# AUTOMATIC CREATION OF KNOWLEDGE GRAPHS FROM DIGITAL MUSICAL DOCUMENT LIBRARIES

Sergio Oramas, Mohamed Sordo and Xavier Serra

*Music Technology Group, Universitat Pompeu Fabra*

Correspondence should be addressed to: sergio.oramas@upf.edu

***Abstract:*** **Most of the current musicological knowledge is present in printed books and manuscripts. In the last years greats efforts have been done in order to digitize and make available these documents in form of Digital Libraries. However, digital documents are mainly stored as raw text, with no more structure than indexes and some metadata. Therefore, implicit knowledge contained in text is not understandable by computers and cannot be processed like that. Automatic processing of text documents may help musicologists in several ways, such as improving navigation through a library, discovering hidden knowledge, accelerating tedious tasks, etc. To apply these techniques to a Digital Library, the information contained in documents should be carefully structured and semantically annotated. Information Extraction is a discipline of computer science focused on the extraction of structured information from unstructured text sources. We propose a method to automatically extract meaningful knowledge from documents present in Digital Musical Document Libraries, by using Information Extraction techniques. Our method has two main steps. First, relevant named entities (e.g. composers, organizations, places, etc.) are identified in the text. Second, words between these entities are syntactically and semantically analyzed to understand the relationship between them. Finally, the extracted knowledge is represented in a machine-readable format as a knowledge graph, where entities are represented as nodes, and relations as edges. The resulting knowledge representation is finally visualized as an interactive graph. With the proposed information visualization, users may go from one document to another by browsing the knowledge graph. We tested our method with a subset of artist biographies present in the Grove Music Online.**

## 1. INTRODUCTION

Exploring connections between music related entities such as artists, works or places is a central activity of musicology [1]. To find these connections, the musicologist generally look for related information in written documents. These documents are mainly organized in compilations and collections. Entire collections have been digitized, and many of them are in a machine-readable format, which has signified a great improvement on the way information is accessed. These digitized collections are stored and managed by information systems where documents can be searched by textual keywords. This improvement have increased significantly the possibilities for musicologists in the way information is gathered. However, these Keyword-based search engines tend to have high precision but low recall [2]. That is, although precise results are often given, many relevant documents may not be obtained. Semantic relations present in text are not understood by search engines, so the actual meaning of text is undertaken. Hence, to take advantage of the epistemic potential of digital libraries, it is not enough to put them online, and make them searchable. We have to transform them from searchable repositories to knowledge environments as the next step in the evolution of digital libraries [3].

The World Wide 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. Nevertheless, Digital Libraries are already one step ahead in that direction, because they use indexes crafted from reliable metadata. Since metadata can express concepts not explicitly occurring in the document, the use of metadata indexes leads to better precision and recall in information services [4]. To build a knowledge environment, data must be interconnected and semantically annotated. This would let computers to understand meanings and relationships between pieces of information. The aim of using common standards from the Semantic Web such as ontologies and the RDF specification[1] is to improve the interoperability and knowledge sharing among libraries, giving rise to a new generation of Digital Libraries: the Semantic Digital Libraries.

A central element of the Semantic Web is the Linked Open Data (LOD) initiative. A huge amount of freely available data have been published following this principles. Digital Libraries may learn two main concepts from the Semantic Web. First, by using standardized knowledge representations. Thus, already existing semantic technologies can be used with Digital Libraries. Second, by taking advantage of interconnection possibilities. That is, by publishing, reusing and interlinking content of Digital Libraries with available datasets from the LOD. To this end, information inside a Digital Library should be carefully annotated and structured. Most of the information present in Digital Libraries is stored as raw text, classified by topics, title or categories. Descriptive metadata, semantic markup or hyperlinks give some structure, but complex and multidimensional structures, patterns and relationships remain implicit, latent or hidden in the library [3]. Normally, you cannot query the system with questions like: who is Mozart's father, or what composers were born in Vienna in the second half of the XVIII century. To achieve this goal, information should be structured and interconnected. Manual annotation of documents is highly time consuming and also very costly in terms of human resources. Information Extraction techniques may help in the automatization of this task.

