frontiers in Digital Humanities

ORIGINAL RESEARCH
published: 20 February 2018
doi: 10.3389/fdigh.2018.00001

Bridging the Gap: Enriching YouTube Videos with Jazz Music Annotations

Stefan Balke[1], Christian Dittmar[1], Jakob Abeßer[2], Klaus Frieler[3], Martin Pfleiderer[3] and Meinard Müller[1]*

[1] International Audio Laboratories Erlangen, Erlangen, Germany, [2] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany, [3] Jazzomat Research Project, University of Music Franz Liszt, Weimar, Germany

Web services allow permanent access to music from all over the world. Especially in the case of web services with user-supplied content, e.g., YouTube™, the available metadata is often incomplete or erroneous. On the other hand, a vast amount of high-quality and musically relevant metadata has been annotated in research areas such as Music Information Retrieval (MIR). Although they have great potential, these musical annotations are often inaccessible to users outside the academic world. With our contribution, we want to bridge this gap by enriching publicly available multimedia content with musical annotations available in research corpora, while maintaining easy access to the underlying data. Our web-based tools offer researchers and music lovers novel possibilities to interact with and navigate through the content. In this paper, we consider a research corpus called the Weimar Jazz Database (WJD) as an illustrating example scenario. The WJD contains various annotations related to famous jazz solos. First, we establish a link between the WJD annotations and corresponding YouTube videos employing existing retrieval techniques. With these techniques, we were able to identify 988 corresponding YouTube videos for 329 solos out of 456 solos contained in the WJD. We then embed the retrieved videos in a recently developed web-based platform and enrich the videos with solo transcriptions that are part of the WJD. Furthermore, we integrate publicly available data resources from the Semantic Web in order to extend the presented information, for example, with a detailed discography or artists-related information. Our contribution illustrates the potential of modern web-based technologies for the digital humanities, and novel ways for improving access and interaction with digitized multimedia content.

Keywords: music information retrieval, digital humanities, audio processing, semantic web, multimedia

OPEN ACCESS

Edited by:
Mark Brian Sandler, Queen Mary University of London, United Kingdom

Reviewed by:
Anna Wolf, Hanover University of Music Drama and Media, Germany
Michael Scott Cuthbert, Massachusetts Institute of Technology, United States

****Correspondence:***
Stefan Balke
stefan.balke@audiolabserlangen.de

Specialty section:
This article was submitted to Digital Musicology, a section of the journal Frontiers in Digital Humanities

Received: *09 October 2017*
Accepted: *16 January 2018*
Published: *20 February 2018*

Citation:
Balke S, Dittmar C, Abeßer J, Frieler K, Pfleiderer M and Müller M (2018) Bridging the Gap: Enriching YouTube Videos with Jazz Music Annotations. Front. Digit. Humanit. 5:1. doi: 10.3389/fdigh.2018.00001

1. INTRODUCTION

Online video platforms, such as YouTube, make billions of videos available to users from all over the world. Many of these videos contain recordings of music performances. Often, these performances are tagged with basic metadata—mainly the artist and the title of the song. However, since this metadata is not curated, it might be incomplete or incorrect. The lack of reliable metadata makes it hard to identify particular recordings, especially for music genres where many renditions of the same musical work exist (e.g., symphonies in Western classical music, ragas in Indian music, or standards in jazz music). Imagine a jazz student who is practicing a jazz solo played by a famous musician and is now interested in the original recording. In the case that the student searches for a musician whose name is not mentioned in the metadata (e.g., because the musician was "only"

a sideman in the band), a textual search may not be successful or may result in too many irrelevant results. Assuming that the student has already a partial or even a complete transcription of the solo available, content-based retrieval techniques could help to resolve this problem. Here, *content-based* means that, in the comparison of music data, the system makes use of the raw music data itself (e.g., from the music recording or the YouTube video), rather than relying on manually generated keywords referring to the artists' names, the song's title or lyrics (Müller, 2007).

Jazz musicians, musicologists, and publishers have made many jazz solo transcriptions publicly available during the last decades, e.g., Hal Leonard's *Omnibook* series.[1] One comprehensive corpus of solo transcriptions is the Weimar Jazz Database (WJD), which consists of 456 (as of May 2017) transcriptions of instrumental solos in jazz recordings performed by a wide range of renowned musicians (Pfleiderer et al., 2017). The solos have been manually transcribed by musicology and jazz students. In addition, the database offers various music-related annotations such as chord sequences or beat positions. We believe that these annotations are a great resource that could help musicians and other researchers in gaining a deeper understanding of jazz music. However, these annotations and the underlying audio material are not directly accessible, mainly for two reasons. First, the audio files originate from commercial music recordings which are protected by copyright and ancillary copyright laws. Therefore, they cannot be made publicly available by scientific institutions. This restricts the usefulness of the dataset for scientific research, where both the annotations and the corresponding audio material are required. Second, the annotations are encoded in a database format which is not easily accessible for users without technical skills. Both problems apply to many scientific datasets which offer musical annotations for commercial music recordings. Simply switching to music recordings that are released under public domain licenses is not an option for research questions which rely on specific music recordings. In our approach, we try to bypass some of these copyright restrictions by using music recordings that are publicly available via YouTube. However, there is no doubt that both musicians and composers should be gratified financially for the music they create according to national and international copyright and ancillary copyright laws. YouTube seems to guarantee this financial entitlement through agreements with national copyright collecting societies. By contrast, for scientific institutions offering music databases, it is very difficult or impossible to handle these legal claims. As a case study, we focus on the recordings that have corresponding annotations in the WJD.

As the main contribution of this paper, we introduce various retrieval methods based on metadata and content-based descriptors and show how these techniques can be applied for identifying and enriching YouTube videos. In the following, we sketch a typical two-stage retrieval scenario which is then described in more detail in the subsequent sections (**Figure 1** provides an overview). In this example, we are interested in the song *Jordu*, recorded by Clifford Brown in 1954 (**Figure 1A**). In the first step, we use the title and the name of the soloist as provided by the WJD to perform a metadata-based search on YouTube (**Figure 1B**). This search results in a list of candidates. Besides relevant music recordings, this list may also contain other recordings by the same artist or

[1]https://www.halleonard.com/search/search.action?seriesfeature=OMNIBK.

