

2.3 Background on MIR techniques

As mentioned before, up to now the community brought up various approaches and solutions towards the automatic processing of musical data such as indexing methods and content-based retrieval. In particular, the indexing methods include different synchronization methods for the automatic linking of two data streams of different formats. Those methods include the linking of CD recordings with symbolic score-like formats such as MIDI [2,18,30,55,58,60], lyrics [69], or scanned sheet music [16,38,49]. Furthermore, different algorithms were developed to determine the structure and repetitions of representative parts of audio, see, e. g., [5,27,42,46,52]. In addition to traditional text-based methods for music search, based on annotations and symbolic music data [12,62], content-based search methods that work directly on the audio data have been proposed recently, see, e. g., [10,39], as well as cross-modal search, see, e. g., [14,21,53,59]. For an overview on the issues of the development of automated music data indexing, we refer to [34,43,51].

3 Software system architecture

The PROBADO Music Digital Library (PMDL) incorporates the storage of as well as the access to digital music documents. For preservation purposes, digital copies of available musical content held by the BSB such as audio recordings, sheet music and other music-related material are made. For indexing purposes, these digital copies are analyzed and annotated by recent state-of-the-art MIR techniques.

One key task is to build up content-based search indexes in order to search for, e. g., lyrics phrases or score and audio fragments. Another key task is to consolidate all available documents for the same piece of music and relate them among each other. Mapping- and synchronization-techniques are used to create alignments between meaningful entities within

- sheet music pages and time segments within audio recordings (score–audio synchronization),
- words and time segments of audio recordings (text–audio synchronization) and
- time segments of different audio recordings of the same piece of music (audio–audio synchronization).

A more detailed view on the topic of the extraction of meaningful entities from scanned sheet music and its mapping to audio recordings is given in [5,21,30]. The resulting benefits are content-based and cross-modal searching capabilities, synchronous, multimodal playback and visualization of pieces of music, as well as advanced cross-modal browsing capabilities. In particular, with score–audio synchronization, a user is on the one hand enabled to visually track within the sheet music representation the currently played back measures within a concrete audio recording he is listening to. On the other hand, the sheets of music can be used to change the playback position within the audio recording. Hence, interactions regarding one modality are reflected within the other modality. A text–audio synchronization allows for a karaoke-like application where the user can see what word is currently sung within the audio recording he listens to. Again, the linkage can be utilized to change the playback position to a specific word. An audio–audio synchronization enables a user to switch between different interpretations (audio-recordings) while maintaining the actual playback position in a musical sense, allowing to draw local comparisons between different interpretations belonging to the same piece of music.

All these functionalities are realized and implemented within a modular system as depicted in Fig. 2. The system is organized as classical three-tier architecture and consists of the *presentation layer*, the *server layer*, and the *repository layer* (left to right). Search indexes, annotations, and linking structures between different modalities are obtained in a preprocessing step which is carried out offline in the repository layer. The access to index structures and synchronization data as well as the delivery of musical content to the user takes place in the server layer. The presentation layer consists of user interface components for accessing musical content, especially content-based searching for musical content, navigation, and browsing within search results, as well as synchronized playback of audio and sheet music or lyrics. For each system interaction such as retrieving search results and accessing musical content, there is a dedicated module, referred to as *Query Engine*. The communication between the presentation and the server layer is provided by a service-oriented architecture (SOA) [67] and utilizes Simple Object Access Protocol (SOAP). SOAP is a network protocol for remote procedure calls and is used for the implementation of Web-Services. Human-readable XML messages are passed between instances on different machines using the

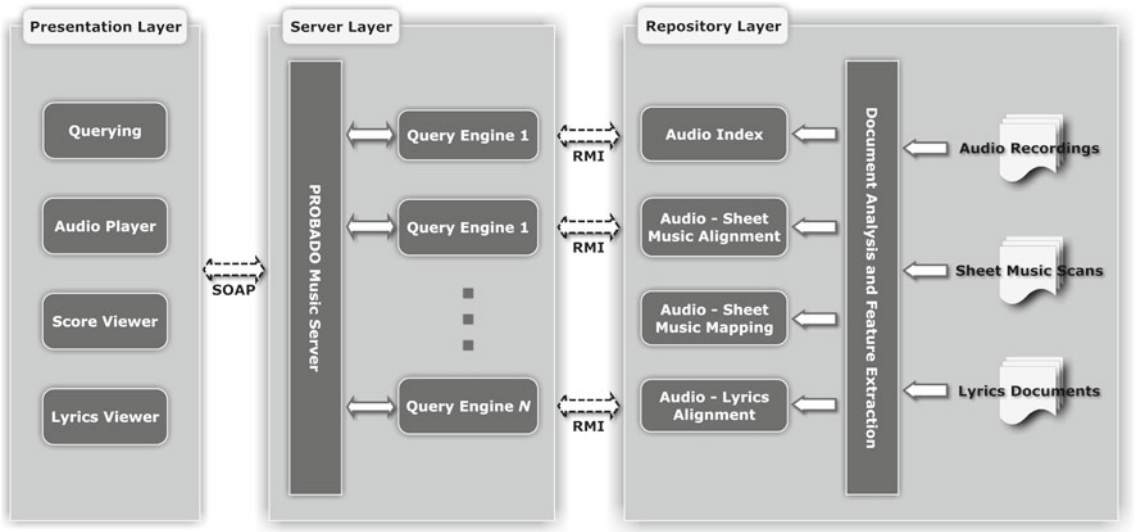

Fig. 2 Overview on the software system architecture of the PROBADO Music Digital Library

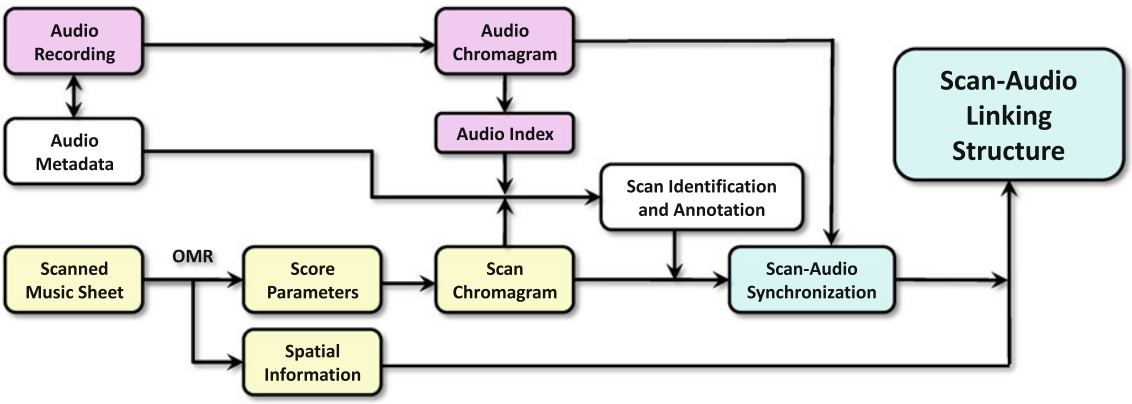

Fig. 3 Overview of the workflow for automatic cross-modal document processing. Two different modalities are considered, which concern the visual (scanned sheet music) and the acoustic (audio recordings) domain

Hypertext Transfer Protocol (HTTP). Remote Method Invocation (RMI) is a Java technology for network interactions between Java classes. There exists a name service, the RMI registry, and objects can call methods of registered remote objects and pass data to them.

4 Cross-modal music processing

In this section, we present the stages of the document processing chain including feature extraction, audio indexing, as well as music identification and synchronization. For an overview, we refer to Fig. 3. In order to compare and relate music data of various types and formats, the objective is to establish cross-modal linking structures that reveal the semantic correspondences of the musical events in the various data streams. To this end, the idea is to transform the various music representations into a common feature representation that allows for a direct comparison of the different types of data. In this context, chroma-based music features have turned out to be a powerful mid-level representation [5,30,43], which will be introduced in Sect. 4.1. In particular, we show how these features can be obtained from audio recordings using signal processing methods as well as from scanned sheet music using OMR.

The features extracted from the audio documents are further processed and suitably organized by means of an inverted file index structure (Sect. 4.2). This audio index can then be used for both identifying individual pages of scanned sheet music (Sect. 4.3) and for content-based music retrieval (Sect. 5). The identification task for sheet music consists of assigning each scanned page to a particular audio recording, see also Fig. 4. For each audio recording, we group the

corresponding pages of sheet music to establish a global correspondence between the audio and the sheet music data on the level of individual tracks, i. e., songs or movements. Finally, using the mid-level chroma representation and Dynamic Time Warping (DTW), the two representations are synchronized, which results in a structure that links the visual and the acoustic domain (Sect. 4.4), see also Fig. 5. This structure lays the foundation for a time-synchronous presentation of sheet music and audio recordings by means of the score viewer interface (Sect. 5.1).