Another significant aspect of Digital Libraries is how the information is accessed. Information visualizations may be useful to explicitly represent complex information, making Digital Libraries more cognitively, accessible and epistemically useful [3]. Inherent relations between entities in a text can be expressed using knowledge graphs. In the graph, entities are represented as nodes, and edges represent relations between those entites [5]. The ability to visualize and navigate this large graphs is crucial for the end-user experience. Therefore, integrating graphical visualizations and knowledge graphs in Musical Digital Libraries may help musicologists to discover new knowledge, browsing the library through the graph.

In this work, an Information Extraction method was applied to artist biographies present in the Grove Music Online digital library. With extracted entities and relations a knowledge graph was created. The graph represents implicit and hidden knowledge present in the text of the biographies. A navigation tool was then developed to browse the library using a graph visualization interface. At the same time, information is also enriched with links to external resources present in the LOD cloud.

The reminder of the paper is structured as follows. Section 2 reviews relevant related work to digital libraries and Information Extraction. Then Section 3 describes the dataset gathered from the Grove Music Online. Section 4 defines the method applied for Information Extraction, and Section depicts the resultant knowledge subspace and its visualization. Finally, Section 6 concludes the paper and points out for future lines of work.

[1] http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

## 2. Related Work

### 2.1. Digital Libraries

There has been a significant emphasis on Digital Libraries research in the last 20 years. There are several journals dedicated to discussion of Digital Libraries issues (e.g. International Journal on Digital Libraries, D-Lib Magazine) and topic specific conferences (e.g. The International conferences on Digital Libraries (ICDL), the European conferences on digital libraries (ECDL)). According to [6], from 1997 to 2006 34,5 per cent of studies present in mainstream peer-reviewed library and information science journals addressed aspects of use and usability. Therefore, the user has been set as the central point of the research field. This happened specially from 2004, where Digital Library systems were mature enough and researchers focused on examining their usability.

To support the creation of a real knowledge environment, information should be semantically annotated. In this direction, semantic technologies can reduce manual overhead for indexing documents, maintaining metadata, or caching for future search [4]. Semantic technologies are tools and methodologies developed around the Semantic Web, and they can be useful in a variety of situations, such as knowledge extraction, entity linking, word sense disambiguation, etc. The concept of Semantic Digital Libraries appeared to define the absorption of semantic principles into Digital Libraries. According to [7], Digital Library Systems suffer from two main issues: inadequate high-level cognition support, and inadequate knowledge sharing facilities. The authors proposed to extend the traditional indexing and searching schema of Digital Libraries from keyword-based to knowledge-based. To this end a two-layered architecture is proposed in order to address users' high order cognitive requests. In this architecture, the information space is subdivided into two different subspaces: a document subspace and a knowledge subspace.

In [2], a traditional knowledge organization system, such as a thesauri, is integrated with digital library architectures using emergent semantic technologies and semantic data. They claim that it is crucial to recognize the metadata schemas, data exchange models, and content standards of the Semantic Web for the creation and interoperation of Semantic Digital Libraries.

