KNOWLEDGE IS OUT THERE: A NEW STEP IN THE EVOLUTION OF MUSIC DIGITAL LIBRARIES

Sergio Oramas and Mohamed Sordo

Introduction

A vast amount of musical knowledge has been gathered for centuries by musicologists and music enthusiasts. Most of this knowledge is implicitly expressed in artist biographies, reviews, facsimile editions, etc. Music digital libraries make this information available and searchable. Documents are indexed and a keyword-based search is generally provided. However, implicit knowledge present in text, scores, and audio is not fully understood by machines, so complex queries cannot be answered.

As a first step towards the approximation of machine understanding to accumulated musicological knowledge, items stored in digital libraries must be semantically annotated. Current descriptive metadata and markup annotations provide some structured information. However, it is insignificant compared with the epistemic potential of the content itself. Once documents are properly annotated, complex structures and meaningful relations between pieces of information may emerge. This supports a paradigm shift, from keyword-based systems to knowledge-based systems, thereby enabling musicologists to formulate more complex queries. As a consequence, digital libraries may turn into real knowledge environments, instead of mere searchable repositories¹.

Manual annotation of documents is very time consuming, and sometimes unwieldy. Thus, the use of reliable, automatic processes is crucial for building knowledge environments. In the last few years several studies and approaches on the use of Semantic Web technologies in digital libraries have been proposed. The result of this intersection has been coined as ‘Semantic Digital Libraries’. Most of the related work is focused on the acquisition of Semantic Web methodologies for knowledge representation, which, among other advantages, facilitates information exchange between multiple sources. In the case of music digital libraries though, tools and methodologies developed around the Semantic Web for automatic knowledge acquisition have received less attention.

This article provides an extensive survey about the applicability and usefulness of state-of-the-art semantic technologies for knowledge acquisition in music digital libraries. In addition, an analysis of the evolution of music libraries from a technological perspective is presented. Finally, a methodology to exploit implicit knowledge present in collections of text documents is proposed and tested on the New Grove Dictionary, providing an overview of the possibilities that a knowledge layer may offer to music digital libraries.

Background

Evolution of Music Libraries

Music libraries can be classified according to their level of technology development. Following this criterion, a pyramid can be constructed, defining the different evolution states of a music library (fig. 1). The base of the pyramid represents traditional libraries, where items are physical, such as books, scores, manuscripts, slate records, etc. Items are often classified into catalogues and indices. In last few decades, digitisation of content has raised a new way of storing and making items available. They can be replicated, and even accessed online from anywhere in the world. Scores and books are scanned and the resulting images are stored in databases. Music can also be stored using digital audio formats. Thus, items can be consulted on a digital device (e.g., computer, smart phone, tablet). This has been the first great revolution for music libraries. Before that, musicologists had to be physically in the library to study texts or listen to audio. Nowadays, thanks to the efforts put into the digitisation and publication of content, musicologists and the general public may have access to items everywhere and at any time.

Digitised items are generally stored in a database along with contextual information about them (e.g., title, author, date of publication, etc.). This contextual information is called metadata, which can be exploited by the search system of a digital library. Information retrieval is a discipline of obtaining information relevant to a query based on metadata or on full text or other content-based indexing. Therefore, information retrieval techniques can be applied to digital libraries to navigate and search within the library. However, these kinds of digitised items are not readable by the computer, so it is not possible to perform searches directly on the content of the item. Hence, there is a need of a further step in the evolution of digital libraries: transforming digitised items into machine-readable ones. To do so, item content must be transcribed into a proper format. Transcription can be done manually by humans, or automatically by a computer. Different techniques can be applied to perform this automatic transcription process. Scanned texts and scores can be converted into machine-readable formats using optical character recognition (OCR) systems. Texts can be stored in TXT, HTML, or XML formats, and scores in MIDI or XML. Low-level audio features (e.g., beats per minute, average pitch, timbre stability) can be extracted from audio files using music information retrieval techniques and stored as metadata. Once the content is transcribed, the library search system may have access to the content itself. In the case of text documents, information retrieval techniques for searching documents can be applied on the whole content of the item, not only on the metadata. Moreover, scores and audio files can be queried using feature values and ranges (e.g., songs with beats per minute higher than 120, songs in a minor key).