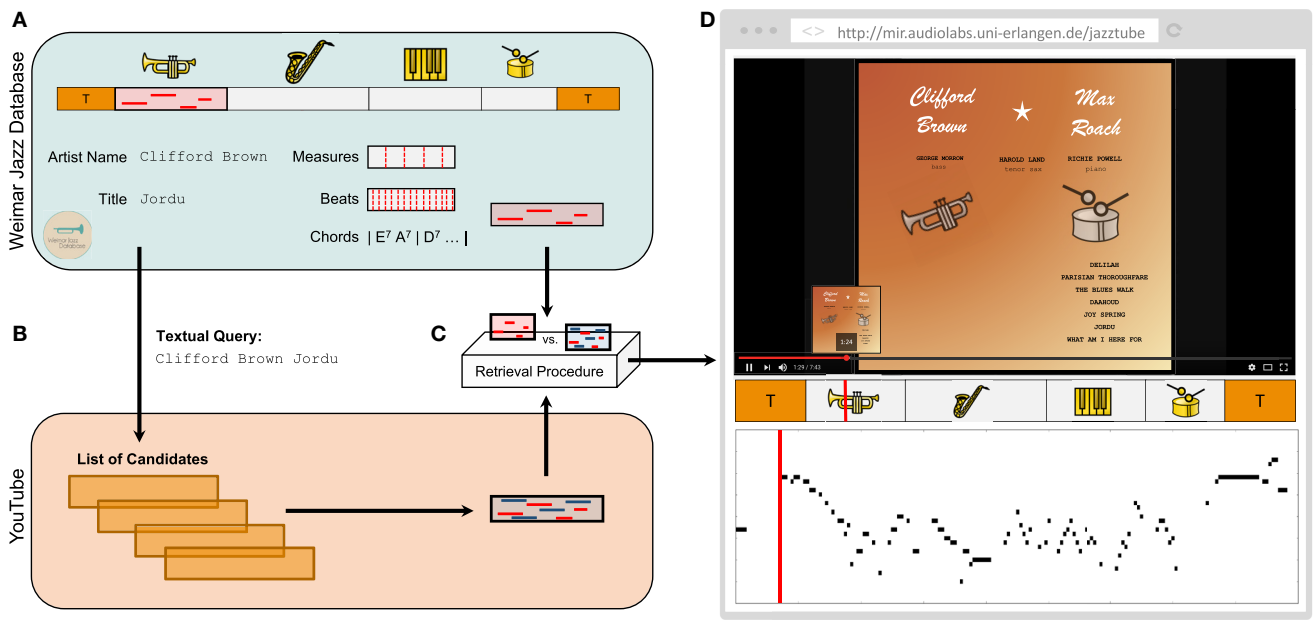

FIGURE 1 | Illustration of the two-stage retrieval scenario applied to retrieve videos from YouTube. **(A)** Overview of the various annotations for Clifford Brown's solo *Jordu* contained in the Weimar Jazz Database. **(B)** First retrieval stage: text-based retrieval on YouTube resulting in a list of candidates. **(C)** Second retrieval stage: content-based retrieval using the solo recording from the WJD as query. **(D)** Identified video embedded in a web-based demonstrator and enriched with the annotations obtained from the WJD. Panel **(D)** has been created by the authors and, therefore, no permission is required for its use in this manuscript.

cover versions by other artists. Using the recording associated to the WJD's annotations, we apply an audio-based retrieval approach to identify the relevant music recordings in this list of candidates (**Figure 1C**). The result of this matching procedure is a list of relevant documents that can be used to link the WJD's annotations to the YouTube videos. The retrieved video is then embedded in a web-based application (**Figure 1D**). In addition, we use the annotations provided by the WJD to further enrich the video, e.g., by offering new navigation possibilities based on the song structure or transcriptions of the song's solo. As a result, the user is able to follow the soloist's improvisation in a piano-roll-like representation. For intuition and hands-on experience with this concept, our web-based application can be accessed under the following address: http://mir.audiolabs.uni-erlangen.de/jazztube.

The remainder of this paper is structured as follows. We start by giving a brief overview of the literature and related projects (Section 2). Then, we introduce the different data resources used in this study (Section 3). Subsequently, we describe the various retrieval procedures that are used to link the WJD to the YouTube videos (Section 4). Finally, we present a web-based service that integrates the introduced data resources in a unifying user interface (Section 5).

2. RELATED WORK

Similar web-based services that aim to enhance the listening experience have been proposed in the past. *Songle*,[2] for instance, lets users explore music from different perspectives (Goto et al., 2011). In this web-based service, computational approaches are used to annotate music recordings (including beats, melodic lines, or chords). Afterward, these generated annotations are presented in a web-based interface. Since the automatically generated annotations may contain errors, the users can correct them or add new ones. The annotations contained in *Songle* can then be used in third-party applications or research projects (e.g., for singing-voice analysis). Another service called *Songrium*,[3] allows users to add lyrics to publicly available videos (e.g., obtained from YouTube). In addition, the lyrics can be visualized and played back along with the linked video similar to karaoke applications. For an overview of other systems by Goto and colleagues, we refer to the literature, see Goto (2011, 2014).

Another project that aims at enhancing the listening experience, especially for classical music, is called *PHENICX* (Performances as Highly Enriched aNd Interactive Concert eXperiences) (Gasser et al., 2015; Liem et al., 2015, 2017; Melenhorst et al., 2015). As one main functionality, suitable visualizations are generated in real-time and displayed during the live performance of an orchestra. Such visualizations may be a rendition of a musical score (score-following applications) or an animation controlled by the baton movements of the orchestra's conductor. Furthermore, as in our scenario, the project offers a web-based service, which allows the playback of enriched videos.[4] In the research project *Freischütz Digital*,[5] user interfaces for dealing with critical editions in an opera scenario were developed (Prätzlich et al., 2015; Röwenstrunk et al., 2015). In this scenario, an essential step is to link the different sheet music editions with the various existing music recordings. These alignments are then used in special user interfaces that may support musicologists in their work on critical editions.

Besides publicly available music recordings or videos, the internet offers additional information (metadata or textual annotations) for music recordings. Many services offer metadata in a structured way, often following standardized data formats as defined in the *Semantic Web* (Berners-Lee et al., 2001). The Semantic Web contains standardized schemas, called *ontologies*, for exchanging different kinds of data. A way to exchange musical annotations is defined in the *Music Ontology* (Raimond et al., 2007). One of the most frequently used services in the Semantic Web is *DBpedia*[6] which offers information from Wikipedia in a structured data format. Popular services for music metadata in general are *MusicBrainz*[7] or *Discogs*.[8] In particular for jazz music, the *JDISC*[9] project aims to provide complete discographies for a number of selected artists. Another related project is called *Linked Jazz*,[10] which offers relationships between jazz musicians in a structured way (Pattuelli, 2012). Besides sharing metadata, researchers have used YouTube as a way of specifying datasets that were used in their experiments (Schoeffler and Herre, 2014). In particular, for audio applications, Google released *AudioSet*, a dataset consisting of over two million 10-s sound clips obtained from YouTube which have then been labeled by human annotators (Gemmeke et al., 2017).

This work follows similar concepts as used in the *SyncPlayer* (Kurth et al., 2005; Thomas et al., 2009; Damm et al., 2012). The SyncPlayer offers various ways of interacting and navigating with a large, multimodal corpus of music recordings, sheet music, and lyrics. Furthermore, users are able to search within this corpus by specifying a short melodic phrase or an excerpt from the lyrics. The results are then presented in an interactive graphical user interface that allows auditioning the results. In previous works, we studied the use of interfaces for two different music scenarios. In Balke et al. (2017a), a web-based user interface motivated by applications in jazz-piano education is presented. In particular, a video recording, a piano-roll representation, and additional annotations are incorporated in a unifying interface that allows the user to simultaneously play back the different media objects. A related approach focusses on the opera *Die Walküre (The Valkyrie)* from Richard Wagner's cycle *Der Ring des Nibelungen (The Ring of the Nibelung)*. The goal of the interface is to supply intuitive functions that allow a user to easily access and explore all available data (including different recordings, videos, lyrics, sheet music) associated with a large-scale work such as an opera.