4.1 Mid-level feature representation

To make the various music representations comparable, one needs to find a suitable mid-level feature representation that satisfies several critical requirements. On the one hand, such a feature representation has to be robust to semantic variations as well as transformation errors. Furthermore, the various types of data should be reducible to the same mid-level representation. On the other hand, the features have to be characteristic enough to capture distinctive musical aspects of the underlying piece of music. In the matching and synchronization context, *chroma-based music features* have turned out to achieve these requirements to a high degree for certain classes of music. Here, the twelve *chroma* correspond to the twelve traditional pitch classes of the equal-tempered scale [5]. In Western music notation, the chroma are commonly indicated by C, $C^{\#}$, ..., B consisting of the 12 pitch spelling attributes. Chroma-based features are well-known to reflect the phenomenon that human perception of pitch is periodic in the sense that two pitches are perceived as similar by the human auditory system (HAS) if they differ by one or more octaves [5].

For an audio recording, the digitized signal is transformed into a sequence of normalized 12-dimensional chroma vectors, where each vector reveals the local energy distribution among the 12 pitch classes. Based on signal processing techniques, a chroma representation can be obtained either using short-time Fourier analysis in combination with binning strategies [5] or using multirate filter bank techniques [43]. Such a representation, in the following also referred to as *audio chromagram*, absorbs variations in parameters such as dynamics, timbre and articulation and corresponds to the rough harmony progression of the underlying audio signal. Figure 5c shows an audio chromagram for the first few measures of an audio recording of the third movement of Beethoven's Piano Sonata Op. 13 ("Pathethique").

The transition from a sheet music representation to a chroma representation consists of several steps. In the first step, musical score symbols such as notes, clefs, key signatures, and time signatures are extracted using OMR, see [9,11]. This process is similar to the well-known OCR, where textual content is extracted from a scanned image of a text document. Note that the OMR extraction step is error-prone and the recognition accuracy strongly depends on the quality of the input image data as well as the complexity of the underlying score. In the context of this paper, we consider high quality scans of sheet music at a resolution of 600 DPI and 1 bit color depth (b/w). In addition to the musical score symbols, the OMR process also provides spatial information. In particular, the exact pixel coordinates of the extracted notes as well as bar line information are available. This allows for localizing all extracted musical symbols within the sheet music.

In the second step, based on the OMR output, we derive a sequence of normalized 12-dimensional chroma vectors, which is also referred to as *scan chromagram*. To this end, we create note events specified by musical onset times, pitches and note durations from the extracted musical symbols. Assuming a constant tempo of 100 BPM, the explicit pitch and timing information can be used to derive a chromagram essentially by identifying pitches that belong to the same chroma class. To this end, we slide across the time axis with a temporal window while adding energy to the chroma bands that correspond to pitches that are active during the current temporal window. Here, a single temporal window equals a single chroma vector. A similar approach has been proposed in [30] for transforming MIDI data into a chroma representation. Note that the particular choice of 100 BPM in our assumption is not crucial, because differences in tempo will be compensated in the subsequent matching and synchronization steps. Figure 5b shows a scan chromagram obtained from a sheet music representation for our "Pathethique" example.

Both the identification of scanned sheet music pages and content-based audio retrieval rely on a mechanism for efficient *audio matching* [39]. Here, given a short query music clip in form of an excerpt taken from an audio recording or in form of some bars of music taken from scanned sheet music, the goal is to automatically retrieve all excerpts that musically correspond to the query from an audio database. As opposed to classical audio identification [1], audio matching allows for semantically motivated variations as they typically occur in different interpretations of a piece of music. The methods for audio matching introduced in [39] work on the basis of chroma representations. As has been shown recently, the chroma features generated from symbolic music representations, e. g., those obtained by the above OMR process, are compatible with audio chromagrams. Therefore, chroma features can be used to perform both audio matching [37] and synchronization of music documents *across* the domains of symbolic music and audio recordings [21].

4.2 Audio indexing and matching

As mentioned before, the key idea we exploit for automatic document analysis is to reduce the two different types of data (visual and acoustic music data) to the same type of representation (chromagram), which then allows for a *direct* comparison *across* the two modalities on the feature level. To also allow for an *efficient* comparison, we further process the chroma features by quantizing the chroma vectors using semantically meaningful codebook vectors as described in [39]. According to the assigned codebook vectors, the features can then be stored in some inverted file index, which is a well-known index structure frequently used in standard text retrieval [71].

In our system, we employ audio matching as described in [39] as an underlying engine for the various music retrieval and identification tasks. The basic matching approach works as follows. Each music document of the repository is converted into a sequence of 12-dimensional chroma vectors. In our implementation, we use a feature sampling rate of 1 Hz. While keeping book on document boundaries, all these chroma sequences are concatenated into a single sequence $(d_0, \ldots, d_{K-1})$ of chroma features. Similarly, a given query music clip is also transformed into a sequence $(q_0, \ldots, q_{L-1})$ of chroma features. This query sequence is then compared to all subsequences $(d_k, d_{k+1}, \ldots, d_{k+L-1})$, $k \in [0 : K - L]$, consisting of L consecutive vectors of the database sequence. Here, we use the distance measure $\Delta(k) := 1 - \frac{1}{L} \sum_{\ell=0}^{L-1} \langle d_{k+\ell}, q_\ell \rangle$, where the brackets denote the inner vector product. The resulting curve Δ is referred to as *matching curve*. Note that the local minima of Δ close to zero correspond to database subsequences that are similar to the query sequence. Those subsequences will constitute the desired *matches* for content-based retrieval, see Sect. 5. Because of the bookkeeping, both document numbers and exact positions of matches within each document can be easily recovered.

So far we have not yet accounted for possible temporal differences between the query clip and corresponding temporal regions within the audio documents. For example, interpretations of the same piece of music often reveal significant local and global differences in tempo as a result of the freedom a musician has in interpreting a piece. Also, when converting a score representation into a feature representation, one needs to assume a tempo that may deviate from a corresponding audio recording (see, e. g., our assumption in Sect. 4.1). To handle such tempo deviations, one can revert to subsequence variants of DTW [43], or one can employ the technique of multiple querying with various chromagrams at different sampling rates [39]. In particular, the latter technique can be supported by the above mentioned index structure facilitating an efficient computation of the audio matches. For the technical details, we refer to [39].

4.3 Identification and annotation of scanned sheet music

After the digitization process, the digitized documents need to be suitably annotated before they can be integrated into the holding of a digital library. In the case of digitized audio recordings, one has to assign metadata such as *title*, *artist*, or *lyrics* to each individual recording. Besides the labor and cost intensive option of manual annotation, one may exploit several available databases that specialize on various types of metadata such as Gracenote [28] or DE-PARCON [35]. Note that in spite of the existence of such databases, it may *not* in general be assumed that the acquisition of metadata is a trivial task because existing databases are frequently incomplete with respect to old recordings, they lack particular types of requested metadata, or contain errors and inconsistencies. An improvement may be achieved by, e. g., tapping and merging multiple data sources. However, this is out of the scope of this work and we rely on the existence of sufficient metadata. This is not a serious restriction as an improvement of both the quantity and quality of metadata is to be expected over the time due to the fact that the supply of high-quality metadata is the business concept of some commercial providers (e. g., Shazam [68]). Ultimately, the libraries mostly do have high-quality metadata and are responsible for the quality assurance of them. For the purpose of this paper, we assume that suitable annotations for the audio recordings are readily available, see Fig. 3.

After the digitization of scanned sheet music—a process that can be done by a scanner with automatic book page turner—each page has to be annotated separately.