In recent years, there have been some research efforts to bring a knowledge-based approach to Digital Music Libraries. It is worth mentioning [1], where a method helps the musicologist to create a linked and extensible knowledge structure over a collection of Early Music metadata and facsimile images. In such a process, the system assists the musicologist in the creation of links between the collection and external Linked Data. Connections are published also as Linked Data, providing the possibility of being reused. Another enlightened study by Gracy et al. [8] reviews current efforts to connect music data already available with the Semantic Web. The authors collected, analyzed and mapped properties used by music Linked Data knowledge bases, library catalogs and various digital collections. They also noticed that dealing with Linked Data, library standards are often bypassed in favor of open structures based on Semantic Web standards, and in the case of music, the Music Ontology has been used as the basis for many of the available data sets. The authors proposed to extend the Music Ontology to include activities such as tagging and annotation, to facilitate the interconnection with Digital Libraries. A similar work [9] tries to define a mapping between MARC fields, subfields, and relator codes to the classes and properties of the Music Ontology. Two recent works [10, 11] relates the concept of Big Data with Music Digital Libraries. Weyde et al. [10] argue that the use of large data collections will enable better understanding of music. The authors claim that new research methodologies should be developed to create and adapt the technology necessary to achieve this goal. In [11] seven big datasets of musical-biographical metadata are aligned, and they show how analysis and visualization of the data might transform musicological understanding. Another effort in this direction is being done by the CompMusic project [2], creating new data collections of non-western music traditions from a knowledge-based perspective [12]. However, to the best of our knowledge, scant research has took advantage of semantic technologies to automatize the process of semantic annotation in Digital Musical Document Libraries, nor in the use of knowledge graphs to represent epistemic relations present in the library. In [13] some initial guidelines are drawn in order to apply these techniques to the music domain.

Once knowledge is well annotated and structured, a second challenge is how to exploit it to improve user experience with a Digital Library. Information visualization can play an important role in the reconceptualization of Digital Libraries as interactive knowledge environments [14]. Initiatives such as the INVENT framework [14] emphasizes the importance of rich interaction with visual representations for supporting knowledge environments.

### 2.2. Information Extraction

Information Extraction (IE) is a subfield of computer science focused on the automatic extraction of structured information from unstructured text documents in machine-readable format. This field is related with Natural Language Processing (NLP), whose major challenge is to understand natural language. IE is usually classified into Traditional Information Extraction and Open Information Extraction. In Traditional IE the vocabulary to be extracted is defined a priori by a template or an ontology, or it is learned from manually annotated training samples. In Open Information Extraction, there is no need to define a pre-specified vocabulary, the objective is to extract all possible relations without requiring any human input [15]. Following the second extraction paradigm, several methods have been published (e.g., [16, 17]). Open IE systems are able to extract arbitrary relations from sentences without the necessity of learning an extractor for each target relation from labeled training examples. Open IE systems have achieved notable success in massive open domain corpora [15].

Another crucial task in the process of IE is to identify which are the entities involved in the extracted relations, thus, determining the identity of entities present in a text. This task of NLP is called Entity Linking (EL), and it should not only identify the entity, but also link it to an existing knowledge-base (e.g. Wikipedia, DBpedia or Freebase). Several EL systems have been released with satisfactory performance, such as DBpedia Spotlight [18], Tag-me [19] or Babelfy [20]. The lack of a proper schema for the output of existing Open IE system limits their suitability to a specific domain, as there is too much ambiguity in extracted facts [21]. The combination of the Open IE paradigm with Entity Linking systems may help in this sense, by reducing ambiguity and, at the same time, improving interlinking with external Linked Data resources.

## 3. Grove Music Online

The Grove Dictionary of Music and Musicians [22] is an encyclopedic dictionary, and one of the largest reference works in Western music. It was first published in four volumes in the last quarter of the XIX century by George Grove. In 1980 a new version called The New Grove [23] was released with 20 volumes, where there are 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 Oxford Music Online as Grove Music Online [3].

### 3.1. Dataset

We manually crawled the text and title of 7,655 artist biographies from the Grove Music Online. These artist biographies were classified in the section People in history in the Grove Music Online. We extracted biographies from different periods of the history of music, from Pre-medieval time to contemporary biographies.

## 4. Semantic annotation

Once the text was extracted and stored locally, an Information Extraction process was performed. This process is divided in two different tasks. First, an Entity Linking process is performed to detect relevant entities in the text and link them to external resources from the LOD. Second, a relation extraction process is applied to

[2] http://compmusic.upf.edu

[3] http://www.oxfordmusiconline.com

detect semantic relations in text between detected entities. As an initial preprocessing step, the text of each biography was segmented (or tokenized) into different sentences. For this segmentation the Stanford NLP tokenizer implementation[4] was used. Then for each sentence both Information Extraction tasks were applied. Finally, all identified entities and extracted relations were added to a unique knowledge graph, which was shared by the whole corpus of biographies.