[2] http://songle.jp.
[3] http://songrium.jp.
[4] http://phenicx.prototype.videodock.com.
[5] http://www.freischuetz-digital.de.
[6] http://www.dbpedia.org.
[7] https://www.musicbrainz.org.
[8] https://www.discogs.com.
[9] http://jdisc.columbia.edu.
[10] https://www.linkedjazz.org.

3. DATA RESOURCES

In this paper, we consider jazz-related data of different modality stemming from different resources. We now introduce the Weimar Jazz Database (WJD), the relevant jazz recordings, the streaming platform YouTube from which we obtain videos, and the used web resources for additional metadata.

3.1. Weimar Jazz Database (WJD)

The WJD is part of the Jazzomat Research Project,[11] which aims at a better understanding of creative processes in improvisations using computational methods (Pfleiderer et al., 2017). The WJD comprises 456 (as of July 2017) high-quality solo transcriptions (similar to a piano-roll representation), extracted from 343 tracks taken from 197 different records. The solos are performed by a wide range of renowned jazz musicians in the period from 1925 to 2009 (e.g., Louis Armstrong, Don Byas, or Chris Potter). All solos were manually annotated by musicology and jazz students at the University of Music Franz Liszt Weimar using the *SonicVisualiser* (Cannam et al., 2006). The annotators had different musical backgrounds but a general familiarity with jazz music, mostly through listening and playing. The produced transcriptions were then inspected with an automated verification procedure that primarily searched for syntactical errors and suspicious annotations, such as beat outliers. In a final step, the transcription was cross-checked by an experienced supervisor and added to the database. **Table 1** lists the number of solo transcriptions grouped by the 13 different occurring solo instruments. As one might expect for jazz music, the database is biased toward tenor saxophone and trumpet solos, which represent about 56% of the currently available solo transcriptions.

Figure 2A shows the distribution of the solos with respect to their durations and recording years. The solos have a minimum duration of 19 s (Steve Coleman's second solo on *Cross-Fade*), a maximum duration of 818 s (John Coltrane's solo on *Impressions*), and an average duration of 107 s. Similarly, **Figure 2B** indicates the distribution of the whole tracks (which usually contain more than a single solo part), with a minimum duration of 128 s, a maximum duration of 1620 s, and an average duration of 354 s. From all 343 tracks, there are 247 tracks with one annotated solo part, 80 tracks with two, 15 with three, and a single track with four annotated solo parts. Summing over the number of annotated note events in all solo transcriptions results in over 200,000 elements.[12]

Figure 3 displays the beginning of *Clifford Brown's* solo on *Jordu* as an example for the data contained in the WJD. **Figure 3A** shows a time-frequency representation (see Section 3.2 for details)

[11] http://jazzomat.hfm-weimar.de.

TABLE 1 | Solo instruments occurring in the WJD.

Abbr.	Instrument	#Solos
cl	Clarinet	15
bcl	Bass clarinet	2
ss	Soprano saxophone	23
as	Alto saxophone	80
ts	Tenor saxophone	157
ts-c	Tenor saxophone in C	1
bs	Baritone saxophone	11
tp	Trumpet	102
cor	Cornet	15
tb	Trombone	26
g	Guitar	6
p	Piano	6
vib	Vibraphone	12
13		∑ 456

The first column introduces an abbreviation, whereas the last column indicates the number of solos of the respective instrument.

[12] Additional statistics: http://mir.audiolabs.uni-erlangen.de/jazztube/statistics/.

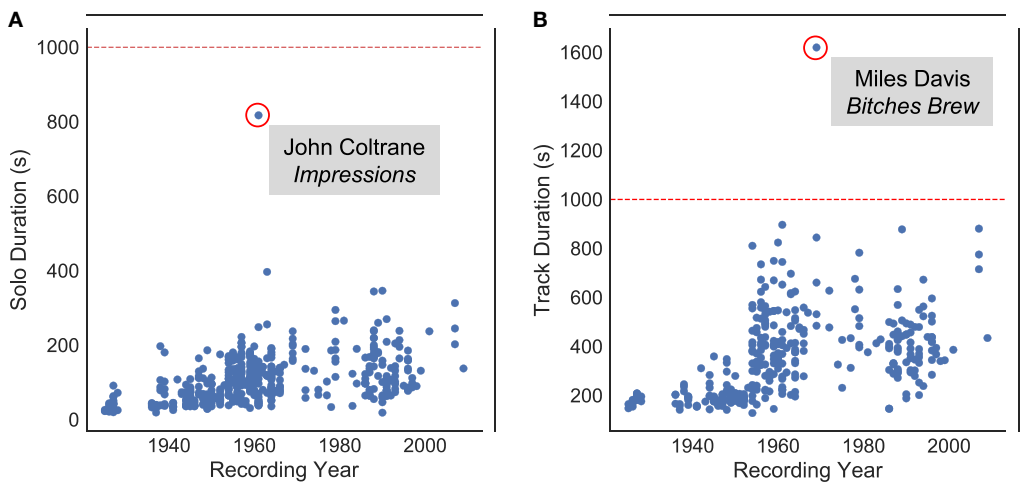

FIGURE 2 | **(A)** 456 solos and **(B)** 343 tracks considered in the WJD, represented according to their duration and respective recording years. The dashed lines indicate the limits for the maximum duration of the considered YouTube videos.