At this stage, digital libraries have significantly increased their ability to provide concrete answers to their users. However, the epistemic potential of the content is not yet being exploited. Computers are able to find patterns of words, but they do not understand the meaning of texts. Users can do simple text queries, but cannot ask complex questions. For example, imagine a digital library with thousands of biographies, scores, and audio files from the Baroque period. A user wants to know the artists that were born in Italy and

---

had worked in France, or pieces in sonata form from German composers born in the second half of the seventeenth century. In their current state, digital libraries could not possibly provide a direct answer; it would be necessary to spend an inordinate amount of time consulting different documents. Therefore, digital libraries should go a step further to make the content understandable by the computer, and provide the user with the tools to exploit this knowledge. One way to do so is by adding semantic annotations to the content. This annotation process can be done either manually or automatically. Manual annotation requires tremendous human effort, and most of the time not feasible. As such, automatic processes for semantic annotation are necessary to extract the knowledge behind the lines present in library texts. A tremendous effort has been made in this direction in the Semantic Web and Natural Language Processing (NLP) communities, developing tools and methodologies for Named Entity Recognition and Disambiguation, Relation Extraction, and Information Extraction\(^3\). The technology to make it possible is already

there, and it is currently being applied to search engines on the Web. However, it has barely been used in current music digital libraries. In the case of music audio files, high-level descriptors (e.g., instrument, style, mood) may be extracted by a properly trained music information retrieval system.

Once library systems are able to understand all their content, the next step would be to become intelligent systems. At this stage, library systems would be able to perform their own reasoning processes by interconnecting knowledge from different documents and sources. They would be more a research partner than a mere search tool, and would be able to perform a complete research task from beginning to end and arrive at their own conclusions. Instead of search boxes, these library systems would display request boxes where musicologists can ask specific questions. The Web retrieval research community is already moving in this direction. Music digital libraries should take advantage of these technological developments to improve their systems and advance towards a better understanding and dissemination of musical knowledge.

The Semantic Web

The Web is currently shifting from a Web of hyper-linked documents to a Web of linked data. This new paradigm is called the Semantic Web, and brings a knowledge-driven framework for data interconnection. The Semantic Web pursues a similar goal as digital libraries: transforming the Web from a searchable repository to a knowledge environment.

To build a knowledge environment, data must be interconnected and semantically annotated. This would allow computers to understand meanings and relationships between items of information. The aim of using common standards from the Semantic Web, such as ontologies and the Resource Description Framework (RDF) specification, is to improve the interoperability and knowledge sharing among libraries, giving rise to a new generation of digital libraries: Semantic Digital Libraries. This interconnection is being employed in the Linked Open Data (LOD) cloud, a core element of the Semantic Web. A significant amount of freely available data has been already published following these principles.

Knowledge Extraction

Knowledge Extraction (KE) can be defined as the creation of knowledge from structured and unstructured text sources. KE is also referred to as Information Extraction (IE) in the Natural Language Processing (NLP) community, whose major challenge is to understand natural language.

A major step towards understanding language is the extraction of meaningful terms (entities) from text as well as the relationships between those entities. The latter is called Relation Extraction, and is an established task in NLP, defined as the process of identifying and annotating relevant semantic relations between entities in text.

Another crucial task in the process of IE is to determine the identity and category of entity mentions present in a text. In NLP, this task is called Entity Linking (EL). EL con-

---

4. Xiao Yu, Hao Ma, Bo-June Hsu and Jiawei Han, “On Building Entity Recommender Systems Using User Click Log and Freebase Knowledge”, in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York: Association for Computing Machinery, 2014).


A NEW STEP IN THE EVOLUTION OF MUSIC DIGITAL LIBRARIES

The idea is to gather the content and metadata from all the documents present in the library and then apply a process of semantic enrichment. This process consists in applying different processes of Information Extraction, such as Entity Linking and Relation Extraction. The goal is to build a Knowledge Base (KB) expressing the relations between entities present in the library. KB is a comprehensive, semantically organised, machine-readable collection of entities, classes, and facts (i.e., attributes and relations). According to Wikipedia, an entity is something that exists in itself, actually or potentially, concretely or abstractly, physically or not. Mozart, the Juilliard School, Baroque, Paris are examples of entities.

