DOREMUS: A Graph of Linked Musical Works

Manel Achichi,[1] Pasquale Lisena,[2] Konstantin Todorov,[1] Raphaël Troncy,[2]
Jean Delahousse[3]

[1] LIRMM, University of Montpellier, CNRS, France
{achichi, todorov}@lirmm.fr
[2] EURECOM, Sophia Antipolis, France
{pasquale.lisena, raphael.troncy}@eurecom.fr
[3] OUROUK, Paris, France
delahousse.jean@gmail.com

Abstract. Three major French cultural institutions—the French National Library (BnF), Radio France and the Philharmonie de Paris—have come together in order to develop shared methods to describe semantically their catalogs of music works and events. This process comprises the construction of knowledge graphs representing the data contained in these catalogs following a novel agreed upon ontology that extends CIDOC-CRM and FRBRoo, the linking of these graphs and their open publication on the web. A number of specialized tools that allow for the reproduction of this process are developed, as well as web applications for easy access and navigation through the data. The paper presents one of the main outcomes of this project—the DOREMUS knowledge graph, consisting of three linked datasets describing classical music works and their associated events (e.g., performances in concerts). This resource fills an important gap between library content description and music metadata. We present the DOREMUS pipeline for lifting and linking the data, the tools developed for these purposes, as well as a search application allowing to explore the data.

1 Introduction

The Linked Open Data (LOD) paradigm for data representation, sharing and publishing has been more and more appealing to the world of museums and libraries over the past years. The LOD project and the semantic web in general offer technological means for data reuse, increased visibility and data sharing on the web, data federation and facilitated exchange of metadata by the creation of links across resources. Attracted by these possibilities, many major actors from the library world, such as the Library of Congress (LOC) or the French National Library (BnF), have embraced semantic web technologies with the goal to open their archives and catalogs to the web. This process has resulted in a number of openly available and explorable RDF graphs reflecting the rich content of numerous libraries and cultural institutions from all over the world [1].

The DOREMUS project follows this line of research and practice, with a particular interest in classical and traditional music, so far relatively under-represented on the LOD.[4] Three major French cultural institutions—the BnF, Radio France (RF) and the Philharmonie de Paris (PP)—have joint efforts with data and social science academics in order to develop shared methods to describe semantically their catalogs of music works and events and open them to the web community. A major contribution of the project is the development of the DOREMUS ontology[5] which extends the well-known CIDOC-CRM and FRBRoo models for representing bibliographic information[6], adapting it to the domain of music, thus filling an important representational gap. A number of shared vocabularies about music-specific concepts (such as musical genres or keys) have been collected or developed, linked and published using the SKOS standard. The data from the catalogs of the three partner institutions comes in MARC or XML formats. Specific tools for data conversion to RDF following the DOREMUS model have been developed. This process results in the construction of several knowledge graphs about music works and events, which have been linked using a specifically developed for this purpose data linking tool. For evaluation purposes, a benchmark has been created manually by the library experts and shared to the semantic web community as part of the Ontology Alignment Evaluation Initiative (OAEI). The data fusion process results in the construction of a pivot graph of shared and unique musical works. Finally, an exploratory search engine is developed that allows to browse the knowledge graphs.

This paper covers the components of the DOREMUS workflow described above, which altogether form a paradigm for lifting, linking and publishing music library metadata. We present in detail the DOREMUS knowledge graphs with a focus on the (re-)used models and vocabularies and the processes that allow for their (re-)production and fusion. The contributions of this work are:

- A model for describing musical works and events extending FRBRoo together with a number of shared and linked music-specific controlled vocabularies.
- Three knowledge graphs about music works that represent the catalogs of three major French cultural institutions.
- An approach to interlink these graphs resulting in the construction of a pivot graph, containing all unique works and links to the original graphs.
- A set of benchmark datasets for data linking evaluation.
- A set of tools for data generation, vocabulary alignment and validation, data linking, pivot graph construction, and data search and exploration.

The remainder of the paper is structured as follows. In the next section, we provide general information about the graphs, their form and content, the (re-) used ontologies and controlled vocabularies, as well as statistics. In Section 3, we detail the different components of the DOREMUS data production pipeline, and in particular, the data conversion and linking approaches. In Section 4, we demonstrate how this resource has already been used and we discuss its wider expected impact. We present related initiatives in Section 5 before we conclude and discuss future work in Section 6.

[4] `http://www.doremus.org`

[5] `http://data.doremus.org/ontology/`

[6] `http://new.cidoc-crm.org/frbroo/`

2 The DOREMUS Datasets of Linked Musical Works

The DOREMUS knowledge graph consists of several datasets, each containing the information coming from a specific database of an institution. In that, a given real-world entity (e.g., a music work) is represented at most once in each graph. Currently, three stable datasets have been published: **(1)** `bnf`: Works and Artists, originally described in MARC records of the BnF; **(2)** `philharmonie`: Works and Concerts, originally described in MARC records of the PP; **(3)** `itema3`: Concerts and Recordings, originally described in XML records of RF. Each dataset has two access points: (1) A specific named graph in the DOREMUS triplestore, accessible through a public SPARQL endpoint. Each graph follows the pattern `http://data.doremus.org/<dataset_name>`. (2) A set of RDF files in Turtle format, available to public download. All datasets are licensed for free distribution, following a Creative Commons Attribution 4.0 license[7] and have a DCAT description in the triplestore itself. All links to DOREMUS datasets or tools are given in Table 1.