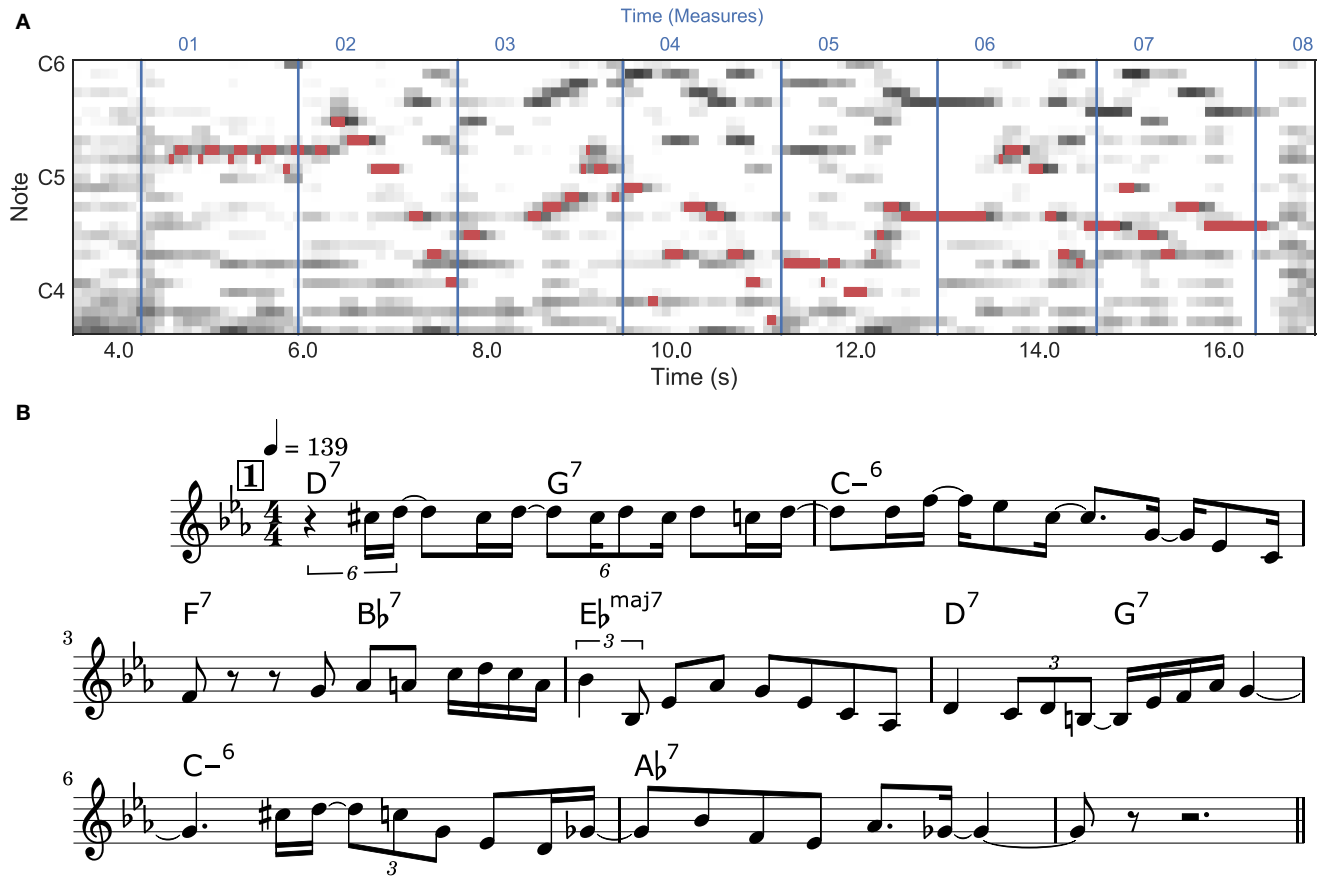

FIGURE 3 | Beginning of *Clifford Brown's* solo on *Jordu*. **(A)** Log-frequency spectrogram, where rectangles (red) indicate the solo transcription and vertical lines (blue) the annotated measure positions. **(B)** Sheet music representation of the transcribed solo.

of this excerpt superimposed by the available solo transcriptions (each note is represented by a red rectangle) and measure positions (represented as blue vertical lines). **Figure 3B** shows a sheet music representation derived from the solo annotations. Note that deriving sheet music from the transcriptions requires algorithms that are able to quantize the onsets and durations of the annotated note events into musically meaningful notes, see Frieler and Pfleiderer (2017).[13]

3.2. Jazz Recordings

A typical jazz recording consists of a soloist who is accompanied by a rhythm section (e.g., double bass, piano, and drums). From an engineering perspective, such a recording is a sequence of amplitude values sampled from a microphone signal (or a mixture of multiple signals). By applying digital signal processing methods, one can analyze and manipulate such signals. A common way to analyze music signals is to transform them into a time-frequency representation, e.g., a spectrogram. For example, in **Figure 2A**, we show an excerpt of a spectrogram from *Clifford Brown's* solo on *Jordu*. There exist different approaches to obtain such a time-frequency representation.[14] In particular, we use a logarithmically spaced frequency axis with a bandwidth of a single semitone per frequency band (row in the spectrogram)—motivated by human's logarithmic perception of frequency and the equal-tempered scale underlying the music. In this representation, one can locate note onsets and durations, as well as harmonic partials generated by the sounding instruments. For an overview of computational approaches and music processing in general, we refer to the literature, e.g., Müller (2015), Knees and Schedl (2016), and Weihs et al. (2016).

3.3. Videos

There exist many different web services that offer users to publish videos. Among these services, YouTube[15] is without doubt the largest and most famous platform for video sharing. For our scenario, we are particularly interested in YouTube videos that contain music—especially the music that underlies the WJD. Some of the offered music videos are official releases by record labels, but the majority are videos uploaded by private platform

[13]The sheet music representation was generated by using the *LilyPond* (http://www.lilypond.org/) export which can be obtained from the WJD by using the *MeloSpyGUI* (http://jazzomat.hfm-weimar.de/download/download.html#download-melospygui).

[14]For the example in Figure 2a, we use the semitone filterbank described in (Müller, 2007; Müller and Ewert, 2011).

[15]http://www.youtube.com.

users. Especially the music videos uploaded by the private users often contain only a static image (cover art) or a slideshow while the audio track is a digitized version of the commercially available record. By embedding YouTube videos in a web service, one relies on the availability of these videos. Due to user deletions, copyright infringements, or legal constraints in some countries, videos may not be available. However, YouTube has a lot of redundancy, i.e., the same music recording may be available in more than one version.

3.4. Additional Metadata

In addition to the solo transcriptions, the WJD contains basic metadata for the music recordings (e.g., artist and record name), as well as a special identifier for the MusicBrainz[16] platform. MusicBrainz is a community-driven platform, which collects music metadata and makes it publicly available. With the identifier available in the WJD, one is able to request a comprehensive list of available metadata from the MusicBrainz platform (e.g., participating musicians, producer's name, and so on). Furthermore, MusicBrainz can serve as a gateway to other web services that offer different kinds of metadata or even other multimedia objects (e.g., pictures of the artist). This "web of data" is often referred to as the Semantic Web (Berners-Lee et al., 2001). By using the web service DBpedia,[17] we can furthermore obtain and integrate content published on Wikipedia (e.g., bibliographic information about the artist).

4. RETRIEVAL AND LINKING STRATEGIES

In this section, we report on experiments where we systematically created links between the annotations contained in the WJD and corresponding YouTube videos. The retrieval task is as follows: given a specific music recording or a solo annotation provided by the WJD as query, identify the relevant videos in the pool of YouTube videos. Since the number of YouTube videos is very large, we follow a two-step retrieval strategy that we describe in the following.

4.1. Retrieval Scenario

We started by formalizing our retrieval task following (Müller, 2015). Let $\mathcal{D}$ be the set of all documents available on YouTube. A YouTube document $D \in \mathcal{D}$ consists of the video and the available metadata. Let $\mathcal{Q}$ be a collection of documents available within the WJD. A WJD document $Q \in \mathcal{Q}$ consists of a solo annotation, the underlying music excerpts, as well as metadata. In our scenario, the document Q served as query, whereas $\mathcal{D}$ was the database to search in. Given a query Q, the retrieval task was to identify the corresponding documents D. In our scenario, we followed a two-step retrieval strategy. First, we performed a metadata-based retrieval using the YouTube search engine. For a query Q, the result of the text-based retrieval is denoted as $\mathcal{D}_Q^{Text} \subset \mathcal{D}$. In the second step, we performed content-based retrieval only based on $\mathcal{D}_Q^{\text{Text}}$ to identify the relevant documents, denoted as $\mathcal{D}_Q^{\text{Rel}} \subseteq \mathcal{D}_Q^{\text{Text}}$.