An item in the library corresponds to an entity in the KB. In addition, a process of Entity Linking is applied to the content of every item, and every entity mention detected is also added to the KB. Entities can be related by different kinds of relations. These relations can be gathered from metadata, or extracted from the text itself by applying Relation Extraction techniques. Moreover, the KB can be related to, or linked to other existing KBs, thus leveraging the power of the Linked Open Data (LOD) initiative.

Once the KB is built, a graph representation of the data can be easily constructed. This representation can be expressed as a simple graph, or as an RDF graph. This is called a Knowledge Graph, which can be exploited in different applications, such as artist relevance, artist similarity, music recommendation, or information visualisation.

---

Experimental Setup

In this section, the proposed methodology is applied to a music digital library. The objective here is to illustrate the possibilities of discovering new knowledge, and to help musicologists to use a digital library.

Hypothetically, a musicologist may want to find out the answer to the following questions:

- What are the most relevant music schools?
- Who are the artists most similar to Schoenberg?
- Is there a migration tendency in artists? If so, to which cities?

The purpose of this experiment is to show how a system using the proposed methodology is able to answer these questions, or at least, provide to the musicologist some insight into finding the proper answer.

All experiments were carried out using the Python programming language, plus some NLP related Web services. The source code of the experiments can be found online. The rest of this section is organised as follows: 1) exposing the dataset used for the experiments; 2) the process for the creation of the knowledge graph is described next; 3) then the obtained data is analysed and some insights from its analysis are shown, and finally, 4) a knowledge graph is used to compute entity relevance and entity similarity.

Dataset: The New Grove

The Grove Dictionary of Music and Musicians is an encyclopedic entity, one of the largest reference works on Western music. George Grove first published it in the last quarter of the nineteenth century. In 1980, a new version called The New Grove was released in twenty volumes, consisting of 22,500 articles and 16,500 biographies. The complete text of the second edition of The New Grove is available in machine-readable format on the online service Grove Music Online.

We automatically gathered the first paragraph of every biography classified in the section People in history in the Grove Music Online. These biographies are related to artists from different periods of the history of music, from pre-medieval time to contemporary. We obtained a total of 16,707 biographies.

Knowledge Graph Creation

The Knowledge Graph was created as follows. First, the subject of every gathered biography was added as an entity to the Knowledge Graph. Then, Entity Linking was applied to every biography text. We used two different Entity Linking procedures. The first consisted in building a gazetteer, a list of artist subjects in the library, and performing a simple string matching on the biography texts, looking for entity mentions. This way, we detected most of the auto references of entities within the library. The second procedure dealt with the detection of other types of entities, such as educational institutions, music genres, places and concert venues. To detect these entities, we used a state-of-the-art

---

Entity Linking tool called Tagme\textsuperscript{12}. Tagme is able to detect mentions of Wikipedia entities in English texts. In addition, Wikipedia pages are classified into categories according to the DBpedia ontology\textsuperscript{13} (DBpedia is a structured version of Wikipedia). We used this classification to filter out entity mentions, keeping only those of the desired type. After the Entity Linking process, we added every identified entity to the graph. Then, we added an edge relating every mentioned entity with the corresponding subject entity of the biography where it was found.

By observing the biographical texts, we detected some common structures. For example, at the beginning of every biography there is a sentence in parentheses with information about the place and date of birth and death. In addition, the second sentence of the biography always describes the role of the biographical subject. Therefore, we applied a process of Information Extraction to obtain the roles, and the place and year of birth and death from every biography. We obtained birth information for 14,355 subjects, and death information for 10,741. For the roles, we collected 434 different ones. The most represented roles in the dataset are shown in Table 1. After the extraction process, we obtained