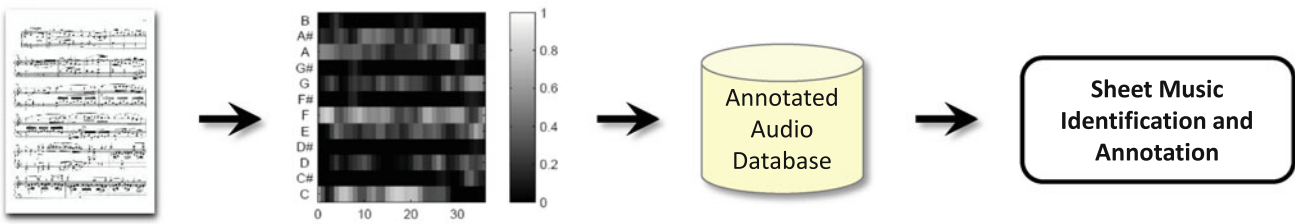

Fig. 4 Overview of the matching procedure for automatic identification and annotation of scanned sheet music using an annotated audio database. The first page of the second movement of Beethoven's piano sonata Op. 2 No. 1 and the resulting scan chromagram are shown

This annotation is usually done in a manual process. Following [21], we now describe how this annotation process can be performed automatically, see also Fig. 4. In our scenario, we assume the existence of an audio database containing annotated digitized audio recordings for all pieces to be considered in the sheet music digitization process. In a preprocessing step, we transform the audio documents into corresponding audio chromagrams and build up an audio index structure. Then, in the annotation step, each scanned page of the sheet music is converted into a separate scan chromagram. Using each scan chromagram as query, we compute the top match within the audio documents as described in Sect. 4.2. Here, we assume that each page is fully contained in a single audio document. Note that this assumption does not generally hold, since a single page may refer to several short pieces or may contain the end and the start of two consecutive movements corresponding to different audio documents. Under our assumption, the top match usually identifies the musically corresponding audio recording with high reliability. As our experiments show, this particularly holds in the case that there are no severe OMR errors. Upon identification, the scanned page can then be automatically annotated by the metadata already associated to the corresponding audio recording, see Fig. 4. Furthermore, the first few top matches usually consist of all passages within the audio recordings that musically correspond to the page. This additional property is exploited for the retrieval and browsing applications as described in Sect. 5.

The content-based comparison of sheet music pages to audio recordings used to establish a correspondence on the level of individual tracks can be supported using various strategies. Firstly, the score is searched for indented grand staffs. Such indentations usually indicate the beginning of a new movement or musical work. Using this information, the scan chromagrams created from pages including such an indentation can be divided at the beginning of the indented grand staff to account for the expected track change. Secondly, title headings that have been recognized in the scores may be used as indicators for the beginning of movements and musical sections as well. Furthermore, the recognized text of these headings can be compared to the known titles of the tracks in the audio database.

Using suitable heuristics, some of the OMR extraction errors can be corrected in a post-processing step prior to the matching step. For example, in the case of piano music, different key signatures for the left and right hand staves can be assumed to be invalid and easily corrected by considering neighboring stave lines. Furthermore, similar to the strategy suggested in [9], one can simultaneously employ various OMR extraction results obtained from different OMR software packages to stabilize the matching result. First experiments show that, based on these strategies, one can achieve a significant improvement of the identification rates.

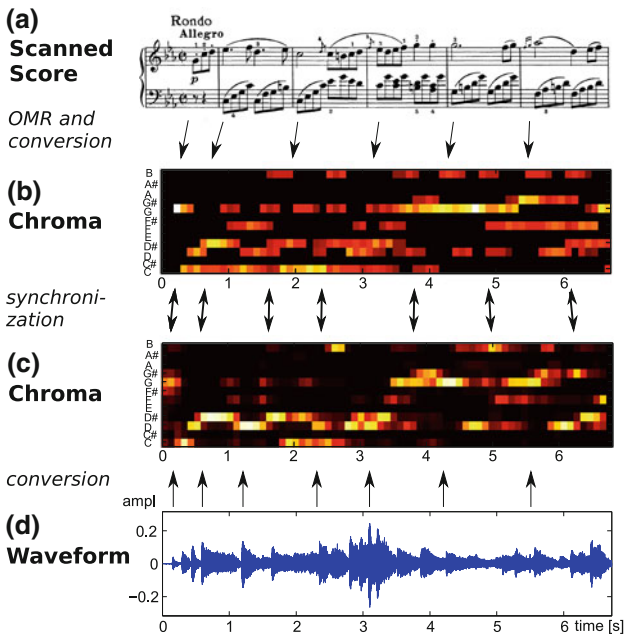

Fig. 5 Data types involved in automatic document processing for the first few measures of Beethoven's Piano Sonata Op. 13 "Pathethique", Rondo (3rd movement): **a** scanned sheet music, **b** scan chromagram, **c** audio chromagram and **d** audio recording (waveform). The scan–audio linking structure (*double-headed arrows*) is obtained by synchronizing the two chromagrams

4.4 Scan-audio synchronization

Once having identified scanned pages of sheet music and corresponding audio recordings, we automatically link semantically related note events across the two types of music representation. In the general context, various alignment and synchronization procedures have been proposed with the common goal to automatically link several types of music representations, thus coordinating the multiple information sources related to a given musical work [2,30,43,47,50,58,61]. In our specific scenario, the problem is referred to as *scan–audio synchronization*, where the objective is to link regions (given as pixel coordinates) within the scanned images of given sheet music to semantically corresponding time positions within an audio recording. Such a procedure has been described in [38].

The basic idea is to convert both the scanned sheet music of a given piece as well as the corresponding audio recording into sequences of chroma features. The resulting scan chromagram and audio chromagram are then synchronized based on standard alignment techniques such as DTW [43]. More precisely, one builds up a cost matrix by computing the pairwise distance between each scan chroma vector and each audio chroma vector. Here, one needs a suitable local distance measure for determining the distance between two chroma vectors. In our implementation, we use the cosine measure

between normalized chroma vectors. Then, an optimum-cost alignment path is determined from this matrix via dynamic programming (DP). In order to handle global tuning shifts in the audio recordings, *chroma cyclic shifting* is performed; see [45] for details. The resulting path through the matrix encodes a temporal alignment of the two chroma sequences. For details on DTW, we refer to the literature [43]. Now, the spatial information of the OMR output allows for assigning each scan chroma vector to a corresponding region within the scanned sheet music image. Combining this spatial information with the scan–audio synchronization result, one can then derive a linking structure between the scanned images and the audio recording. An example of the discussed scan–audio synchronization is shown in Fig. 5, where the resulting linking structure is indicated by the double-headed arrows. The importance of such linking structures has been emphasized in the literature [19]. In Sect. 5, we will introduce user interfaces that exploit the scan-audio alignments in order to facilitate cross-modal music presentation and navigation.

We conclude this section by discussing some challenges that arise in the music synchronization context and give rise to future research. As mentioned in [23], the quality of the resulting synchronization depends on several factors. One of these factors are structural differences such as missing or additional repeats of musical sections. For example, the score might contain a section that is not played in the audio recording or the audio recording might contain an extra repeat that is either not present or not recognized in the score. Such differences in structure may be caused by OMR errors or stem from the fact that a performance is not required to strictly follow the structure suggested by the musical score. Furthermore, for a given audio recording, it is not guaranteed that the recorded performance is actually based on the particular score edition that is to be synchronized with. Differences in structure violate the boundary and/or monotonicity assumptions made in DTW, see [43]. Such differences may be handled in a manual preprocessing step or by partial matching strategies [44]. In [24], a novel variant of DTW is introduced, referred to as JumpDTW, which allows jumps and repeats in the alignment and significantly improves synchronization results. However, the general problem of partial similarities between music representations to be synchronized still pose many open research problems. Further dissimilarities of more local nature are musical events in the audio and sheet music representations with deviating pitch or duration. Problematic are also note ambiguities in the score such as arpeggios, trills, grace notes, or other ornaments. Generally, differences of this class tend to have little impact on the overall synchronization result as long as they stay local and are enclosed by sections without mismatches and errors. Significant differences in tempo may also cause problems in the synchronization procedure. Recall that for computing a mid-level representation from a symbolic music representation, one needs to decide on the tempo to be used in the transformation from musical onset times like beats and bars to physical onset times like seconds and milliseconds. Since tempo directives of music notation are often not output by OMR systems, the tempo then has to be guessed or estimated. For classical music, the tempo can vary over a wide range from about 25 to 200 BPM. Differences between the estimated tempo and the actual tempo of the audio recording are usually handled by the DTW-based alignment strategy. However, DTW starts to loose flexibility and accuracy when the tempo differences become too large. For a detailed examination of such issues, we refer to [22].

The practical impact of the issues discussed above are heavily data-dependent and unforeseeable in general. This holds especially in the case of great structural differences between the score and associated audio recordings, combined with other issues such as a high degree of polyphony and pitch and tempo deviations. In the worst case, synchronization results may become so degenerate that large parts of the calculated positions do not match the actual positions (in a musical sense). However, serious errors affecting the synchronization accuracy occur only rarely for large parts of the data set processed in this work. We experienced a small degradation with an increasing degree of polyphony. While synchronization results for piano sonatas are virtually flawless, some of the more complex orchestral works need to be corrected manually before the synchronization is calculated.