4.2. Text-Based Retrieval

In the first step of our retrieval strategy, we extracted a subset of possible video candidates from YouTube. These were retrieved by performing two text-based queries (per solo) using the standard YouTube search engine (using YouTube's default settings). The first text-based query term consisted of the name of the soloist and the song title (e.g., `John Coltrane Kind of Blue`). Since the soloist is not always the artist who released the record, we performed a second text-based query that consisted of the artist's name under which the record was released, followed by the song title (in our example: `Miles Davis Kind of Blue`). From each retrieval result, we took the top 20 candidates (or less, depending on the number of YouTube search results). Furthermore, we only considered videos that are shorter or equal to 1000 s to avoid videos where users uploaded, for instance, complete records to YouTube (rather than individual songs).

In our experiments, using the first text-based query terms for all 456 solos considered in the WJD led to a pool of 4,114 video candidates. The second text-based query resulted in a pool of 4,069. In a next step, we fused the two candidate pools together, where we removed duplicates by using the video identifiers attached to every YouTube video. Our final candidate pool comprised 5,199 video candidates—resulting in approximately 12 candidates per query (solo). Note that $\mathcal{D}_Q^{\text{Text}}$ may still contain cover versions and other irrelevant documents. The following audio-based retrieval step is intended to resolve this issue.

4.3. Audio-Based Retrieval

In a second step, we used the audio recordings from the WJD to refine the list of candidates $\mathcal{D}_Q^{\text{Text}}$ obtained from the text-based retrieval. This task is also known as *audio identification* and can be approached in many different ways, see, e.g., Cano et al. (2005), Müller (2015), and Arzt (2016). Our method is based on chroma features and diagonal matching which is easily extendable to retrieval scenarios with different query and database modalities (e.g., matching solo transcription against audio recordings or matching audio excerpt against sheet music representations). Furthermore, since we performed our retrieval only on the small subsets $\mathcal{D}_Q^{\text{Text}}$, we do not consider efficiency issues here. In particular, we used a chroma variant called CENS with a feature rate of 5 Hz (Müller et al., 2005; Müller and Ewert, 2011). [18] We compared a query Q with each of the documents $D \in \mathcal{D}_Q^{\text{Text}}$ by using diagonal matching. This comparison yields a distance value $\delta_{Q,D} \in [0:1]$ for each pair (Q,D), where $\delta_{Q,D} = 0$ refers to a perfect match and $\delta_{Q,D} = 1.0$ to a poor match. By sorting the documents $D \in \mathcal{D}_Q^{\text{Text}}$ by $\delta_{Q,D}$ in an ascending order, one receives a ranked list. In this ranked list, the most similar documents (w.r.t. to the used distance function) are listed on top. In the case of extracting the relevant documents, one has to further process this ranked list. For instance, one may mark a document as relevant if $\delta_{Q,D}$ is smaller than a threshold $\tau \in [0:1]$. All relevant documents that fulfill this condition are then collected in the subset $\mathcal{D}_Q^{\text{Rel}} \subseteq \mathcal{D}_Q^{\text{Text}}$.

[16] http://www.musicbrainz.org/.

[17] http://wiki.dbpedia.org.

[18] All computations can be done by using the implementations provided by the Python library *librosa* (McFee et al., 2017).

Using this retrieval approach with a threshold $\tau = 0.1$, we were able to identify 988 relevant videos for 329 solos on YouTube (on average 3 relevant videos per solo, min = 1, max = 9). For 92 queries, we retrieved 1 relevant document, for 67 queries 2, for 60 queries 3, and for 110 queries more than 3 documents. However, for 124 queries, we were not able to find any relevant videos. We found different reasons for this from manually inspecting some candidate lists. One obvious reason is that the metadata-based retrieval step did not return any relevant documents in $\mathcal{D}_Q^{\text{Text}}$ (e.g., for the textual `David Murray Ask me Now`). Sometimes, only other versions of the same song are available on YouTube, for instance, the textual query `Art Pepper Anthropology` yields mainly results for the version of this song from the record *Art Pepper + Eleven: Modern Jazz Classics* instead of the relevant version from the record *The Intimate Art Pepper*. Furthermore, in many instances, we found that relevant documents were present in $\mathcal{D}_Q^{\text{Text}}$, but not recognized since the distance value $\delta_{Q,D}$ surpassed the chosen threshold $\tau = 0.1$ by a small margin.

4.4. Solo-Based Retrieval

The previous experiments were based on the assumption that we have access to the music recordings underlying the WJD annotations. However, in certain scenarios this might not be the case, for instance, when only a score representation of the piece or the solo is available. In this case, audio identification is no longer possible and one needs more general retrieval strategies. In the following experiment, we simulate a retrieval scenario by using the WJD's solo transcriptions (see **Figure 3A**) as query and convert them to chroma features. This constitutes a challenging retrieval task, where one needs to compare monophonic queries (the solo transcriptions) against polyphonic audio mixtures (music recordings contained in the YouTube videos).

In a first experiment for this advanced retrieval task, we took the same list of candidates $\mathcal{D}_Q^{\text{Text}}$ and parameters as used in Section 4.3 and only exchanged the audio-based queries against solo-based queries. In order to evaluate the results, we took the results $\mathcal{D}_Q^{\text{Rel}}$ from the audio-based retrieval as reference. The solo-based retrieval is considered as correct if among the top-K documents in the ranked list, there is at least one relevant document. The results for this Top-K evaluation measure are shown in **Table 2**.

We retrieve for 85% of the queries a relevant document at rank 1 (Top-1). For 99% of the queries, the first relevant document is within the Top-5 matches. The mean reciprocal rank for the first matches for all queries is 0.91 ($\sigma = 0.22$). Although the solos and audio recordings vary in their degree of polyphony, we reach respectable results. The main reason is that the solo transcriptions are relatively long and perfectly aligned, leading to a high "discriminative power." Furthermore, the queries are very unique, since they stem from an improvisation. When the queries get shorter, usually the discriminative power decreases rapidly, as they may represent more frequently used patterns.

TABLE 2 | Top-K matching rate for the solo-based retrieval.

K	1	3	5	10	15	20
Top-K	0.85	0.97	0.99	1.00	1.00	1.00

The Top-K matching rate is calculated by dividing the Top-K matches by the 329 retrieved solos from the audio-based retrieval.

