DOREMUS: Doing Reusable Musical Data

Manel Achichi[1], Rodolphe Bailly[2], Cécile Cecconi[2], Marie Destandau[2],
Konstantin Todorov[1], Raphaël Troncy[3]

[1]University of Montpellier, [2]Philharmonie de Paris, [3]Eurecom

Abstract. This paper introduces DOREMUS—a semantic web project aiming to provide common knowledge models and shared multilingual vocabularies to cultural institutions, publishers, distributors and users in the musical domain. The project develops methods to describe, publish, connect and contextualize music catalogs on the web of data. Our focus is on the description of classical and traditional musical works as well as their interpretations (events).

1 Introduction

Music is everywhere—played, recorded, broadcasted and listened to. Files of recorded music are all over the web—stored, streamed, shared or sold. The knowledge about musical works is described in detail in the information systems of several cultural and media organizations around the world. The leading actors in France—the BnF (*Bibliothèque Nationale de France*), the *Philharmonie de Paris* and *Radio France*, as well as the *Meaning Engines* company join forces with three academic institutions within the DOREMUS project[1] in order to make this knowledge available and re-usable on the web of data. The project covers three main objectives: (1) to publish and to link musical data on the web of data including the description of musical works as well as their interpretation (event); (2) to develop tools assisting the selection of pieces of music, for instance, to generate an original playlist for a specialized radio station or to illustrate a musical period or genre; and (3) to construct and validate educational tools that enable the deployment of standards, vocabularies and the underlying semantic web technologies in the cultural institutions in France and worldwide.

In this paper, we describe our approach for publishing and interlinking descriptions of musical data based on ontological models and vocabularies and we discuss several applications. The goal is to exploit as much as possible the rich data which is available and to open it to the public. The main contributions consist in (1) the provision of advanced ontological models for describing musical works and events and (2) the generation of automatic methods for opening musical data that are initially given in a specific input format, so far, disconnected from the semantic web world.

---

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

2 Modeling Musical Data

Musical works are complex objects. Expressing them comprehensively requires the description of their physical manifestations (recordings, scores) and all the events that define them (creation, publication, performance). The first aspect is relatively well-mastered, in library catalogues as well as in the music industry. Several models can be used to describe a musical work, some being specific to the music domain (MusicOntology), while others being more broadly designed for libraries in general (FRBR[2], UNIMARC[3]). The second aspect is rather new, although there is a growing need and interest in using even-based models. Several ontologies aim specifically to define events (Event[4], LODE[5]), but there are fewer examples of using them to describe the creation and publication processes.

Fig. 1. Modeling a musical work: M-classes and U-properties are extensions of FRBRoo

One of the difficulties with musical works is that although their expressions may differ significantly, they are still regarded as a single work. Modeling it requires to express the singleness of the work as well as the specificities of its expressions, and to show how events are connected to these expressions. Another issue is that an arrangement may be considered as an expression or as a new work, depending on the data producer. Our model shall also make it possible to specify the relations between works, for instance when one work is derived from another.

We are currently working on a model enabling the expression of all these aspects in a coherent way, while being flexible and powerful enough to interoperate

with information systems dealing with any kind of cultural data. It occurred to us that the FRBRoo ontology had all the required qualities. It is based on the FRBR and CIDOC-CRM models, with a concern for the necessity to describe how events occur through the life of a thing and its resulting states. It is a formal ontology, based on the articulation of events and things, containing the notion of work, and offering precise elements to describe physical items. Its structure is adaptable and allows to model cultural metadata with fine granularity. Since the work, as defined in this model, is not specifically musical, we are extending the model by additional classes and properties that are recurrent and important in the description of musical works (Figure 1).

3 Data Lifting

As in many other domains, the DOREMUS project faces the problem of processing a very large amount of heterogeneous (in terms of formats and languages) and semi-structured data in order to transform them into RDF, describe them by (re-)using appropriate vocabularies, link these data to existing datasets and publish them on the web of data—a process also coined as *data lifting* [1]. We have followed and adapted the workflow of the Datalift platform (Figure 2). The methods for extracting RDF triples need to be adapted to the data source in the musical field, which consists of documents in the MARC (MAchine-Readable Cataloging) format, as well as XML documents.
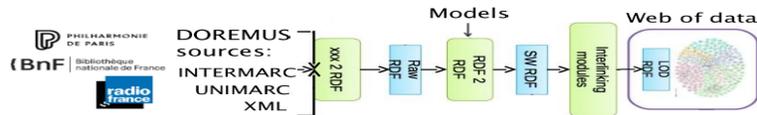

Fig. 2. Data semantization and publishing workflow (adapted from [1]).

There exist several approaches for extracting RDF triples from XML documents [2]. However, MARC (and its variants) documents has, so far, been rarely processed by semantic web applications. We have developed a two steps approach for transforming MARC data into RDF, while dealing with two variants of the MARC format: UNIMARC and INTERMARC. First, we propose an approach to extract raw RDF at the highest possible level of granularity by exploiting subareas coded in MARC with an alphanumeric character and a delimiter "$" (Figure 3). The principle is that each area and sub-area will be represented by an RDF property and the RDF object has the value of the area or one of its subfields. Second, we perform a manual mapping between the UNI/INTER-MARC areas and sub-areas with classes and properties defined in the models discussed in Figure 1. Thus, the description of musical works is performed at a lower level of granularity (due to merging of certain subfields taken as values of a property), while ensuring completeness of the information.
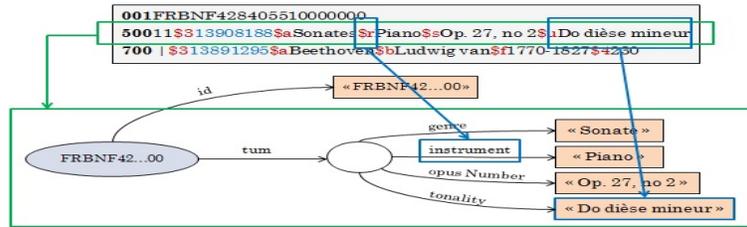
Fig. 3. Transforming the UNIMARC version of the uniform musical title (TUM) of the *"Moonlight Sonata"* by *L.V.Beethoven* to RDF

Several documents within a single institution or across institutions describe the same entity (a particular musical work), but from different perspectives (scores, sound recordings, videos). We aim to develop generic data connectors [1] to automatically detect relationships (equivalence, transcription, interpretation) between musical entities. The multilingual nature of data—a characteristics that should be preserved in the data lifting process—will be handled by automating the alignment process of multilingual ontologies specific to the field of music.

4   Applications and Future Work

Once the musical data coming from the cultural institutions is on the web of data, we plan to develop a musical recommendation application. It will serve as an evaluation of our model, and enable us to control the quality of data, after being modeled, lifted and interlinked. It will also be an example of how to implement the DOREMUS model for other cultural institutions.

The application will offer features such as the creation of original playlists for specialized radio, or the selection of appropriate musical works and interpretations to illustrate the biography of a composer, a historical period, culture or genre. The targeted end-users range from a child building his/her musical culture, through music lovers or curious discoverers, to expert musicologists.

Acknowledgments

References

1. F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, et al. Enabling linked data publication with the Datalift platform. In *AAAI Semantic Cities WS*, 2012.
2. D. Van Deursen, C. Poppe, G. Martens, E. Mannens, and R. Walle. XML to RDF conversion: a generic approach. In *IEEE AXMEDIS*, pages 138–144, 2008.