5 User interfaces for multimodal music access

In this section, we present the user interfaces for accessing musical content in a cross- and multimodal way. One of our key contributions is the multimodal presentation of and the cross-modal navigation in music documents. For this purpose, a special document viewer, presented in Sect. 5.1, provides various views and modes the user can operate on music documents (cf. Figs. 6, 7, and 8). In particular, it first provides the simultaneous, synchronized playback of audio recordings and various visualizations, including scores, lyrics and videos. Second, it allows for a cross-modal navigation in music documents. For this, the document viewer utilizes synchronization data that links (connects) semantically meaningful musical entities of one representation to corresponding ones of another representation belonging to the same piece of music across different modalities; e. g., image regions within scanned sheets of music and time segments within the audio recording. In the case of sheet music and audio recordings, the user is on the one hand enabled to track visually the currently played measure of a piece of music within its sheet music representation while listening to an associated audio recording. On the other hand, he has the option to navigate through the sheets of

music and select a specific measure in order to change the playback position within the audio recording accordingly. This can be useful, since the sheet music representation gives a more suitable possibility to search for specific parts within a piece of music than winding in the audio recording.

As another key contribution, we propose how to incorporate a concept for combined multimodal queries into the typical stages of a query-retrieval chain, particularly query formulation (Sect. 5.2), content-based retrieval and ranking (Sect. 5.3), presentation of query results (Sect. 5.4), and mechanisms for user feedback and navigation (Sect. 5.5). For this purpose, our system enables the user to formulate a query that may consist of different modalities, including the textual, visual and auditory modality. In particular, he is enabled to query a combination of metadata, lyrics and audio fragments. For this, he formulates single, unimodal queries and adds them successively to his search. These queries are gathered in a special structure for representing sets of queries, referred to as *Query Bag*. After the Query Bag is submitted to the retrieval system, the user is presented a list of pieces of music that match in at least one modality regarding his query. To organize the result list we employ a ranking approach, introduced in [14], based on a combination of multiple result lists ensuring that pieces of music containing more matching modalities are given a higher rank, see Sect. 5.3. The results, i. e., pieces of music, may be viewed in detail and played back in the document viewer module. Additionally, both the result list and the document viewer can be used for querying, query refinement and document navigation, see Sects. 5.2 and 5.5.

To give a user-friendly and intuitively operable interaction environment, our approach was to incorporate the look-and-feel of well-tried Internet search platforms that a wide range of users is familiar with. The user interface for the retrieval, browsing, playback and navigation in musical content is completely web-based and runs in virtually every state-of-the-art JavaScript/Java-enabled Internet browser. Figure 9 shows a snapshot of a typical system configuration. Similar to popular existing query engines, the top part of the user interface contains the query formulation area while the result view area is located below. Both areas are further subdivided to facilitate the subsequently described functionalities. The query formulation area is split into both a tab cards region and the Query Bag region. The result area is divided into the result list pane and the document viewer.

In the following, we give a detailed view on the document viewer and its particular audio-visual playback and navigation capabilities. Subsequently, a further in-detail look at each particular stage of the query-retrieval chain is given. Throughout the whole section, the piece of music "Gefrorne Tränen" belonging to the song cycle "Winterreise" by Franz Schubert will serve as our running example.

5.1 Multimodal music presentation and navigation

In this subsection, we give a detailed view on the document viewer, the central component for multimodal music presentation and navigation. The document viewer allows for the simultaneous, synchronized playback of musical content associated to a currently selected piece of music, including audio recordings, sheets of music, lyrics and videos. More precisely, besides the playback of audio recordings, it provides three visualization modes, including a score, a lyrics and a video visualization. For example, while an audio recording is played back, available sheets of music or lyrics are displayed synchronously; i. e., the user can visually track the currently played measure or the currently sung words, respectively, within the audio recording (cf. Figs. 6 and 7, respectively). Due to this style of enjoying music in a multimodal way, the document viewer may be thought of as being a video player, but in addition it provides sophisticated user interaction options such as navigation and query refinement, which are examined in Sect. 5.5.

The document viewer is divided into three areas, the top, the center, and the bottom area. The top area consists of tab cards, from where the various visualization modes can be chosen, the title of the currently selected piece of music, as well as buttons for exchanging score books or audio recordings, respectively, used for the audio-visual playback; cf., e. g., the tab cards depicted in the upper left corner of Fig. 6. For each modality except the auditory one, there exists a designated tab card to the left of the top area. In case that more than one document per modality is available, the user can freely exchange which documents are used for the audio-visual presentation of a piece of music. This can be achieved by clicking either the score book icon or the album cover art icon, respectively, whereafter a corresponding pop-up menu lists all available contents associated to the piece of music, from where the user can choose which audio or visual content, respectively, is used for playback. For example, if different audio recordings of a piece of music are available, the user has the choice to decide which specific performance he wants to listen to. With this functionality, he is also allowed to switch between different performances while retaining the actual musical playback position. Thus, the user can additionally draw local comparisons between different interpretations of a piece of music. Sheet music books may also be exchanged, if more than one is available. In the center area, the various visualizations are displayed. Here, the user is presented either the score view, the lyrics view or the video view, depending on the currently selected visualization mode, i. e., a specific visualization can be activated by clicking its appropriate tab card in the top area. The bottom shows a timeline bar that enables the user to adjust the playback position by moving the slider knob. Below the timeline bar, there are further buttons to control the playback state (start/pause, stop) as

well as the playback position. While the control buttons retain their positions, the labels are exchanged depending on the currently selected visualization mode (cf. Figs. 6, 7, and 8).

5.1.1 Score visualization mode

In the score visualization mode, the document viewer presents sheet music and associated audio recordings to the user, as illustrated in Fig. 6. Here, a virtual score book showing two pages of our running example, is shown. When starting the playback of an audio recording, corresponding measures within the sheet music are synchronously highlighted by a colored rectangle. When reaching the end of an odd-numbered page during playback, the page is turned over automatically, if score-following is enabled (cf. the "Score Following On/Off" button displayed in the upper right). For navigation purposes, the user may thumb easily through the sheet music book by pointing and clicking on the edge of an individual page. When scrolling through the score book, the score-following is disabled—otherwise it would scroll back to the page containing the currently played measure. Furthermore, a convenient way of changing the playback position of the audio recording is to select a specific measure within the sheet music. By clicking on a measure, the playback position is changed accordingly to where the measure begins to play. Additionally, the score visualization supports smooth zooming, allowing the user to view the sheets of music from near or from far away. Besides those presentation and navigation features, the score of an underlying piece of music, particularly consecutive measures, can also be used for content-based retrieval purposes. This characteristic is considered further in Sect. 5.5.

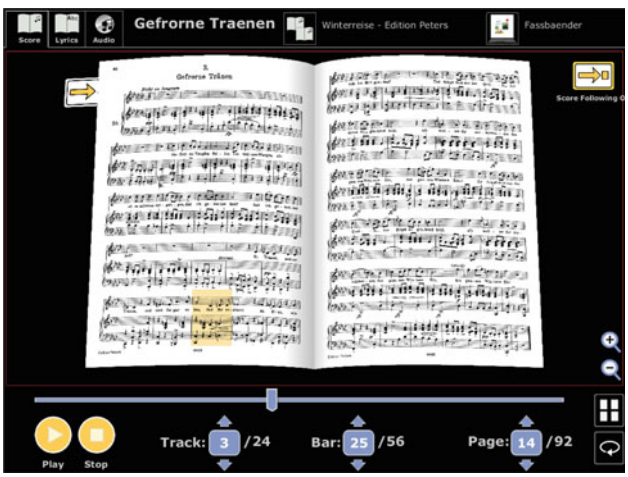

Fig. 6 Document viewer in score visualization mode. A scorebook is displayed in which the current measure is *highlighted*. The user can alternatively thumb through the pages

5.1.2 Lyrics visualization mode

In the lyrics visualization mode, the document viewer presents lyrics—in the context of classical music also referred to as libretti—and associated audio recordings to the user, as illustrated in Fig. 7. Here, a textual excerpt of the lyrics of our running example is displayed in a teleprompter-like manner. When starting the playback of an audio recording and words are sung within the latter, corresponding words within the lyrics text are synchronously highlighted by displaying them underlined and in a special color, similar to the score visualization. Again, similar to the score visualization, the lyrics text scrolls appropriately during playback, if text-following is enabled (cf. the "Following On/Off" button displayed in the upper right). For navigation purposes, the user may scroll the text by using the vertically aligned scrollbar to the right of the text. When scrolling the text, the text-following is disabled, otherwise the text-pane would jump back and display the one section that contains the currently sung (highlighted) word. Furthermore, a convenient way of changing the playback position of the audio recording is to select a specific word within the lyrics text. By clicking on a word, the playback position is changed accordingly to where the word is sung. Additionally, the text size can be adjusted. Besides those presentation and navigation features, the lyrics of an underlying piece of music, particularly consecutive words, can also be used for text retrieval purposes. This characteristic is considered further in Sect. 5.5.