### *4.1. Entity Linking*

To detect named entities in our biography corpus we apply two different techniques. First, we use an existent tool for Entity Linking. From available tools we selected DBpedia Spotlight. Although DBpedia Spotlight does not obtain better precision and recall than other state-of-the-art tools, we selected it as it links entities directly with DBpedia and it is free available. DBpedia [5] is a knowledge-base with structured information extracted from Wikipedia. DBpedia Spotlight looks for named entities in the text that refer to resources present in DBpedia. The output of DBpedia Spotlight contains three relevant values: the URI, the types of the entity, and a score between 0 and 1. A URI (Uniform resource identifier) is a string of characters used to identify a resource unequivocally. DBpedia resources are classified into different classes of the DBpedia ontology and other schemas. This information is very useful to determine the type of entity detected. Only entities of the following types were taken into account by our system: Artist, Place, Agent, Person, Concept, Work, Language, Music Genre, Event and Award. The given score denotes an estimation of the correctness of the match. We also filtered results by score, keeping only entity matchings with a score higher than 0.5. Apart from DBpedia Spotlight, a co-reference resolution technique was applied taking into account the specificity of our dataset. By observing biographical texts, we detected that the vast majority of times a subject pronoun appeared, it was referring to the biographied. Therefore, each occurrence of the subject was treated as an entity, and linked to the correspondent DBpedia resource. To determine the gender of the subject an external Web service called GenRe API [6] was used. This service is able to determine the gender of a person name in any language with high accuracy.

### *4.2. Relation Extraction*

Once all entities were detected and annotated in the text, a process of relation extraction was applied. We used an ad-hoc rule-based method based on dependency parsing trees. The applied method looks for paths among entities in the syntactic structure of the sentence, and filter them out according to the linguistic category of the words in between. Relations are then lexicalized with the stemmed version of those words. When a relation between two entities is detected in a sentence, the entities are added to the knowledge graph as nodes, and the relation as an edge relating both nodes. The following example shows a relation extracted from the biography of Joseph Hofman:

> **Sentence**: *He* became director of the recently founded *Curtis Institute of Music*
> **Entity 1**: He (Joseph Hofman) `http://www.oxfordmusiconline.com/subscriber/article/grove/music/13172`
> **Entity 2**: Curtis Institute of Music `http://dbpedia.org/resource/Curtis_Institute_of_Music`
> **Relation label**: become director of

We identified in our dataset 9,736 entities that were involved at least in one relation. The number of entities found for each of the 12 different categories was: 5,718 Artists, 1,101 Places, 1,075 Agents, 814 Things, 590 People, 241 Concepts, 109 Works, 20 Languages, 35 Music Genres, 19 Events and 11 Awards. After applying our relation extraction method, we found 13,648 different relations between the 9,736 identified entities.

## 5. The knowledge subspace

### *5.1. Knowledge Graph*

Once entities were detected and relations defined we created our knowledge graph with the extracted information. Entities were stored as nodes and relations as edges. For each node we stored the name of the entry, the type of entity, the DBpedia URI, and the link to Grove Music Online (only if the entity was a biographied). The edges contained the lexicalization extracted from the relation extraction process, explaining the relation between the involved entities.

Once the process was complete we used the generated knowledge graph to create a knowledge subspace for the Digital Library. This knowledge subspace is complementary to the already available document subspace. To this end we created a new layer of Web navigation based on the graph using information visualization techniques.

### *5.2. Information Visualization*

To visualize and navigate through the knowledge graph we developed a Web interface using the D3js javascript library. In the interface the whole graph is first drawn as a network without any label. When the user hovers the mouse over a node, the name of the entity and their relations are highlighted. If the user clicks in a node a new graph is drawn with this node as its center, and all related nodes around it, as showed in Fig 1. Node and edge labels are also shown in the figure. If the user clicks on a related node the focus change to that node as the new center of the subnetwork. At the same time an info box is showed at the top right of the interface displaying information about the entity, such as names of related entities, a link to the DBpedia resource, and a link to Grove Music Online. Therefore, the user can switch in any moment to the traditional navigation of the document subspace. An online demo of the visualization tool is available online[7].

