DOing REusable MUSical data (DOREMUS)

Pasquale Lisena
EURECOM
Sophia Antipolis, France
pasquale.lisena@eurecom.fr

Raphaël Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

ABSTRACT

The aim of this tutorial is first to provide in-depth explanations of DOREMUS, a model for describing music metadata. We will demonstrate how real data coming from musical libraries can be converted to this model by presenting the whole DOREMUS tools chain. We will illustrate how the DOREMUS data can be used for query answering and consumed through various applications including an exploratory search engine and music recommender systems.

CCS CONCEPTS

• **Information systems → Ontologies**; **Recommender systems**; *Semantic web description languages*; *Music retrieval*;

KEYWORDS

Ontology, Music Metadata, Linked Data, Recommender System, Graph Embeddings

ACM Reference format:
Pasquale Lisena and Raphaël Troncy. 2017. DOing REusable MUSical data (DOREMUS). In *Proceedings of Knowledge Capture Conference, Austin, Texas USA, December 2017 (K-CAP'17),* 4 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Music information can be very complex. Describing a classical masterpiece in all its form (the composition, the score, the various publications, a performance, a recording, the derivative works, etc.) is a complex activity. An even more challenging task consists in describing jazz and ethnic music for which the performance plays a central role, the music is generally not written and the authorship is not well defined. In the context of the DOREMUS research project[1], we develop tools and methods to manage music catalogues on the web using semantic web technologies.

In this tutorial, we show strategies and tools for managing music knowledge. In the Section 2, we present the DOREMUS model for describing music, together with music specific controlled vocabularies. In the Section 3, we present tools for converting music datasets, taking as example the ones coming from the rich musical archives of three leading cultural institutions in France – the Bibliothèque Nationale de France (BnF), the Philharmonie de Paris (PP) and Radio France (RF) – describing musical works, publications, performances and concerts. We demonstrate the expressiveness of the model showing how complex music-specific queries can be answered. Finally, we describe strategies for data visualisation and recommendation in the Section 4.

[1] urlhttp://www.doremus.org

K-CAP'17, December 2017, Austin, Texas USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 A MUSIC DATA MODEL

Among the music ontologies, the most known example is the Music Ontology [9] that provides a set of music-specific classes and properties for describing musical works, performances and tracks, together with fragments of them. Further needs in exploit the music knowledge coming from libraries led to the definition of a new ontology.

2.1 The DOREMUS Ontology

The DOREMUS model[2] is an extension of FRBRoo, for describing cultural objects [4], applied to the specific domain of music. This is a dynamic model, in which the abstract intention of the author (called Work) exists only through an Event (i.e. the composition event) that realises it in a distinct series of choices called Expression. This Work-Expression-Event triplet can also describe different parts of the life of a work, like the Performance, the Publication or the creation of a derivative Work, each one incorporating the expression from which it comes from.

On top of the FRBRoo original classes and properties, specific ones have been added in order to describe aspects of a work that are specifically related to music, such as the musical key, the genre, the tempo, the medium of performance (MoP), etc. [3].

Each triplet contains an information that, at the same time, can live autonomously and be linked to the other entities. Thinking about a classic work, we will have a triplet for the composition, one for any performance event, one for every manifestation (i.e. the score), etc., all connected in the graph. A jazz improvisation that consists in an extemporaneous creation of a new work, will have only the triplet for the Performance Work, Performance Expression and Performance Creation, in absence of the moment of composition and writing of the score that are almost mandatory for classical music and without the need to be attached to any other entity. It is considered a work *per se*. All the Work entities of each triplet are then connected to a Complex Work, a class that has the objective of collecting together all the representations — both the conceptual and sensory ones (manifestation) — of the same creative idea.

The result is a model that, if on one side is quite complex and hard to adopt, on the other hand has a very detailed expressiveness.

[2] http://data.doremus.org/ontology/

The graph depicted inFigure 1 shows a real example from our data: Beethoven's *Sonata for piano and cello n.1*[3].