5.1.3 Video visualization mode

In the video visualization mode, the document viewer acts like a video player, as illustrated in Fig. 8. Here, the moving images of a video performance of our running example

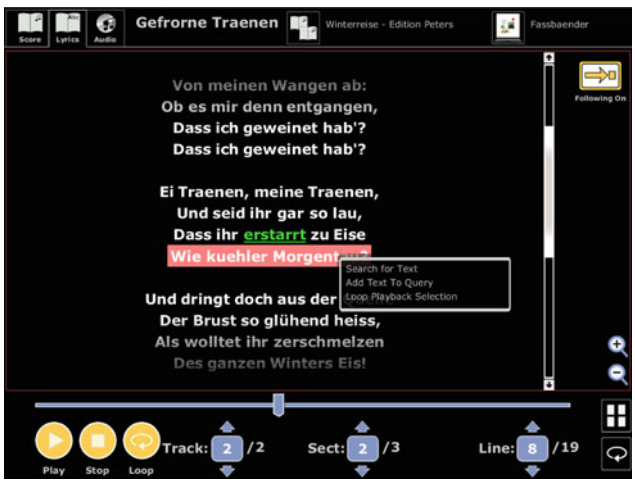

Fig. 7 Document viewer in the lyrics visualization mode. Selected text can be queried

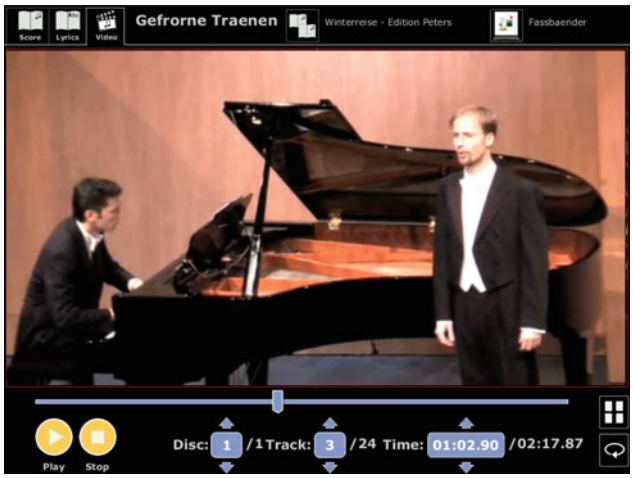

Fig. 8 Document viewer in video playback mode. A video recording of a performance of the selected song is shown

are played back. For navigation purposes, the user can orient himself on the moving images in order to find specific parts of the currently loaded piece of music. Until now, the video visualization does not support retrieval capabilities.

5.2 Multimodal query formulation and interface

This subsection gives some more detail on the provided querying functionality. As mentioned before, a key task in the context of the PMDL is to enable content-based search using lyrics phrases or score fragments as queries. Due to the consolidation of all musical content belonging to the same piece of music, each content-based search may also be viewed as cross-modal. This is, one can use either of the visual or textual modalities as queries, while aiming to find matches in the other modality. Up to now, two distinct options for content-based querying are available, *lyrics-based retrieval* as proposed in [48] and *score-based retrieval* using audio matching as proposed in [39]. Both approaches use indexing techniques to achieve a high retrieval efficiency. In the following, we firstly give a brief overview on the raised querying options. Subsequently, a detailed look at the query formulation and its interface is given.

The lyrics-based retrieval allows for formulating a query in the textual modality by entering a few words in order to find the positions within audio recordings where the words

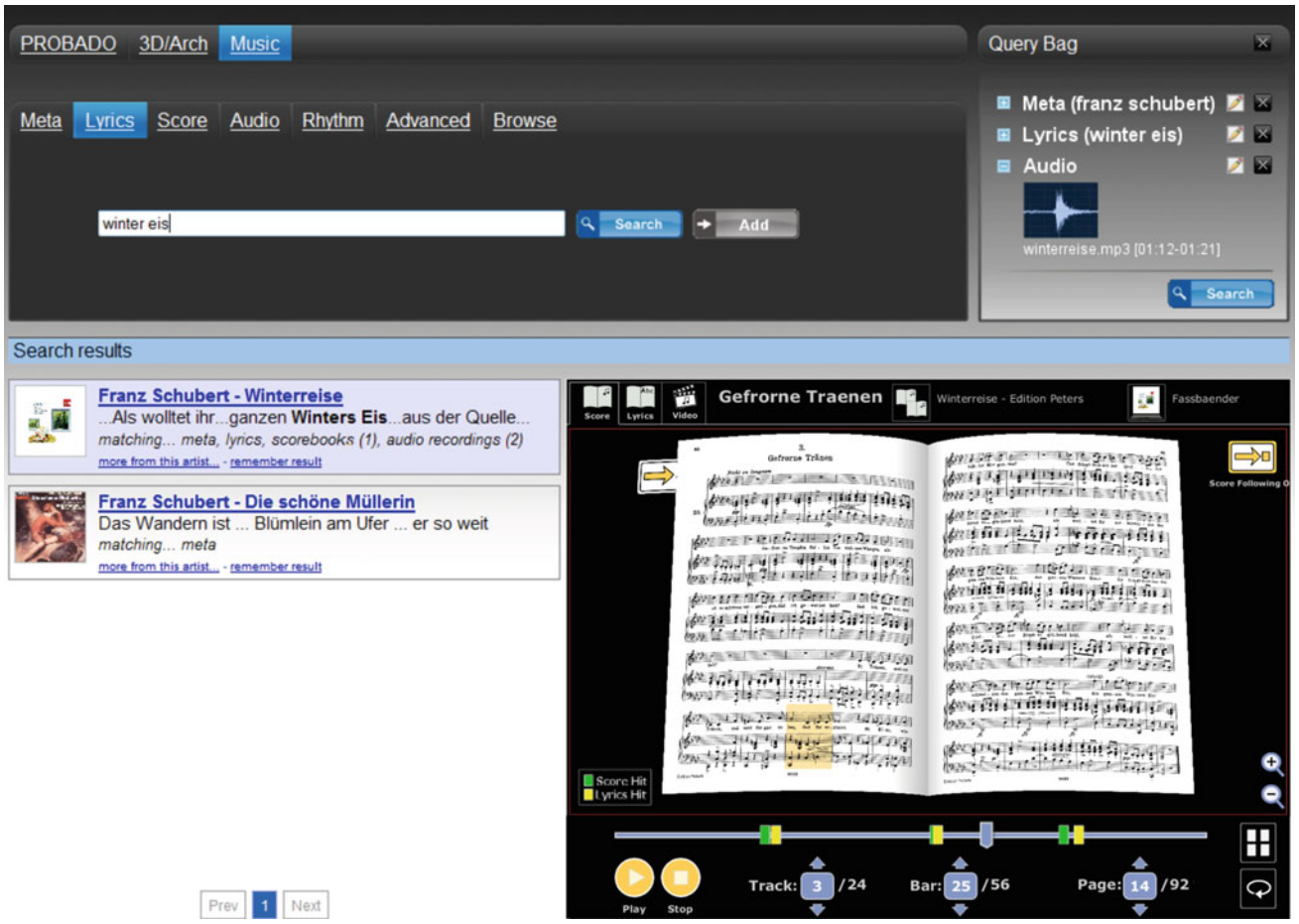

Fig. 9 Web-based user interface for query formulation (*top left*), Query Bag (*top right*), display of the results (*bottom left*) and the document viewer (*bottom right*)

are sung. The mapping of positions within the lyrics text document to time segments within an audio recording is performed using lyrics-enriched MIDI files as described in [48]. Here, onset times of individual words or syllables are explicitly given within a musical context. This information, in turn, is then used to synchronize the lyrics to the audio recording. The subsequently used indexing technique is based on inverted files which are well known from classical full-text retrieval [71] and enhanced for the special case of a lyrics search. The search is fault-tolerant w.r.t. misspelled or omitted words in both the query as well as the lyrics, see [48].

The score-based retrieval follows the query-by-example paradigm. A query is formulated in the visual modality by selecting a portion of a sheet music page, particularly a few consecutive measures. The system retrieves all occurrences of the selected music excerpt within the indexed audio recordings. Note that the sheets of music are images obtained from scanned analogue pages and thus the actual musical content or semantic is expressed in the visual modality. Exploiting the previously described synchronization, instead of querying the selected score excerpt, the according snippet of an associated (synchronized) audio recording is used for the search process. Here, a sequence of audio features is extracted from the snippet and subsequently a feature-based search on an audio features index is performed. Due to the extraction of consecutive features that reflect the chromatic harmonic progression of the underlying audio snippet at a coarse level, the audio retrieval system is robust against changes in timbre, instrumentation, loudness and transposition and therefore musically similar snippets can be found regardless of a particular performance [5,30]. For a more detailed view, we refer to [38] and the references therein.