Table 1. Most Representative Roles in the Dataset

<table>
<thead>
<tr>
<th>Role</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>composer</td>
<td>2,618</td>
</tr>
<tr>
<td>teacher</td>
<td>1,065</td>
</tr>
<tr>
<td>conductor</td>
<td>968</td>
</tr>
<tr>
<td>pianist</td>
<td>704</td>
</tr>
<tr>
<td>organist</td>
<td>676</td>
</tr>
<tr>
<td>singer</td>
<td>404</td>
</tr>
<tr>
<td>violinist</td>
<td>285</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>musicologist</td>
<td>144</td>
</tr>
<tr>
<td>critic</td>
<td>133</td>
</tr>
</tbody>
</table>

\begin{thebibliography}{99}
\bibitem{Tagme} http://tagme.di.unipi.it/, accessed 16 September 2016.
\bibitem{DBpedia} http://wiki.dbpedia.org/services-resources/ontology, accessed 16 September 2016.
\end{thebibliography}
a Knowledge Graph with 47,367 nodes (i.e., entities) and 274,333 edges (i.e., relations) relating those nodes. An excerpt of the graph is shown in Figure 3.

**Analysing the Data**

With the amount of births per decade a histogram could be built (fig. 4). In this histogram the distribution of artist births alongside history is represented. We easily observe that there is a peak of births in the second half of the sixteenth century and in the second half of eighteenth century. Note that the scope of this research is not to extract musicological conclusions, but to show in what way semantic technology may help the musicologist.

We also counted the number of births and deaths by country. Table 2 shows the number of births and deaths and the perceptual difference between both values for five countries with more births. We observe that the United States and Italy have a negative percentage. This means that there is a migratory tendency of their artists. However, France has a positive value, so it has absorbed artists from abroad. The same count for the five cities with more births is shown in Table 3. All these cities have an absorbing tendency. Perhaps the case of Paris draws more attention with an increase of 137%.

**Entity Relevance**

Entity mentions identified in the biographies can be seen as hyperlinks that connect to other entities. Therefore, algorithms to measure the relevance of nodes in a network of
hyperlinks can be applied to our semantic graph. In order to measure entity relevance, we applied the PageRank algorithm to the resultant graph. We calculated the most

---

relevant entities of five different categories: cities, venues, educational institutions, biography subjects, and genres.

- **Venue**: Covent Garden Theatre, King’s Theatre, Drury Lane, Carnegie Hall, Théâtre de la Monnaie, Stadttheater, Theatre Royal
- **Educational Institution**: Paris Conservatoire, Moscow Conservatory, Juilliard School, St Petersburg Conservatory, Prague Conservatory, Leipzig Conservatory, Vienna Hochschule für Musik
- **Biography Subject**: Haydn, Claude Debussy, Arnold Schoenberg, Robert Stevenson, Paul Hindemith, Giovanni Pierluigi da Palestrina, Gustav Mahler, Maurice Ravel, Jean-Philippe Rameau
- **Genre**: chamber music, a capella, jazz, folk music, avant garde, baroque music, electronic music, musical theatre, plainchant

PageRank calculates relevance according to the number of in-links and the relevance of the entities from which those in-links came from. Looking at our graph, we can observe for instance that the Paris Conservatoire has an in-link that comes from its citation in the Claude Debussy biography, who indeed is an entity with a very high relevance. Following this idea, the relevance of every entity in the graph is computed.

Initially, the results seem to be quite coherent. However, there are some clear errors. For example, Robert Stevenson appears to be one of the most important subjects in the *New Grove*. The problem here is that there are three different Robert Stevensons. Moreover, Robert Stevenson is the name of the author of 300 biographies in the *New Grove*, and his name appears at the end of these biographies. Thus, our system faces two problems. First, it cannot disambiguate between the different Robert Stevensons, and second, it takes the name of the author of the biography as a referenced entity, which should be omitted. The same happens with Haydn, there are different Haydns in the *New Grove* (e.g., Joseph Haydn, Haydn Draper, Haydn Wood, etc.). Therefore, entity relevance must be further analysed and the introduction of a disambiguation algorithm in the Entity Linking step is necessary.