2.2 Music Controlled Vocabularies

A large number of properties that are involved in the music description are supposed to contain values that are shared among different entities: different composition can have as genre "sonata", different performer can play a "bassoon", different authors can have as function "composer" or "lyricist". These labels can be expressed in multiple languages or in alternative forms (i.e. "sax" and "saxophone", or the French keys "Do majeur" and "Ut majeur"), making reconciliation hard. Our choice is to use controlled vocabularies for those common concepts. A controlled vocabulary is a thematic thesaurus of entities, each one being again identified with a URI. We are using SKOS [8] as representation model, that allows to specify for each concept the preferred and the alternative labels in multiple language, to define a hierarchy between the concepts (so that the "violin" is a narrower concept with respect to "string"), and to add comments and notes for describing the entity and help the annotation activity. Each concept becomes a common node in the musical graph that can connect a musical work to another, an author to a performer, etc.

Different kinds of vocabularies are required for describing music. Some of them are already available on the web: this is the case of MIMO[4] for describing musical instruments, or RAMEAU[5] for musical genres, ethnic groups, etc. Some others are not published in a suitable format for the Web of Data, or the version published is not as complete as other formats that are available to libraries or in online sources: this happens with the vocabularies published by the International Association of Music Libraries (IAML), [6] that have been published after the start of the project and for which we sometimes provide more details (labels, languages, etc.). Finally, there is also the case of vocabularies that do not exist at all and that we generate on the base of real data coming from the partners, enriched by an editorial process that involved also librarians. As a result, we collected, implemented and published 15 controlled vocabularies belonging to 6 different categories[7].

3 DATA CONVERSION

Both the French National Library (BnF) and Philharmonie of Paris make use of the MARC format for representing the music metadata. The flat structure of MARC, which consists in a succession of fields and subfields (Figure 2), reflects the purpose of converting printed or handwritten records in a computer form. Although MARC is a standard, its adoption is restricted to the library world, making its serialization to other formats (usually XML) a need for an actual use. MARC fields are also not labeled explicitly, but encoded with numbers, with the consequence of having to use a manual for deciphering the content. The semantics of these fields and subfields is not trivial: a subfield can change its meaning depending on the field, under which it is found, and on the particular variant of MARC (UNIMARC and INTERMARC). A field or subfield can contain information about different entities, like the first performance and the first publication combined in the same field of the notes, without a clear separation. Often, the information is represented in the form of a free text [10].

The benefits of moving from MARC to an RDF-based solution consist in the interoperability and the integration among libraries and with third party actors, with the possibility of realizing smart federated search [1, 2]. In order to achieve these goals, two tasks are necessary: data conversion and data linking.

3.1 From MARC to RDF

For the conversion task, we rely on MARC2RDF, [8] an open source prototype we developed for the automatic conversion of MARC bibliographic records to RDF using the DOREMUS ontology [6]. The conversion process relies on explicit expert-defined transfer rules (or mappings) that indicate where in the MARC file to look for what kind of information, providing the corresponding property path in the model as well as useful examples that illustrate each transfer rule, as shown in Figure 3. The role of these rules goes beyond being a simple documentation for the MARC records, embedding also information on some librarian practices in the formalisation of the content (format of dates, agreements on the syntax of textual fields, default values if the information is absent).

The converter is composed of different modules, that works in succession. First, a *file* parser reads the MARC file and makes the content accessible by field and subfield number. We implemented a converting module for both the INTERMARC and UNIMARC variants. Then, it builds the RDF graph reading the fields and assigning their content to the DOREMUS property suggested in the transfer rules.

Then the *free-text interpreter* extracts further information from the plain text fields, that includes editorial notes.This amounts to do a knowledge-aware parsing, since we search in the string exactly the information we want to instantiate from the model (i.e. the MoP from the casting notes, or the date and the publisher from the first publication note). The parsing is realized through empirically defined regular expression, that are going to be supported by Named Entity Recognition techniques as a future work. Finally, the *string2vocabulary* component performs an automatic mapping of string literals to URIs coming from controlled vocabularies. All variants for a concept label are considered in order to deal with potential differences in naming terms. As additional feature, this component is able to recognise and correct some noise that is present in the source MARC file: this is the case of musical keys declared as genre, or fields for the opus number that contain actually a catalog number and vice-versa. These cases and other typos and mistakes have been identified thanks to the conversion process and the visualization of the converted data, supporting the source institution in they work of updating and correcting constantly their data.