Now we want to turn towards the query formulation interface. The query formulation area, shown in Fig. 9 (top), consists of various query formulation forms per modality which are organized as tab cards (top left). It further contains the Query Bag (top right), where single queries can be added to, viewed, revised or removed. Currently, the user is enabled to formulate metadata, lyrics and audio fragment queries within the appropriate, designated tab card. Additional query formulation types such as entering a melody by a virtual piano, humming a melody or tapping a rhythm are planned to be integrated in the future.

From within any tab card the user has the choice to either perform an immediate, unimodal search using the just formulated query (classical query) or to add the latter to the Query Bag and continue with the formulation of another query in order to gather a couple of unimodal queries. The Query Bag stores all queries and offers an overview representation of all gathered queries. So, the user at any time is informed about which queries he has collected so far. Each single query inside the Query Bag can be examined more precisely by clicking the plus-sign icon left to the query. To the right of each query there are icons for reformulating the query and for removing it from the Query Bag as well, by clicking either the pencil- or the "x"-icon, respectively. By clicking the pencil icon, the corresponding query formulation tab opens for editing. Once the user has finished assembling the individual queries, the search button at the bottom of the Query Bag can be clicked in order to submit them to the search engine as one integrated, multimodal query. Subsequently, a multimodal search is performed.

5.3 Multimodal content-based retrieval and ranking

Once the Query Bag is submitted, the system disassembles it and delegates each contained single query to an appropriate query engine which is capable of handling the particular type of query. The query engines act independently from each other and for each modality a homogeneous list of matches is returned. In this, each match consists of a document ID, the position of the matching segment, and a ranking value $r \in [0, 1]$. In the case of content-based queries, the latter segments are generally short parts of the document. If, however, a document matches due to its metadata description, the document is said to match at every position; i. e., a matching segment ranges from the beginning to the end of the document.

Due to the synchronization of different document types such as audio recordings, sheets of music and lyrics documents, all matching segment boundaries can be expressed in the time domain, i. e., translated to a start timestamp and an end timestamp. Thus, all segments are directly comparable, which will be exploited in the subsequent combined ranking and merging. For merging and ranking of multiple result lists returned by the different query engines into a single, integrated result list we use a straight-forward bottom-up approach explained in the following.

Each result list returned by a query engine consists of document IDs and for each document ID there exists a list of matching segments. These segment lists are inserted into a hashtable, where a single data entry stores a piece of music's ID together with related segment lists. For each inserted segment list, the respective modality is stored as well. With this, all inserted segment lists associated to the same piece of music are clustered and stored within a single hashtable data entry. Subsequently, for each entry of the hashtable a merging of the contained segment lists is performed. This step is now described in detail. Let M be the global number of queried modalities and m the local number of non-empty segment lists stored in a currently considered hashtable entry.

We now consider the merging step of two segment lists, $L^1 := \{(b_1^1, e_1^1, r_1^1), \ldots, (b_{|L^1|}^1, e_{|L^1|}^1, r_{|L^1|}^1)\}$ and $L^2 := \{(b_1^2, e_1^2, r_1^2), \ldots, (b_{|L^2|}^2, e_{|L^2|}^2, r_{|L^2|}^2)\}$, where the ith entry of list k is a segment $s_i^k = (b_i^k, e_i^k, r_i^k)$ consisting of a start

timestamp b_i^k and an end timestamp e_i^k as well as a ranking value r_i^k. In this, we assume that each segment list corresponding to a single modality does only contain non-overlapping segments. Let L be the merged, integrated segment list. For merging two lists L^k and L^l into L, we consider two cases. If a segment s_i^k does not overlap in time with any segment s_j^l of the other list, s_i^k is simply copied to L. If there is a temporal overlapping of a segment $s_i^k := (b_i^k, e_i^k, r_i^k) \in L^k$ and a segment $s_j^l := (b_j^l, e_j^l, r_j^l) \in L^l$, s_i^k and s_j^l are merged into a new segment $s := (\min(b_i^k, b_j^l), \max(e_i^k, e_j^l), r)$ which is inserted into L. Note that overlaps do reflect simultaneously arising hits and for this reason, we want them to get higher ranked in general. To additionally promote small individual ranking values r_i^k, r_j^l in the latter case of segmental overlap, the assigned ranking value is defined as $r := (r_i^k + r_j^l) \cdot f_{\text{boost}}$, where $1 \leq f_{\text{boost}} \leq M$ is a constant global boosting factor. The merging of the m segment lists is done iteratively until no residual segment list remains. Note that the factor f_{boost} is applied only once during the processing of the segment lists. When all m segment lists are merged into a single, integrated segment list, all of the segments' ranking values are normalized by applying the factor $1/(M \cdot f_{\text{boost}})$ resulting in a final ranking value in the interval $[0, 1]$. This algorithm can be implemented in a straight-forward manner with a time complexity linear in list lengths, as long as each list L^k is sorted ascending w.r.t. the start timestamps b_i^k of its matching segments $s_i^k := (b_i^k, e_i^k, r_i^k)$.

In the end, for every piece of music there results an individual, integrated list of multimodal matching segments along with assigned ranking values. The overall ranking value for a piece of music is determined by the maximum ranking value of its integrated segment list. Finally, the pieces of music are put into a new result list and sorted in descending order of their respective ranking values. This means that the final result list is organized such that the more modalities within pieces of music do match, the higher their assigned ranking values are. Therefore they occur at earlier positions in the list. In turn, pieces of music matching in less modalities occur at later positions in the list.

5.4 Integrated presentation of query results

Typically, available search engines provide the user with a flatly organized result list only where the list entries commonly consist of single documents. However, in the case of the music domain, there are multiple document types (in our case audio recordings, sheet music and lyrics documents) representing a piece of music using different modalities. As in our applications we have multiple documents of the different types available for a piece of music, we believe that it is of special interest to present all those documents in a collective manner, even if some of them do not match a user's query. Therefore, we took this consideration into account concerning the presentation of query results.

The bottom area of Fig. 9 shows the result list (left) and the document viewer (right). While the result list shows the matching pieces of music regarding the query, the document viewer offers access to the entire indexed content belonging to the currently selected piece of music. It furthermore gives a more detailed view on matching regions within its multimodal content and is also responsible for playing back the latter. As mentioned before, the resulting matches are presented to the user not at document level. Instead, the user is offered every piece of music where at least one document representing that piece contains one or more matches to the current query. All documents belonging to the same piece of music that match the user's query are summarized within a single list entry. The entry shows the artist's name and the title of the piece of music, a lyrics excerpt as well as the matching documents along with their number (in brackets). Additionally, at the bottom there are links to show more titles of the artist and to save the result (see also Sect. 5.5). A more detailed view of the single matching documents as well as the exact matching positions therein, is given in the document viewer.

Another key feature of the document viewer is the integrated display of matching segments along the timeline bar at the bottom. Besides the adjustment of the current playback position by using the slider knob, it is used to show all matching positions within the currently selected multimodal contents used for playback. The matching positions are indicated by colored boxes along the timeline bar, where the color and brightness of the boxes encode modality and ranking value, respectively. Additionally, matching segments within "inactive" documents, i. e., others than those ones used for playback, are displayed as gray boxes.

5.5 Query refinement and cross-modal navigation

From within the result list, for each retrieved piece of music, the user is enabled to request more titles of the same artist by choosing the appropriate "get more titles from artist…"-link which is available from the context menu. Once the user selects this option, the Query Bag is flushed, rebuilt with only a simple metadata query consisting of the artist's name and a subsequent new search is performed, what finally results in an updated list that displays all pieces of music by this artist contained in the database.

Moreover, the user can utilize content-based searching capabilities from within visual content following the query-by-example paradigm. When the user selects a portion of either a sheet music page or the lyrics text, he can use this excerpt for a new query, see Figs. 6 and 10. He has the option to start either a completely new search based only on the selected portion, or to add the query as an additional partial

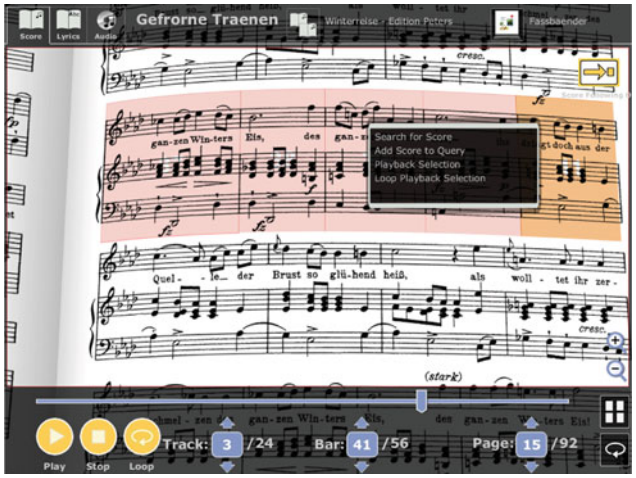

Fig. 10 Document viewer in score visualization mode. Multimodal content of selected measures can be queried

query to the Query Bag. In the case of sheet music, a portion may consist of two modalities, score and text (cf. Fig. 6). Here, the user can choose whether he queries both modalities together or separate from each other.