4.5. Perspectives

So far, our approach for retrieving videos from YouTube relies on either audio recordings or very clean solo transcriptions taken as queries. This is exactly the situation we had in our WJD scenario. In other scenarios, one may have to deal with imperfect or less specific queries. For instance, a query might be a person humming the solo, which then requires an extra step for extracting the fundamental frequency from the hummed melody (query-by-humming), see, e.g., Pauws (2002), Ryynänen and Klapuri (2008), and Salamon et al. (2013). Furthermore, the query may have a different tuning, may be transposed to another key, or played with rhythmic variations. A possible solution could be to use multiple queries, e.g., transposing the query in all possible 12 keys and performing a separate retrieval for each resulting version. Another way of handling some of these issues is to use different feature representations, e.g., features that are robust against temporal deformations, see Arzt (2016), Sonnleitner and Widmer (2016), and Sonnleitner et al. (2016).

In a related retrieval scenario, described in Balke et al. (2016), audio recordings were retrieved from a database containing Western classical music recordings by using monophonic queries with a duration of only a few measures. Besides the discrepancy in the degree of polyphony between query and database documents, tuning, key, and tempo deviations, which frequently occur in Western classical music performances, make this retrieval task very challenging. A common preprocessing step, which targets the "polyphony gap" between query and database document, is to enhance the predominant melody in audio recordings. In Salamon et al. (2013), the authors used a so-called *salience representation* in a query-by-humming system which led to a substantial increase in performance (Salamon and Gómez, 2012). In Balke et al. (2017b), a data-driven approach is used to estimate a salience representation for jazz music recordings which showed a similar performance as the aforementioned, salience-based method.

Another untapped resource for jazz music retrieval is the many publicly available solo transcriptions. However, these transcriptions are typically not available in a machine-readable format. In this case, one could use Optical Music Recognition (OMR) systems to convert sheet music images to symbolic music representations. This conversion may introduce errors, such as missing notes, wrongly detected clefs, key signatures, or accidentals, see Byrd and Schindele (2006), Bellini et al. (2007), Fremerey et al. (2009), Raphael and Wang (2011), Rebelo et al. (2012), Balke et al. (2016). A recently proposed approach for score-following tries to circumvent the difficult OMR step by directly working on the scanned images of the sheet music (Dorfer et al., 2016). Two Convolutional Neural Networks (CNN)—one applied to the sheet music and a second one to the audio recordings—are used for feature extraction. In an extra layer, these features are then combined to retrieve temporal relationships between the two modalities, for instance, with a learned embedding space (Raffel and Ellis, 2016; Dorfer et al., 2017). Currently, new OMR approaches based on deep neural networks show promising

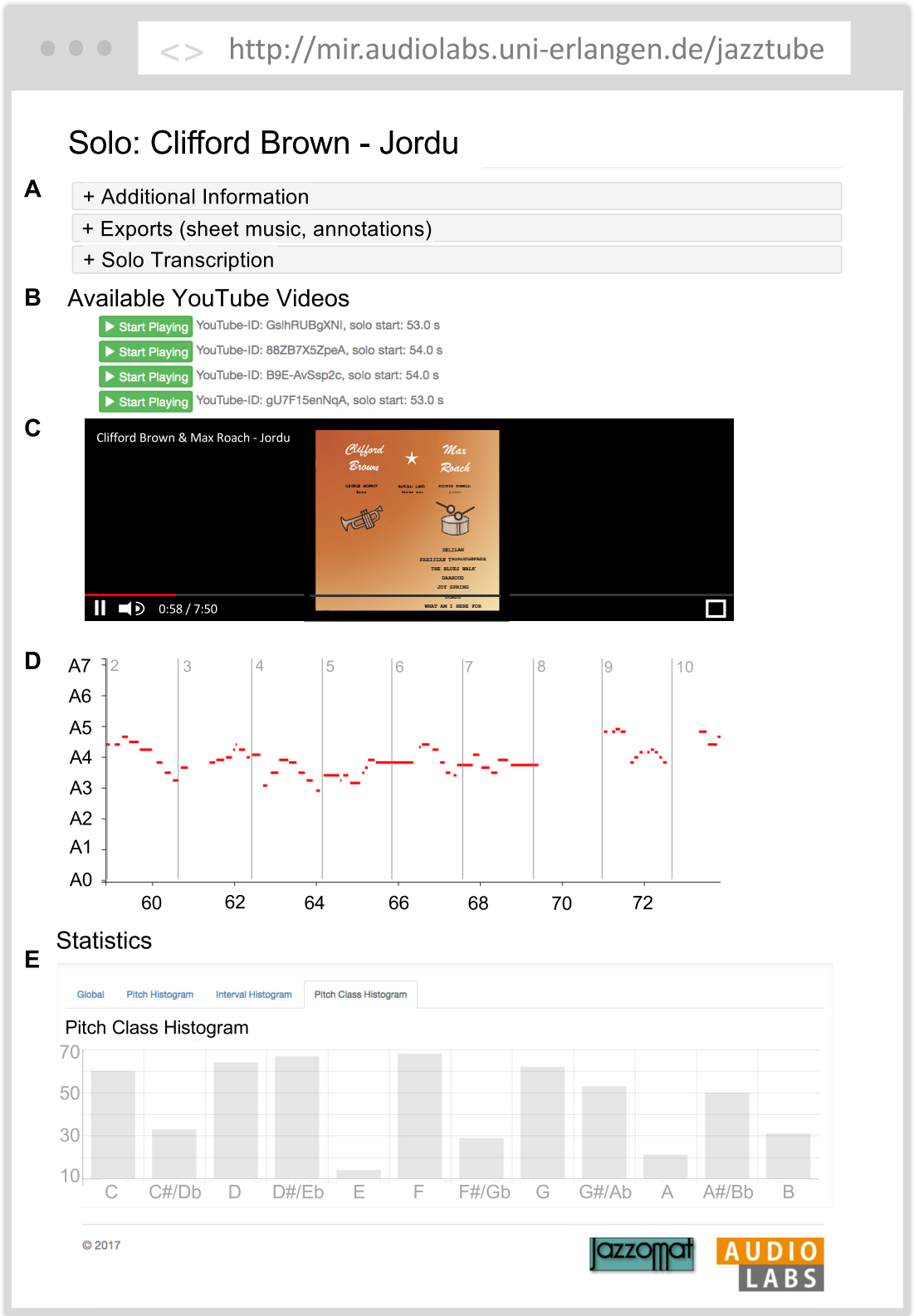

FIGURE 4 | Screenshot of our web-based interface called *JazzTube*. **(A)** Metadata and export functionalities. **(B)** List of linked YouTube videos. **(C)** Embedded YouTube video. **(D)** Piano-roll representation of the solo transcription synchronized with the YouTube video. **(E)** Additional statistics. Panel **(C)** has been created by the authors and, therefore, no permission is required for its use in this manuscript.

results and may lead to a significant increase in conversion quality (Hajič and Dorfer, 2017).

5. APPLICATION

In this section, we present the functionalities of our web-based application, called *JazzTube*, which allows users to easily access the WJD's annotations, as well as the corresponding YouTube videos, in different interactive ways. The application offers various ways to access the WJD. First, tables of the compositions, soloists, and transcribed solos contained in the WJD are given in the form of suitable tables. Furthermore, one can access the information on the record, the track, and at the solo level.