3.2 Dealing With Heterogeneous Formats

Apart from MARC, we are converting other source bases (in XML), that are too specific to be handled by a single converter. Therefore, we developed *ad hoc* software that have a generic workflow:

[3] http://data.doremus.org/expression/614925f2-1da7-39c1-8fb7-4866b1d39fc7
[4] http://www.mimo-db.eu/
[5] http://rameau.bnf.fr/
[6] http://iflastandards.info/ns/unimarc/
[7] https://github.com/DOREMUS-ANR/knowledge-base/tree/master/vocabularies
[8] https://github.com/DOREMUS-ANR/marc2rdf

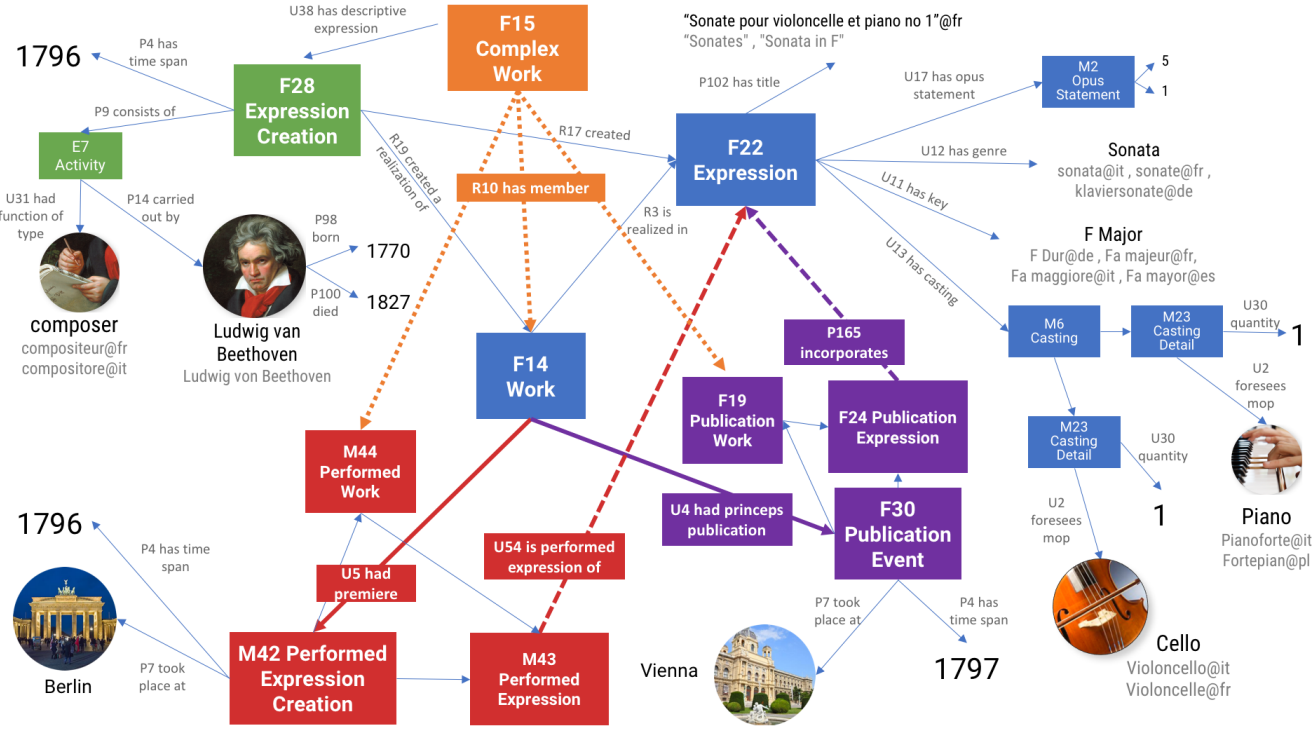

Figure 1: Beethoven's *Sonata for piano and cello n.1* represented as a graph using the DOREMUS ontology

```
001 FRBNF139081882FR
100 $313891295$w.0..b.....$aBeethoven$mLudwig van$d1770-1827
144 $w....b.fre.$aSonates$bPiano$pOp. 27, no 2$tDo dièse mineur
                                  NUM  SUB
```

Figure 2: An excerpt of a UNIMARC record.

UNIT OF INFORMATION	F22 Expression: Opus Number
PATH	F22 Self-Contained Expression U17 has opus statement M2 Opus Statement [U42 has opus number M12 Opus Number] + [U43 has opus subnumber M13 Opus Subnumber]
INTERMARC BNF	TUM : 144 $p, chain of digits TUM : 144 $p, chain of digits before the comma
TRANSFER RULE	Remove the abbreviation "Op." before the number
EXAMPLE	144 $pOp. 352 --> M12 = 352 144 $pOp. 27, no 2 --> M12 = 27, M13 =2

Figure 3: Example of mapping rules describing the opus number and sub-number of a work

parse the input file and collect the required information, create the graph structure in RDF, run the *string2vocabulary* module described previously. This procedure creates different graphs, one for each source. Those source databases are complementary but also contain overlaps (e.g. two databases that describe the same work or the same performance with complementary metadata). We have started to automatically interlink the datasets, so that the resulting knowledge graph provides a richer description of each work.

3.3 Answering complex queries

Before the beginning of the project, a list of questions have been collected from experts of the partner institutions[9]. These questions reflect real needs of the institutions and reveal problems that they face daily in the task of selecting information from the database (e.g. concert organisation or broadcast programming) or for supporting librarian and musicologist studies. They can be related to practical use cases (the search of all the scores that suit a particular formation), to musicologist topics (the music of a certain region in a particular historical period), to interesting stats (the works usually performed or published together), or to curious connections between works, performances or artists. Most of the questions are very specific and complex, so that it is very hard to find their answer by simply querying the search engines currently available on the web. We have grouped these questions in categories, according to the DOREMUS classes involved in the question.