As matching segments within multimodal contents are displayed as boxes along the timeline bar at the bottom of the document viewer, they can be simultaneously utilized for navigation purposes. By clicking on a box, the playback is started or resumed at the corresponding time position. This functionality enables the user to jump directly to the found segments matching the user's query.

6 The PROBADO Music Digital Library

The PDLI is a research effort to develop next-generation digital library support for non-textual documents with the aim to contribute to all parts of the digital library workflow and processing chain from content acquisition to semi-automatic indexing, search, and presentation of content. As part of the PDLI, the PMDL is set up in the real-life scenario of the BSB Munich with the long-term goal of its incorporation into the business transition. To achieve this goal, several practice-relevant requirements have to be identified in order to work out practicable software solutions.

The digitalization efforts of the BSB so far produced a total of about 95,000 scanned pages of sheet music and 4,000 digital audio recordings. The data set here consists mainly of classic-romantic pieces of music, including piano sonatas, string quartets, and orchestral works. In joint work with the BSB, we have developed an entity-relationship data model [17] based on the FRBR model [32]. In contrast to the widely used Online Public Access Catalogue (OPAC) or Machine-Readable Cataloging (MARC) models, our model offers a more complex, favorable description of music documents. A key benefit is that it takes into account the complex connection between various expressions of a musical work and parts of it. Our data model is heavily utilized especially in the context of the multimodal music access, ranging from cross-modal indexing to the cross- and multimodal retrieval.

During the last year, the proposed music repository has gradually been integrated into the BSB workflow. While the indexing process of piano sonatas works quite well so far, orchestral works are the biggest challenge especially due to the occurrence of transposing instruments. As it will be hardly possible to avoid all possible types of errors, we are currently developing an interactive dialog system for the maintenance of new documents in case of the failure of the indexing algorithms.

7 Conclusions and future work

In this paper, we presented a framework for a digital music repository that has been developed for use in a real-world library scenario. As its main components the framework comprises

- a modular repository architecture,
- a data model for managing bibliographic data, metadata and the available heterogeneous document types,
- a workflow for automated document processing, particularly addressing the steps of content analysis, annotation, and indexing, as well as
- a user-interface that allows for the content-based navigation, browsing, and searching the underlying document collection.

For content-based document analysis, browsing, and navigation we put into practice several state-of-the-art methods from the field of MIR and hence bridge the gap between basic research, on one hand, and real-world applications, on the other hand.

As a major scientific contribution and underlying principle of the proposed system, cross-modal document processing is an integral part of all the stages of the document processing chain from entering new music documents into the system up to search, retrieval, and delivery of linked musical content. As a second major contribution, to facilitate cross- and multimodal retrieval, we propose an enhanced retrieval strategy offering composite queries.

We illustrated how the proposed framework is currently set up at the BSB Munich as part of the PDLI. Note that the developed software system and workflows are not restricted to work only within the BSB; rather, they are realized for generic application in real-life libraries. Both the developed software system and workflows have been designed

as *generic* components and may hence be equally used in a wide range of real-world music library scenarios. In fact, the underlying PDLI framework comprises a generic digital library architecture for generalized document types and is currently installed for additional document collections such as 3D models from the field of architecture data [8].

Required future work is manifold and ranges from several open basic research tasks to be addressed, over the adaptation of existing research results to become feasible solutions for everyday use in a library, to the improvement of the proposed workflow and its adaptation to further relevant processing modes and document types, as well as the detailed evaluation of preprocessing times, search times, and search quality. As some important examples, we mention the (fully) automatic synchronization of textual lyrics to their actual occurrences in the CD-audio, the aforementioned robust alignment of scanned sheet music to CD-audio material considering structural differences, variabilities or errors in the different documents, or the improvement of OMR/OCR results by using additionally available side information. Furthermore, automatic detection and processing of only partially available documents or inter-document inconsistencies is a challenging task for future work. Additionally, the review of the synchronization results is an important part of quality improvement and assurance. This specific task can neither be achieved automatically nor is a systematic execution by the library staff reasonable for very large data sets. Thus, a system for user feedback should be established to report incorrect synchronizations. Subsequently, the erroneous synchronizations can be revised. While the PMDL presented in this paper, for the first time, employs various singular MIR mechanisms such as music synchronization, matching and indexing, the systematic use of such techniques to achieve a fully automatic content indexing for music collections comprising solving the latter tasks, remains a largely open challenge.

Acknowledgments This work was supported in part by the Deutsche Forschungsgemeinschaft under Grant INST 11925/1-1.

References

1. Allamanche, E., Herre, J., Fröba, B., Cremer, M.: AudioID: Towards content-based identification of audio material. In: Proceedings of the 110th Audio Engineering Society (AES) Convention (2001)
2. Arifi, V., Clausen, M., Kurth, F., Müller, M.: Synchronization of music data in score-, MIDI- and PCM-format. Comput. Musicol. **13**, 9–33 (2004)
3. Baggi, D., Barate, A., Haus, G., Ludovico, L.A.: NINA—navigating and interacting with notation and audio. In: Proceedings of the 2nd International Workshop on Semantic Media Adaptation and Personalization (SMAP), pp. 134–139. IEEE Computer Society, Washington, DC, USA (2007). doi:10.1109/SMAP.2007.28
4. Bainbridge, D., Thompson, J., Witten, I.H.: Assembling and enriching digital library collections. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 323–334. IEEE Computer Society, Washington, DC, USA (2003)
5. Bartsch, M.A., Wakefield, G.H.: Audio thumbnailing of popular music using chroma-based representations. IEEE Trans. Multimed. **7**(1), 96–104 (2005)
6. Birmingham, W.P., Pardo, B., Meek, C., Shifrin, J.: The MusArt music-retrieval system: an overview. D-Lib Magazine **8**(2) (2002). doi:10.1045/february2002birmingham. URL http://www.dlib.org/dlib/february02/birmingham/02birmingham.html
7. Birmingham, W.P., O'Malley, K., Dunn, J.W., Scherle, R.: V2V: a second variation on query-by-humming. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), pp. 380–380. IEEE Computer Society, Washington, DC, USA (2003)
8. Blümel, I., Krottmaier, H., Wessel, R.: The PROBADO framework: a repository for architectural 3D-models. In: International Conference on Online Repositories in Architecture. Fraunhofer irb Verlag (2008)
9. Byrd, D., Schindele, M.: Prospects for improving OMR with multiple recognizers. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), pp. 41–46 (2006)
10. Cano, P., Battle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. In: Proceedings of the 5th IEEE Workshop on Multimedia Signal Processing (MMSP) (2002)
11. Choudhury, G., DiLauro, T., Droettboom, M., Fujinaga, I., Harrington, B., MacMillan, K.: Optical music recognition system within a large-scale digitization project. In: Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR) (2000)
12. Clausen, M., Kurth, F.: A unified approach to content-based and fault-tolerant music recognition. IEEE Trans. Multimed. **6**(5), 717–731 (2004)
13. D'Aguanno, A., Vercellesi, G.: Automatic music synchronization using partial score representation based on IEEE 1599. J. Multimed. **4**(1), 19–24 (2009)
14. Damm, D., Kurth, F., Fremerey, C., Clausen, M.: A concept for using combined multimodal queries in digital music libraries. In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL) (2009)
15. Damnjanovic, I., Reiss, J., Barry, D.: Enabling access to sound archives through integration, enrichment , and retrieval. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo (ICME), pp. 1597–1598 (2008). doi:10.1109/ICME.2008.4607756
16. Dannenberg, R.B., Raphael, C.: Music score alignment and computer accompaniment. In: Pardo, B. (ed.): Special Issue: Music Information Retrieval, vol. 49, pp. 38–43. ACM, New York, NY, USA (2006). doi:10.1145/1145287.1145311
17. Diet, J., Kurth, F.: The PROBADO music repository at the Bavarian State Library. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), pp. 501–504 (2007)
18. Dixon, S., Widmer, G.: MATCH: A music alignment tool chest. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR) (2005)
19. Dunn, J.W., Byrd, D., Notess, M., Riley, J., Scherle, R.: Variations2: Retrieving and using music in an academic setting. In: Pardo, B. (ed.): Special Issue: Music Information Retrieval, vol. 49, pp. 53–58. ACM, New York, NY, USA (2006). doi:10.1145/1145287.1145314
20. European Union: EUROPEANA (2007). http://www.europeana.eu/portal/index.html
21. Fremerey, C., Müller, M., Kurth, F., Clausen, M.: Automatic mapping of scanned sheet music to audio recordings. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), pp. 413–418. Philadelphia, USA (2008)
22. Fremerey, C., Clausen, M., Ewert, S., Müller, M.: Sheet music-audio identification. In: Proceedings of the 10th International