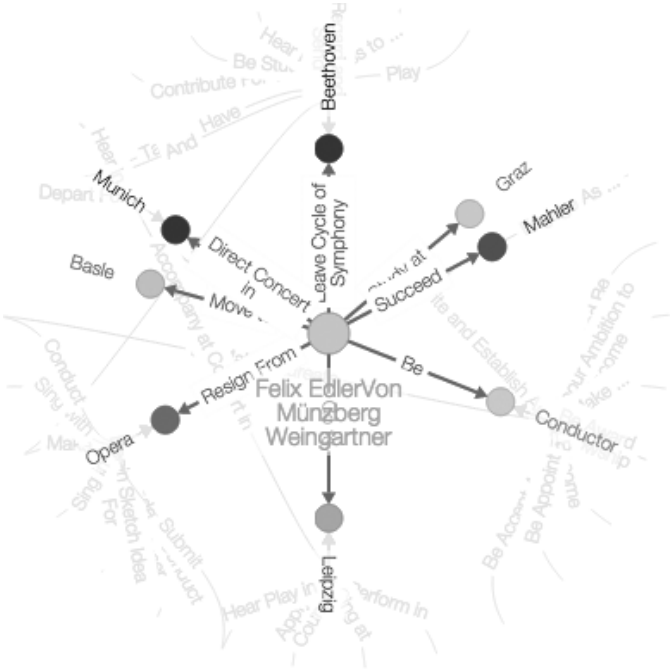


**Figure 1:** Graph Visualization Example

[4] `http://nlp.stanford.edu/software/`
[5] `http://dbpedia.org/`
[6] `http://namesorts.com/api/`
[7] `http://mtg.upf.edu/grove-browsing`

## 6. Conclusion

Tendencies in the development of Digital Libraries are moving from searchable environments to knowledge environments. Descriptive metadata and markup annotations give some structure to libraries, but complex structures and relationships remain implicit, latent or hidden. In order to discover and publish this latent knowledge, semantic technologies and standard data representations from the Semantic Web are being taken. This fact has given rise to a new generation of Digital Libraries called Semantic Digital Libraries. The use of semantic technologies may help to accelerate the process of semantic annotation of documents, and the publication of data in RDF format through the LOD cloud may enhance the possibilities for knowledge interchange between different libraries.

Therefore, we propose a method that takes advantage of semantic technologies in order to automatize a process of semantic annotation of a Music Digital Document Library. To test our approach, a process of Information Extraction has been applied to a subset of artist biographies from the Grove Music Online Digital Library. With the extracted information a knowledge graph was created, representing connections and relations between entities. Finally, this graph was graphically visualized in a Web interface. As a consequence of that, the library can be navigated through the knowledge graph.

The benefits of this approach are twofold. First, the creation of a knowledge subspace improve the way knowledge may be discovered by musicologists. Second, the use of information visualization over the knowledge subspace improves the user experience in the navigation through the Digital Library.

Although those initial results are promising, further development is already necessary to improve the Information Extraction process and the Information Visualization tool. Moreover, Information Extraction tools are usually developed and trained for domains different than music. Therefore, if extraction models had been trained over texts from the music domain, the performance would have been better. More exhaustive evaluation is also necessary to assess the quality of the information extracted and the satisfaction of final users.

## 7. Acknowledgments

This research was funded by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

## References

[1] T. Crawford, B. Fields, D. Lewis, and K. Page: *Explorations in Linked Data practice for early music corpora*. In *Digital Libraries 2014*, 2014.

[2] P. Babu: *Knowledge Organization Systems for Semantic Digital Libraries*. In *International Conference on Trends in Knowledge and Information Dynamics*, volume II:10–13, 2012.