Apart from the described errors, the results are quite accurate. If we would ask different musicologists for the most relevant entities in these categories, there may not, perhaps, be agreement on the order. However, there would be a set of entities that would be categorised as relevant for all of them. The most important insight from the calculated set of entities is not the order or the exact names, but the coherence of the results obtained by the system without any human intervention. There are some errors and some missing entities, but the results are quite accurate and may provide a quick insight into the data for a musicologist. Further work on the applied algorithms would lead to a better accuracy.
of the results, but this is out of scope for this research. In the section on FlaBase below, an evaluation of the entity relevance methodology is carried out on a similar Knowledge Graph.

Entity Similarity

Another application of the Knowledge Graph is to compute the list of top-N similar entities of a given entity. We used a semantic-based approach\(^\text{16}\) for computing the list of the top five more similar entities to all the artist entities. This approach is based on the comparison of the neighborhoods of entities within the graph. An example of the list of most similar entities of four different artists is illustrated here.

- **Arnold Schoenberg**: Anton Webern, Paul Hindemith, Gustav Mahler, Alban Berg, Claude Debussy
- **Guido Adler**: Heinrich Jalowetz, Eusebius Mandyczewski, Robert Fuchs, Karl Weigl, Anton Wranitzky
- **Manuel de Falla**: Ricardo Viñes, Juan Vicente Lecuna, Enrique Granados, Miguel Llobet Soles, Luigi Russolo
- **Miles Davis**: Dizzy Gillespie, Herbie Hancock, Paul Chambers, Tony Williams, Cannonball Adderley

Schoenberg is related to artists from the same period and also with a strong relation. If we look into the common entities in the neighborhood graph of Schoenberg and Webern, we discover that they are directly related. In addition, they are related by many other entities, such as Theodor W. Adorno, Alban Berg, Vienna, or the Darmstadt School. In the list of Guido Adler’s similar entities it is relevant to see that they are mostly musicologists. For instance, Guido Adler and Heinrich Jalowetz are directly related as Adler taught Jalowetz in Vienna. In the same way, de Falla’s most similar entities are Spanish composers, while Miles Davis is related to jazz musicians that played with him.

These lists might not be completely accurate to all musicologists, but it is clear that all the similar entities shown are closely related to their seed entity. Therefore, the system has again been able to show meaningful results from a set of texts without any human aid.

Mixing Different Data Sources

When there is information coming from different data sources instead of only one digital library, there are some issues that must be resolved before applying the semantic enrichment process described above. Thus, an extension of the previous methodology has been defined (see fig. 5). We added two more steps at the beginning of the pipeline. First, we gathered all available content from the different data sources. Second, it was mandatory to decide whether two entities coming from different sources were in fact the same. This later process is called Entity Resolution, and it is focused on the problem of identifying and linking different manifestations of the same real world object\(^\text{17}\). Once the data is properly harvested and mapped, we can proceed with the semantic enrichment processes.


FlaBase

To illustrate an example of a KB built from different sources we used FlaBase. FlaBase, or Flamenco Knowledge Base, is the acronym of a new KB of flamenco music that we have created. Its ultimate goal is to gather all available online editorial, biographical, and musicological information related to flamenco music. FlaBase combines information gathered from general KBs (e.g., Wikipedia, DBpedia), online music encyclopedias (e.g., MusicBrainz [https://musicbrainz.org/, accessed 16 September 2016]), and specialised flamenco Web sites (e.g., http://www.andalucia.org [accessed 16 September 2016]). FlaBase does not contain any information from the New Grove, as this KB was created only with content extracted from open data repositories.

The most difficult problem in the creation of the KB was related to the Entity Resolution step. When two or more data sources (e.g., Wikipedia and MusicBrainz) have an entry about the same entity (e.g., Camarón, the flamenco singer), but those entries are identified with different labels according to the source (e.g., Camarón, Camarón de la Isla, or Jesús López Monje), how can our system automatically determine if they refer to the same artist? Several techniques can be applied, from string similarity to graph theory, by exploiting the entity label or contextual information of the entities. To build FlaBase, we applied a pair-wise string similarity approach. Further information about the creation process of FlaBase is available in an earlier publication18.

Once the data was gathered and combined, a process of semantic enrichment was applied following a similar methodology like that of creating a Knowledge Graph discussed above. Based on the resulting graph, we computed entity relevance by applying the PageRank algorithm.