5.1. Solo View

Figure 4 shows a screenshot of the core functionality of our interactive, web-based user interface. In the top panel, some general information about the solo (**Figure 4A**) is shown. Many of these entries are hyperlinks and lead to the artist's overview page or the corresponding track. Furthermore, several possibilities of exporting the solo transcription, either as comma-separated values (CSV) or as sheet music, are offered. The conversion from the annotations to the sheet music is obtained by using the algorithm described in Pfleiderer et al. (2017). Below this basic information, all available YouTube videos are listed (**Figure 4B**). Having more than one match gives alternatives to the user. Note that YouTube videos may have different recording qualities or may disappear from YouTube. After pressing the play button, the corresponding YouTube video is automatically retrieved and embedded in the website (**Figure 4C**). Below the YouTube player, a piano-roll representation of the solo transcription is presented running synchronously with the video playback (**Figure 4D**). Finally, at the bottom, additional statistics about the solo (e.g., pitch histograms) are provided (**Figure 4E**).

5.2. Soloist View

Starting from an overview table of available soloists, the user can navigate to the soloist view containing additional details about the artist. Here, one can also find the available solo transcriptions for the given soloist. Furthermore, *Semantic Web* technologies are used to perform a search query on *DBpedia* to retrieve further details. Usually the received response is very rich in information. Currently, a short biography and a link to the corresponding *Wikipedia* entry for further reading are included. In addition to the biographical data, further relationships to other artists, obtained from the *LinkedJazz* project, are embedded.

5.3. Technical Details

Our web-based demonstrator is a typical client-server application. The client uses the Hypertext Transfer Protocol (HTTP) to perform requests to the server (e.g., by entering an URL through a web browser). These requests are then processed by the server and the response is displayed in the user's web browser. For setting the layout, we use the open-source framework *Bootstrap*.[19] This framework allows for designing a website for different devices (e.g., laptops, tablets, or smartphones). Interactions and animations within the client are realized with *JavaScript*.[20] In particular, a framework called *D3 (Data-Driven Documents)*[21] for visualizing the piano-roll is employed. For the server backend, the Python framework *Flask* is used.[22]

5.4. Possible Advancements for JazzTube

In the case of the Weimar Jazz Database, looking at the scrolling piano-roll visualization of a jazz improvisation while simultaneously listening to the recording could be both of high educational value and a great pleasure. To relate sounding music to moving pitch contours, rhythms, changing event densities, and recurring or contrasting motifs and patterns, which are easily recognizable from a piano-roll visualization, can enrich and deepen the understanding of the tonal, rhythmical, and formal dimensions of the music in an inimitable way. Moreover, recognizing musical passages visually immediately before listening to the sounds can contribute to the play with musical expectancies (or "sweet anticipation" (Huron, 2006)), which lie at the heart of the pleasures of listening to music.

For the future, several extensions to the current form of *JazzTube* are desirable. The piano-roll representation could be extended with different layers of annotations, such as phrases, midlevel units, chords, choruses, form part, or tone formation, which are already available in the WJD. Coloring or annotating events with respect to different functions, e.g., roots of underlying chords, passing tones, or melodic accents, would give an even deeper insight into the inner structure of an improvisation. In the case of jazz, automated identification and annotation of patterns and licks would provide options for analysis not easily achievable with traditional paper and pencil tools. Furthermore, retrieving patterns or motifs from the database would be of great value, for example, by selecting a few tones in a solo and finding and displaying all cross-references in the corpus. Finally, adding options of score-following would be of great help, since music notation is still the standard communication and representation tools of musicians and musicologists. On a different footing, the vast educational implications could be further exploited by adding specialized display options or specifically designed course materials and tutorials (e.g., on jazz history) based on the contents and possibilities of *JazzTube*.

6. CONCLUSION

With *JazzTube*, we offer researchers and music lovers novel possibilities to interact with and navigate through the content of the WJD. With JazzTube's innovative approach to link scientific music databases, including metadata, transcriptions, and further annotations to the corresponding audio recordings that are publicly available via YouTube, copyright restrictions can be bypassed in an elegant way. The approach of *JazzTube* could open up a way

[19] http://www.getbootstrap.com.

[20] https://www.javascript.com.

[21] https://www.d3js.org.

[22] http://flask.pocoo.org.

for music projects to connect metadata and annotations with audio recordings that cannot be freely provided on the internet but can be used for searching for the corresponding audio recordings at YouTube. This could be a way to easily link, e.g., the recording metadata provided within the JDISC[23] project with YouTube recordings. Furthermore, we envision that *JazzTube* is a source for inspiration and fosters the necessary dialog between musicologists and computer scientists to further advance the field of *Digital Humanities*.

AUTHOR CONTRIBUTIONS

Many people have contributed to this paper in various ways. In collaboration with all authors, SB and MM developed the presented concepts and wrote the manuscript. SB carried out the retrieval experiments and realized the web-based interface. CD was involved in technical discussions and writing. MP, JA, and KF were responsible for the data generation and curation of the annotations used in this study. All authors contributed to revisions and additions of the manuscript.

[23] http://jdisc.columbia.edu.

ACKNOWLEDGMENTS

We would like to thank all student annotators of the Jazzomat research project and Patricio López-Serrano for proof-reading the manuscript. SB wants to thank Brian McFee for giving inspiration on web-based systems with his work on the JDISC project together with Dan P. W. Ellis (Columbia University, New York City).

FUNDING

This work was supported by the German Research Foundation (DFG) under grant numbers: MU 2686/6-1 (SB and MM), MU 2686/11-1 (SB and MM), MU 2686/12-1 (SB and MM), MU 2686/10-1 (CD and MM), and PF 669/7-1 (KF, JA, and MP).

REFERENCES

Arzt, A. (2016). *Flexible and Robust Music Tracking*. Ph.D. thesis, Universität Linz, Linz.

Balke, S., Arifi-Müller, V., Lamprecht, L., and Müller, M. (2016). Retrieving audio recordings using musical themes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 281–285. Shanghai, China.

Balke, S., Bießmann, P., Trump, S., and Müller, M. (2017a). Konzeption und Umsetzung webbasierter Werkzeuge für das Erlernen von Jazz-Piano. In *Proceedings of the GI Jahrestagung*, 61–73. Chemnitz, Germany.

Balke, S., Dittmar, C., Abeßer, J., and Müller, M. (2017b). Data-driven solo voice enhancement for Jazz music retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 196–200. New Orleans, USA.

Bellini, P., Bruno, I., and Nesi, P. (2007). Assessing optical music recognition tools. *Comput. Music J.* 31: 68–93. doi:10.1162/comj.2007.31.1.68

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American* 284: 28–37. doi:10.1038/scientificamerican0501-34

Byrd, D., and Schindele, M. (2006). Prospects for improving OMR with multiple recognizers. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 41–46. Victoria, Canada.

Cannam, C., Landone, C., Sandler, M., and Bello, J.P. (2006). The sonic visualiser: a visualisation platform for semantic descriptors from musical signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 324–327. Victoria, Canada.