Conference on Music Information Retrieval (ISMIR), pp. 645–650. Kobe, Japan (2009a)
23. Fremerey, C., Müller, M., Clausen, M.: Towards bridging the gap between sheet music and audio. In: Selfridge-Field, E., Wiering, F., Wiggins, G.A. (eds.) Knowledge Representation for Intelligent Music Processing, no. 09051 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Germany, Dagstuhl, Germany (2009b). http://drops.dagstuhl.de/opus/volltexte/2009/1965
24. Fremerey, C., Müller, M., Clausen, M.: Handling repeats and jumps in score-performance synchronization. In: Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR). Utrecht, the Netherlands (2010)
25. Good, M.: MusicXML: An internet-friendly format for sheet music. In: Proceedings XML Conference and Exposition (2001). http://www.idealliance.org/papers/xml2001/papers/html/03-04-05.html
26. Google Inc.: Google Book Search (2007). http://books.google.com
27. Goto, M.: A chorus-section detecting method for musical audio signals. In: Proceedings of the IEEE Internatinal Conference on Acoustics, Speech, and Signal Processing ICASSP, pp. 437–440 (2003)
28. Gracenote: Music Search (2008). http://www.gracenote.com/
29. Hankinson, A., Pugin, L., Fujinaga, I.: Interfaces for document representation in digital music libraries. In: Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR), pp. 39–44 (2009)
30. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proceedings of the 4th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2003)
31. Huber, D.M.: The MIDI Manual. Focal Press, Boston (1999)
32. IFLA Study Group: Functional requirements for bibliographic records: Final report. UBCIM Publications-New Series **19** (1998). http://www.ifla.org/VII/s13/frbr/frbr.htm
33. Kahle, B.: Internet Archive (1996). http://www.archive.org/index.php
34. Klapuri, A., Davy, M. (eds.): Signal Processing Methods for Music Transcription. Springer, New York (2006)
35. Krajewski, E.: DE-PARCON softwaretechnologie (2008). http://www.de-parcon.de/
36. Krottmaier, H., Kurth, F., Steenweg, T., Appelrath, H.J., Fellner, D.: PROBADO—a generic repository integration framework. In: Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL) (2007)
37. Kurth, F., Müller, M., Fremerey, C.: Audio Matching für symbolische Musikdaten. In: Fortschritte der Akustik, Tagungsband der DAGA (2007a).http://www.cs.uni-bonn.de/~meinard/publications/07_KuMuFr_DAGA_SymbAudioMatch.pdf
38. 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), pp. 261–266 (2007b)
39. Kurth, F., Müller, M.: Efficient index-based audio matching. IEEE Trans. Audio Speech Lang. Process. **16**(2), 382–395 (2008)
40. Landone, C., J., H., Reiss, J.: Enabling access to sound archives through integration, enrichment and retrieval: the EASAIER project. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), pp. 159–160 (2007)
41. Ludovico, L.A.: IEEE 1599: a multi-layer approach to music description. J. Multimed. **4**(1), 9–14 (2009)
42. Maddage, N.C., Xu, C., Kankanhalli, M.S., Shao, X.: Content-based music structure analysis with applications to music semantics understanding. In: Proceedings of the ACM Multimedia, pp. 112–119. New York, NY, USA (2004). doi:10.1145/1027527.1027549
43. Müller, M.: Information Retrieval for Music and Motion. Springer, New York (2007)
44. Müller, M., Appelt, D.: Path-constrained partial music synchronization. In: Proceedings of the 34th International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 65–68. Las Vegas, Nevada, USA (2008)
45. Müller, M., Clausen, M.: Transposition-invariant self-similarity matrices. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), pp. 47–50 (2007)
46. Müller, M., Kurth, F.: Towards structural analysis of audio recordings in the presence of musical variations. EURASIP J. Appl. Signal Process. **2007**(89686), 18 (2007)
47. Müller, M., Kurth, F., Röder, T.: Towards an efficient algorithm for automatic score-to-audio synchronization. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR), pp. 365–372. Barcelona, Spain (2004)
48. Müller, M., Kurth, F., Damm, D., Fremerey, C., Clausen, M.: Lyrics-based audio retrieval and multimodal navigation in music collections. In: Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL) (2007)
49. Orio, N.: Alignment of performances with scores aimed at content-based music access and retrieval. In: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), pp. 479–492. Rome, Italy (2002)
50. Orio, N., Lemouton, S., Schwarz, D.: Score following: State of the art and new developments. In: Proceedings of the Conference of New Interfaces for Musical Expression (NIME), pp. 36–41. Montreal, CA (2003)
51. Pardo, B.: Introduction. In: Pardo, B. (ed.): Special Issue: Music Information Retrieval, vol. 49, pp. 28–31. ACM, New York, NY, USA (2006). doi:10.1145/1145287.1145309
52. Peeters, G., Burthe, A.L., Rodet, X.: Toward automatic music audio summary generation from signal analysis. In: Proceedings of the 3th International Conference on Music Information Retrieval (ISMIR) (2002)
53. Pickens, J., Bello, J.P., Monti, G., Crawford, T., Dovey, M., Sandler, M.: Polyphonic score retrieval using polyphonic audio queries: a harmonic modeling approach. In: Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR), pp. 140–149. Paris, France (2002)
54. Pinto, A.: Multi-model music content description and retrieval using IEEE 1599 XML standard. J. Multimed. **4**(1), 30–39 (2009)
55. Raphael, C.: A hybrid graphical model for aligning polyphonic audio with musical scores. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR) (2004)
56. Rauber, A., Frühwirth, M.: Automatically analyzing and organizing music archives. In: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Springer Lecture Notes in Computer Science. Springer, Darmstadt, Germany (2001). http://www.ifs.tuwien.ac.at/ifs/research/publications.html
57. Selfridge-Field, E. (ed.): Beyond MIDI: The Handbook of Musical Codes. MIT Press, Cambridge (1997)
58. Soulez, F., Rodet, X., Schwarz, D.: Improving polyphonic and poly-instrumental music to score alignment. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR) (2003)
59. Suyoto, I.S.H., Uitdenbogerd, A.L., Scholer, F.: Searching musical audio using symbolic queries. IEEE Trans. Audio Speech Lang. Process. **16**(2), 372–381 (2008). doi:10.1109/TASL.2007.911644
60. Turetsky, R.J., Ellis, D.P.: Force-aligning MIDI syntheses for polyphonic music transcription generation. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR) (2003a)

61. Turetsky, R.J., Ellis, D.P.W.: Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR) (2003b)
62. Typke, R., Wiering, F., Veltkamp, R.C.: A survey of music information retrieval systems. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), pp. 153–160 (2005)
63. Union der deutschen Akademien der Wissenschaften: Neue Mozart Ausgabe (2007). http://www.nma.at/
64. United States: World Digital Library (2009). http://www.wdl.org/en/
65. University of Chicago Library: Chopin Early Edition (2004). http://chopin.lib.uchicago.edu/
66. University of Rochester Libraries: UR research—Sibley Music Library (2009). https://urresearch.rochester.edu/home.action
67. W3C: Web Services. http://www.w3.org/2002/ws/
68. Wang, A.L.C.: An industrial-strength audio search algorithm (2003). http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf
69. Wang, Y., Kan, M.Y., Nwe, T.L., Shenoy, A., Yin, J.: LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics. In: Proceedings of the 12th annual ACM International Conference on Multimedia, pp. 212–219. ACM Press, New York, NY, USA (2004). http://doi.acm.org/10.1145/1027527.1027576
70. Wiener Wissenschafts-, Forschungs- und Technologiefonds: Schubert-Autographe. http://www.schubert-online.at/
71. Witten, I.H., Moffat, A., Bell, T.C.: Managing Gigabytes. 2nd edn. Van Nostrand Reinhold, New York (1999)
72. Witten, I.H., Mcnab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: Proceedings of the 5th ACM International Conference on Digital Libraries (2000). http://citeseer.ist.psu.edu/witten99greenstone.html