Table 1 provides an overview of how many queries we can currently write for each category. The implementation of recordings, scores, performance that is still work in progress – along with the interconnection to the LOD repositories – is one important reason for which some questions have not yet been translated into SPARQL and other ones have not results.

[9]https://github.com/DOREMUS-ANR/knowledge-base/tree/master/query-examples

Category	Query / Questions
A. Works	23 / 29
B. Artists	1 / 3
C. Performances	6 / 9
D. Recordings	0 / 11
E. Publications	0 / 5

Table 1: For each category of questions, we provide the ratio of the number of converted queries

4 EXPLORATION AND RECOMMENDATION

We consider exploration and recommendation as two sides of the same medal. With the first one, we let the user browse the datasets, discover connections on his own, understand how we build the knowledge. Through recommendation, we remove this responsibility to the user with the purpose of presenting what he needs in a particular moment.

4.1 Visualizing the Complexity

We developed the first version of Overture, a web prototype of an exploratory search engine for DOREMUS data. The application makes requests directly to our SPARQL endpoint[10] and provides the information in a nice user interface.

At the top of the user interface, the navigation bar allows the user to navigate between the main concepts of the DOREMUS model: expression, performance, score, recording, artist. The challenge is in giving to the final user a complete vision on the data of each class and letting him/her understand how they are connected to each other. We keep as example Beethoven's *Sonata for piano and cello n.1* [11]. Aside from the different versions of the title, the composer and a textual description, the page provides details on the information we have about the work, like the musical key, the genres, the intended MoP, the opus number. When these values come from a controlled vocabulary, a link is presented in order to search for expressions that share the same value (for example, the same genre or the same musical key). A timeline shows the most important events related to the work (the composition, the premiere, the first publication). Other performances and publications can be represented below. The background is a portrait of the composer that comes from DBpedia. It is retrieved thanks to the presence in the DOREMUS database of `owl:sameAs` links. These links comes in part from the International Standard Name Identifier (ISNI) service[12], in part thanks to an interlinking realised by matching the artist name, birth and death date in the different datasets.