Cano, P., Batlle, E., Kalker, T., and Haitsma, J. (2005). A review of audio fingerprinting. *J. VLSI Signal Process.* 41: 271–84. doi:10.1007/s11265-005-4151-3

Damm, D., Fremerey, C., Thomas, V., Clausen, M., Kurth, F., and Müller, M. (2012). A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction. *Int. J. Digit. Libr.* 12: 53–71. doi:10.1007/s00799-012-0087-y

Dorfer, M., Arzt, A., and Widmer, G. (2016). Towards score following in sheet music images. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 789–795. New York, USA.

Dorfer, M., Arzt, A., and Widmer, G. (2017). Learning audio-sheet music correspondences for score identification and offline alignment. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 115–122. Suzhou, China.

Fremerey, C., Müller, M., and Clausen, M. (2009). Towards bridging the gap between sheet music and audio. In *Knowledge Representation for Intelligent Music Processing*, Edited by E. Selfridge-Field, F. Wiering, and G.A. Wiggins, 9051. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik.

Frieler, K., and Pfleiderer, M. (2017). Onbeat oder offbeat? Überlegungen zur symbolischen Darstellung von Musik am Beispiel der metrischen Quantisierung. In *Proceedings of the GI Jahrestagung*, 111–125. Mainz, Germany.

Gasser, M., Arzt, A., Gadermaier, T., Grachten, M., and Widmer, G. (2015). Classical music on the web – user interfaces and data representations. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 571–577. Málaga, Spain

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., et al. (2017). Audio set: an ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 776–780. New Orleans, USA.

Goto, M. (2011). Music listening in the future: augmented music-understanding interfaces and crowd music listening. In *Proceedings of the Audio Engineering Society (AES) Conference on Semantic Audio*, Ilmenau, Germany.

Goto, M. (2014). Frontiers of music information research based on signal processing. In *Proceedings of the International Conference on Signal Processing (ICSP)*, 7–14. Hangzhou, China.

Goto, M., Yoshii, K., Fujihara, H., Mauch, M., and Nakano, T. (2011). Songle: a web service for active music listening improved by user contributions. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 311–316. Miami, Florida.

Hajič, J. Jr., and Dorfer, M. (2017). Prototyping full-pipeline optical music recognition with musicmarker. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking Session*, Suzhou, China.

Huron, D.B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press, Cambridge, Massachusetts.

Knees, P., and Schedl, M. (2016). *Music Similarity and Retrieval*. Springer Verlag, Berlin, Heidelberg.

Kurth, F., Müller, M., Damm, D., Fremerey, C., Ribbrock, A., and Clausen, M. (2005). SyncPlayer – an advanced system for multimodal music access. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 381–388. London, UK.

Liem, C.C.S., Gómez, E., and Schedl, M. (2015). PHENICX: innovating the classical music experience. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–4. Torino, Italy.

Liem, C.C.S., Gómez, E., and Tzanetakis, G. (2017). Multimedia technologies for enriched music performance, production, and consumption. *IEEE MultiMedia* 24: 20–3. doi:10.1109/MMUL.2017.20

McFee, B., McVicar, M., Nieto, O., Balke, S., Thomé, C., Liang, D., et al. (2017). *Librosa 0.5.0*. Zenodo. doi:10.5281/zenodo.293021

Melenhorst, M.S., van der Sterren, R., Arzt, A., Martorell, A., and Liem, C.C. (2015). A tablet app to enrich the live and post-live experience of classical concerts. In

Proceedings of the International Workshop on Interactive Content Consumption (WSICC), Brussels, Belgium.

Müller, M. (2007). *Information Retrieval for Music and Motion*. Springer Verlag, Berlin, Heidelberg.

Müller, M. (2015). *Fundamentals of Music Processing*. Springer Verlag, Berlin, Heidelberg.

Müller, M., and Ewert, S. (2011). Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 215–220. Miami, Florida.

Müller, M., Kurth, F., and Clausen, M. (2005). Chroma-based statistical audio features for audio matching. In *Proceedings of the IEEE Workshop on Applications of Signal Processing (WASPAA)*, 275–278. New Paltz, NY.

Pattuelli, M.C. (2012). Personal name vocabularies as linked open data: a case study of Jazz artist names. *J. Info. Sci.* 38: 558–65. doi:10.1177/0165551512455989

Pauws, S. (2002). CubyHum: a fully operational query by humming system. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 187–196. Paris, France.

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., and Burkhart, B. eds. (2017). *Inside the Jazzomat. New Perspectives for Jazz Research*. Mainz, Germany: Schott Campus.

Prätzlich, T., Müller, M., Bohl, B.W., and Veit, J. (2015). Freischütz digital: demos of audio-related contributions. In *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Malága, Spain.

Raffel, C., and Ellis, D.P.W. (2016). Pruning subsequence search with attention-based embedding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 554–558. Shanghai, China.

Raimond, Y., Abdallah, S., Sandler, M., and Giasson, F. (2007). The music ontology. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 417–422. Vienna, Austria.

Raphael, C., and Wang, J. (2011). New approaches to optical music recognition. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 305–310. Miami, FL.

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R.S., Guedes, C., and Cardoso, J.S. (2012). Optical music recognition: state-of-the-art and open issues. *Int. J. Multimedia Information Retr.* 1: 173–90. doi:10.1007/s13735-012-0004-6

Röwenstrunk, D., Prätzlich, T., Betzwieser, T., Müller, M., Szwillus, G., and Veit, J. (2015). Das Gesamtkunstwerk Oper aus Datensicht – Aspekte des Umgangs mit einer heterogenen Datenlage im BMBF-Projekt "Freischütz Digital". *Datenbank Spektrum* 15: 65–72. doi:10.1007/s13222-015-0179-0

Ryynänen, M., and Klapuri, A. (2008). Query by humming of MIDI and audio using locality sensitive hashing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2249–2252. Las Vegas, NV.

Salamon, J., and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio Speech Lang. Process.* 20: 1759–70. doi:10.1109/TASL.2012.2188515

Salamon, J., Serrà, J., and Gómez, E. (2013). Tonal representations for music retrieval: from version identification to query-by-humming. *Int. J. Multimedia Info. Retr.* 2: 45–58. doi:10.1007/s13735-012-0026-0

Schoeffler, M., and Herre, J. (2014). The influence of audio quality on the popularity of music videos: a YouTube case study. In *Proceedings of the International Workshop on Internet-Scale Multimedia Management*, 35–38. Orlando, Florida, USA.

Sonnleitner, R., Arzt, A., and Widmer, G. (2016). Landmark-based audio fingerprinting for DJ mix monitoring. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 185–191. New York City, NY.

Sonnleitner, R., and Widmer, G. (2016). Robust quad-based audio fingerprinting. *IEEE Trans. Audio Speech Lang. Process.* 24: 409–21. doi:10.1109/TASLP.2015.2509248

Thomas, V., Fremerey, C., Damm, D., and Clausen, M. (2009). SLAVE: a score-lyrics-audio-video-explorer. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 717–722. Kobe, Japan.

Weihs, C., Jannach, D., Vatolkin, I., and Rudolph, G. (2016). *Music Data Analysis: Foundations and Applications*. CRC Press, Abingdon, UK.

Conflict of Interest Statement: This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.