**DOREMUS: Doing Reusable Musical Data**

Manel Achichi1 , Rodolphe Bailly2 , C´ecile Cecconi2 , Marie Destandau2 , Konstantin Todorov1 , Rapha¨el Troncy3

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 publish, share, connect, contextualize and update music catalogues on the web of data.

**1 Introducing and Motivating DOREMUS**

Music is everywhere—played, recorded, broadcast 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 institutions around the world. The leading actors in this field in France—the *BnF* (*Bibliothèque Nationale de France*), the *Philharmonie de Paris* and *Radio France*, as well as the company *Meaning Engines*, together with three scientific labs gather forces under the DOREMUS project1 in order to make this knowledge available and 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, (2) to develop tools assisting the selection of pieces of music, for instance in order to propose an original musical program to a specialized radio station or choose pieces of music in order 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 to publishing and linking semantic musical data based on ontological models and vocabularies, which allow for the representation of musical works and events and we discuss the envisioned applications. The goal is to exploit all available data in the best possible way, valorizing these data and opening them to the public. The main contributions at this stage of the project 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](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 nowadays, in library catalogues as well as in the industry. Several models can be used to describe it. Some are specific to music (MusicOntology), others are more generalist (FRBR2 , UNIMARC3 ). The second aspect is rather new, although there is a growing need and interest in it. Several ontologies specifically define events (Event4 , DOLCE5 , etc.), but there are few examples of them being used to describe processes of creation or publication. One of the difficulties with musical works is that although their expressions may differ significantly from one to another, 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 a 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.



**Fig. 1.** Unambiguous minimal description of a musical work. The M-classes represent the extension of the existing model.

The data modeling working group of Doremus is 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*,

2 <http://www.ifla.org/publications/functional-requirements-for-bibliographic-records>

3 <http://www.ifla.org/publications/unimarc-formats-and-related-documentation>

4 <http://motools.sourceforge.net/event/event.html>

5 <http://www.loa.istc.cnr.it/old/DOLCE.html>

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 (as shown as the fragment in Fig. 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 coined as *data lifting* [1]. We have followed and adapted the workflow of the Datalift project, given in Fig.

2. The methods for extraction of RDF triples need to be adapted to the source data in the musical field, which consist of documents in the MARC (MAchine- Readable Cataloging) format, as well as XML documents. Taking into account multilingualism and data duplication are two of challenges that we face.



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

There exist several approaches for extracting RDF triples from XML documents (e.g., [2]). However, the content described by using the MARC format and its variants, which is the vast majority of the DOREMUS data, has so far remained unknown for the semantic web and therefore unused and unaccessible for semantic web applications. We have developped an approach for transforming MARC data into RDF, allowing to open data in that format to the web and build services on top of them. Recall that the DOREMUS project focuses on datasets describing classical or traditional music and the main two entities of interest are those of a *musical work* and an associated *event*.

We are dealing with two variants of the MARC format — UNIMARC and INTERMARC. Two solutions emerge when considering how to proceed in order to extract the semantic content from data delivered in these formats. At 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 ”$” (*cf.* Fig. 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.



**Fig. 3.** Transforming the UNIMARC version of the uniform musical title (TUM) of the *”Moonlight Sonata”* by *L.V.Beethoven* to RDF.

Further on, this approach was adapted to the set of models discussed above (Fig. 1), by the help of explicit mappings between the model classes and the corresponding UNI/INTER-MARC areas and sub-areas. 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.

Once the semantic content is extracted from the data of our partners, the question of managing data heterogeneity in the data linking and publishing processes arises. Several documents within a single institution or across institutions describe the same entity (a given piece of music), but from different perspectives (scores, sound recordings, videos). Generic data connectors [1] will be extended 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**

Once the musical data coming from the partner institutions is on the web of data, we plan to develop several applications. The main objective is to build tools in support of the selection of musical works, able to suggest original musical programming for specialized radia, choosing works and interpretations to illustrate the biography of a composer, a historical period, culture or genre. Groups targeted by the partner institutions vary from a child building his/her musical culture, through music lovers or curious discoverers, to expert musicologists.

**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 *AXMEDIS’08*, pp. 138–144, IEEE, 2008.