We built an ordered list with the top ten entities of different artist categories present in the KB: cantaor (i.e., flamenco singer), guitarist and bailaor (i.e., flamenco dancer) (see Table 4).

For evaluation purposes, we asked a flamenco expert to build a list of top ten artists for each category according to his knowledge and the available literature. Thus, the expert

---

Table 4. Top Five List of Relevant Artists Determined by the Algorithm

<table>
<thead>
<tr>
<th>Cantaor</th>
<th>Guitarist</th>
<th>Bailaor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonio Mairena</td>
<td>Paco de Lucía</td>
<td>Antonio Ruiz Soler</td>
</tr>
<tr>
<td>Manolo Caracol</td>
<td>Ramón Montoya</td>
<td>Rosario</td>
</tr>
<tr>
<td>La Niña de los Peines</td>
<td>Niño Ricardo</td>
<td>Antonio Gades</td>
</tr>
<tr>
<td>Antonio Chacón</td>
<td>Manolo Sanlúcar</td>
<td>Mario Maya</td>
</tr>
<tr>
<td>Camarón de la Isla</td>
<td>Sabicas</td>
<td>Carmen Amaya</td>
</tr>
</tbody>
</table>

provided us with this hypothetical list of the top ten artists by category, and we considered it as basic truth. We define precision as the number of identified artists in the resulting list obtained by our system that are also present in the basic truth list provided by the expert, divided by the length of the expert list. We evaluated the output of the algorithm by calculating precision over the entire list (top ten), and over the first five elements (top five), obtaining a value of 0.633 and 0.933, respectively. A high value of precision, especially for the top five list, indicates that the content gathered in FlaBase is highly accurate, as is the method applied for computing artist relevance.

Conclusions

This article presents an analysis on the evolution of music digital libraries based on recent advancements in technology, and further explained that music digital libraries are still in the early stage of development compared to the latest developments in Web searching. In addition, a methodology was proposed to exploit implicit knowledge present in collections of text documents from digital libraries. This methodology has been applied on a corpus of artist biographies gathered from the New Grove. Results illustrate the suitability of this methodology for knowledge discovery, and how it can help musicologists to seek out information in a large amount of documents. Finally, an extension of the methodology was proposed to combine different data sources, and was illustrated by the creation of a flamenco music Knowledge Base.

It is clear that music digital libraries can strongly benefit from these knowledge-based approaches, allowing musicologists to not only search content, but also to discover new or hidden insights in the data.

English Abstract

Musicological knowledge is spread between the lines of thousands of texts stored in hundreds of music libraries. Technology, and more recently semantic technologies, may play a key role in the way information is retrieved. In this article, an analysis of the evolution of music digital libraries from a technological perspective is presented. Then, a methodology to exploit implicit knowledge present in collections of text documents is proposed. The described methodology is applied over a
set of 16,707 artist biographies gathered from *Grove Music Online*. A knowledge graph is then created after applying a process of information extraction. Several insights are extracted from the data to illustrate the possibilities of the proposed methodology for musicologists. Finally, an extension of the methodology is proposed to combine different data sources, and illustrated with the creation of a flamenco music knowledge base.

**French Abstract**
Les connaissances musicologiques sont étalées entre les lignes de milliers de textes qui sont conservés dans des centaines de bibliothèques. La technologie, et plus récemment les technologies sémantiques, pourraient jouer un rôle clé dans la manière dont les informations sont recherchées. Cet article présente une analyse de l'évolution des bibliothèques musicales numériques d'un point de vue technologique, et propose une méthodologie qui vise à exploiter les connaissances présentes dans les collections de documents textuels. La méthodologie décrite s'applique à plus de 16 707 biographies d'artistes rassemblées à partir du *Grove Music Online*. Un graphique représentant les différents détails, d'après les requêtes, est ensuite établi par voie d'extraction d'information. Plusieurs aperçus sont extraits des données afin d'illustrer les différentes possibilités proposées par la méthodologie, à destination des musicologues.

Enfin, une extension de la méthodologie est proposée pour pouvoir combiner diverses sources de données, illustrée par la création d'une base de connaissance de flamenco.

**German Abstract**