[3] K. V. Fast and K. Sedig: *Interaction and the epistemic potential of digital libraries*. In *International Journal on Digital Libraries*, volume 11(3):169–207, 2011.

[4] S. Tönnies and W.-t. Balke: *Using Semantic Technologies in Digital Libraries - A Roadmap to Quality Evaluation*. In *13th European Conference on Digital Libraries (ECDL 2009)*, 2009.

[5] I. Herman, I. C. Society, G. Melanc, and M. S. Marshall: *Graph Visualization and Navigation in Information Visualization : A Survey*. In *Journal IEEE Transactions on Visualization and Computer Graphics*, volume 6(1):24–43, 2000.

[6] C. L. Liew: *Digital library research 1997-2007: Organisational and people issues*. In *Journal of Documentation*, volume 65(2):245–266, 2009.

[7] L. Feng, M. a. Jeusfeld, and J. Hoppenbrouwers: *Beyond information searching and browsing: acquiring knowledge from digital libraries*. In *Information Processing & Management*, volume 41(1):97–120, 2005.

[8] K. F. Gracy, M. L. Zeng, and L. Skirvin: *Exploring Methods To Improve Access to Music Resources by Aligning Library Data With Linked Data : A Report of Methodologies and Preliminary Findings*. In *Journal of the Association for Information Science and Technology (JASIST)*, volume 64(10):2078–2099, 2013.

[9] M. L. Zeng, K. F. Gracy, and L. Skirvin: *Navigating the Intersection of Library Bibliographic Data and Linked Music Information Sources: A Study of the Identification of Useful Metadata Elements for Interlinking*. In *Journal of Library Metadata*, volume 13(2-3):254–278, 2013.

[10] T. Weyde, S. Cottrell, D. Wolff, D. Tidhar, N. Gold, M. Plumbley, and M. Mahey: *Big Data for Musicology*. In *International Workshop on Digital Libraries for Musicology*. 2014.

[11] S. Rose and S. Tuppen: *Prospects for a Big Data History of Music*. In *International Workshop on Digital Libraries for Musicology*. 2014.

[12] X. Serra: *Exploiting Domain Knowledge in Music Information Research*. In *Stockholm Music Acoustics Conference 2013 and Sound and Music Computing Conference 2013*, 2013.

[13] S. Oramas: *Harvesting and Structuring Social Data in Music Information Retrieval*. In *Lecture Notes in Computer Science*, volume 8465:817–826, 2014.

[14] K. Fast: *The INVENT framework: Examining the role of information visualization in the reconceptualization of digital libraries*. In *Journal of Digital Information*, 2005.

[15] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni: *Open Information Extraction from the Web*. In *International Joint Conferences on Artificial Intelligence*, pages 2670–2676. 2007.

[16] A. Fader, S. Soderland, and O. Etzioni: *Identifying Relations for Open Information Extraction*. In *Empirical Methods in Natural Language Processing*. 2011.

[17] Mausam, M. Schmitz, R. Bart, S. Soderland, and O. Etzioni: *Open Language Learning for Information Extraction*. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012.

[18] P. Mendes and M. Jakob: *DBpedia spotlight: shedding light on the web of documents*. In *International Conference on Semantic Systems*, pages 1–8. 2011.

[19] P. Ferragina, D. Informatica, and U. Scaiella: *TAGME : On-the-fly Annotation of Short Text Fragments ( by Wikipedia Entities )*. In *19th ACM International Conference on Information and Knowledge Management*, pages 1625–1628, 2010.

[20] A. Moro, A. Raganato, R. Navigli, D. Informatica, and V. R. Elena: *Entity Linking meets Word Sense Disambiguation : a Unified Approach*. In *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[21] A. Dutta and M. Schuhmacher: *Entity Linking for Open Information Extraction*. In *Lecture Notes in Computer Science*, volume 8455:75–80, 2014.

[22] G. Grove: *A Dictionary of Music and Musicians*. 1878.

[23] S. Sadie: *The New Grove Dictionary of Music and Musicians, 2d ed.*, 2001.