4.2 Music Recommendation Using Graph Embeddings

What should we suggest to a user listening Beethoven? Similar musicians should share with the German composer some features: the period, similar properties on the compositions (genre, key, casting) or similar instrument played (the piano itself, or also the harpsichord that is in the same family). But how to define a similarity measure that take into acount these concepts? We propose a solution based on graph embeddings generated at different levels:

(1) For simple features (e.g. genre, key, instrument), we compute for each term an embedding applying *node2vec* [5] on two sub-graphs: the one of the controlled vocabularies and the one corresponding to the usage of their values in the DOREMUS dataset;
(2) For complex features (e.g. artist), we generate the embeddings by the combination of its corresponding feature embedding. In the case of artists, we will generate a vector composed of the period (mapped in »0; 1¼) and the averages of the vector of the genre, key and casting (instrument) of his composition, together with the one of the played instrument, after having reduced their dimensionality;
(3) Finally, for the work, we combine again simple and complex feature embedding, following the same rules.

Using graph embeddings reduces the similarity problem as the reverse of an euclidean distance. if some properties are missing, we apply a penalisation computed as percentage of missing feature in the target vector with respect to the seed one [7].

The biggest advantage of this method is that the embeddings computation is required only for the simple features: each embedding is re-used in different combination. Because different weights can be assigned to each property in order to tune up the recommendation, we plan to experiment with neural networks in order to discover the best weighting strategy.

[10] http://data.doremus.org/sparql
[11] http://overture.doremus.org/expression/614925f2-1da7-39c1-8fb7-4866b1d39fc7
[12] The ISNI database contains authority information about people involved in creative processes (i.e. artists). It is managed by the ISNI Quality Team, which the BnF is a member of, and artists record in the BnF database contains generally an ISNI reference.

ACKNOWLEDGMENTS

This work has been partially supported by the French National Research Agency (ANR) within the DOREMUS Project, under grant number ANR-14-CE24-0020.

REFERENCES

[1] Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler. 2012. Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World* 113, 11/12 (2012), 549–570.
[2] Gillian Byrne and Lisa Goddard. 2010. The strongest link: Libraries and linked data. *D-Lib magazine* 16, 11 (2010), 5.
[3] Pierre Choffé and Françoise Leresche. 2016. DOREMUS: Connecting Sources, Enriching Catalogues and User Experience. In *24th IFLA World Library and Information Congress*. Colombus, USA.
[4] Martin Doerr, Chryssoula Bekiari, and Patrick LeBoeuf. 2008. FRBRoo: a conceptual model for performing arts. In *CIDOC Annual Conference*. Athens, Greece, 6–18.
[5] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA.
[6] Pasquale Lisena, Manel Achichi, Eva Fernandez, Konstantin Todorov, and Raphaël Troncy. 2016. Exploring Linked Classical Music Catalogs with OVERTURE. In *15th International Semantic Web Conference (ISWC)*. Kobe, Japan.
[7] Pasquale Lisena and Raphaël Troncy. 2017. Combining Music Specific Embeddings for Computing Artist Similarity. In *18th International Conference on Music Information Retrieval (ISMIR), Late-Breaking Demo Track*. Suzhou, China.
[8] Alistair Miles and José R Pérez-Agüera. 2007. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly* 43, 3-4 (2007), 69–83.
[9] Yves Raimond, Samer A. Abdallah, Mark B. Sandler, and Frederick Giasson. 2007. The Music Ontology. In *15th International Conference on Music Information Retrieval (ISMIR)*. 417–422.
[10] Roy Tennant. 2002. MARC must die. *Library Journal* 127, 17 (2002), 26–27.