Name of resource or tool and description, URL
Data
`bnf`: Works and Artists from the BnF, `http://data.doremus.org/bnf`
`philharmonie`: Works and Concerts from the PP, `http://data.doremus.org/philharmonie`
`itema3`: Concerts and Recordings from RF, `http://data.doremus.org/itema3`
DOREMUS triplestore, `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/data`
DOREMUS sparql endpoint, `http://data.doremus.org/sparql`
Example queries, `http://data.doremus.org/queries.html`
DOREMUS ontology, `http://data.doremus.org/ontology`
DOREMUS vocabularies, `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/vocabularies`
Vocab. Alignments, `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/vocabularies/alignments`
DOREMUS linked data, `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/linked-data`
DOREMUS benchmarks 2016, `http://islab.di.unimi.it/content/im_oaei/2016/#doremus`
DOREMUS benchmarks 2017, `http://islab.di.unimi.it/content/im_oaei/2017/#doremus`
Tools
`marc2rdf` converter, `https://github.com/DOREMUS-ANR/marc2rdf`
`itema3` converter, `https://github.com/DOREMUS-ANR/itema3converter`
`euterpe` converter, `https://github.com/DOREMUS-ANR/euterpe-converter`
Legato: instance matcher, `https://github.com/DOREMUS-ANR/legato`
DOREMUS pivot graph constructor, `https://github.com/DOREMUS-ANR/pivot-graph-constructor`
Overture search engine, `http://overture.doremus.org`
YAM++ vocabulary mapping and validation, `http://yamplusplus.lirmm.fr`
Learning materials, `https://github.com/DOREMUS-ANR/training`

Table 1: Links to DOREMUS resources and tools.

2.1 Content and Form of the Resource

The DOREMUS knowledge graphs contain information about works (referred to as *expressions*) and related entities from the field of classical and traditional music. Each entity is identified by an univocal persistent URI, which follows the pattern `http://data.doremus.org/<group>/<uuid>`, where the group is determined by the class of the entity (e.g. `expression`) and the UUID (Universal Unique Identifier) is generated at conversion time in a deterministic way

[7] `https://creativecommons.org/licenses/by/4.0/`

using the dataset name, the class and the identifier of the source record as seed.[8] Currently, the resource is shaped by three knowledge graphs of music works (one per institution), that are linked together in a pivot graph, which is the union of all unique expressions across the three bases, together with `owl:sameAs` links to the original graphs (cf. Section 3). We find information pertaining to the instruments, genre and key of a music work (e.g., "piano", "sonata", "A-flat major"), its composer and title(s), date of creation, catalog numbers, opus numbers, etc. As an example, we can find all this information linked to `http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea`, representing Beethoven's *Moonlight Sonata*. Jazz music would contain a different kind of information, as in `http://data.doremus.org/expression/cc7fc9a6-124d-3cc1-\95e7-5644ecb394a6`, representing Coltrane's *Naima*.

2.2 (Re-)used Ontologies and Vocabularies

The DOREMUS model is an ontology for the description of music catalogs. It is an extension for the music domain of the FRBRoo model for describing librarian information, which has in turn been born as a dialog of the librarian FRBR model and the CIDOC-CRM ontology for representing museum information, putting togheter the distinction between Work, Expression, Manifestation, Item of the former with the centrality of creation events for describing the cultural object lifecycle coming from the latter [2]. On top of the FRBRoo original classes and properties, specific ones have been added in order to describe aspects of a work that are related to music, such as the musical key, the genre, the tempo, the medium of performance (MoP), etc. [3]. The model is ready to be used for describing the interconnection of different arts: it is the case of the soundtrack of a movie, or a song that uses the text of a poem.

DOREMUS imports the Work-Expression-Event triplet[9] pattern of FRBRoo: the abstract intention of the author (Work) exists only through an Event (i.e. the composition) that realizes it in a distinct series of choices called Expression(s). This pattern ensures that each step of the life of a musical work can be modeled separately, following the same triplet structure. Thinking about a classic work, we will have a triplet for the composition, one for any performance event, one for every manifestation (e.g., the score), all connected in the graph. This means also that each part of the music production process is considered as an Event that gives birth to a new Work and a new Expression: this leads to the creation of classes like Performance Work or Recording Expression. Each triplet contains information that at the same time can live autonomously and be linked to the other entities. This provides the freedom of representing, for example, a jazz improvisation as extemporaneous performance not connected to a particular pre-existing work, or to collect all the recordings of a piece of world music. The result is a model, which is quite complex and hard to adopt if we look at the levels of distribution of information: from an Expression, one has to pass through Event

[8] `https://github.com/DOREMUS-ANR/marc2rdf/blob/master/URI.patterns.md`

[9] Not to be confused with an RDF triple.

and Activity to reach a composer, or through Casting and Casting detail to get the MoP. On the other hand, the model has a very detailed expressiveness that allows, for instance, to describe different kinds of contributors (not only authors or performers), to detail the casting of a composition (with number, roles, notes for each instrument/voice), to specify performers at level of single performance inside a whole concert. As an extension of FRBRoo, the model appears familiar to librarian catalogers (documentation: `http://data.doremus.org/ontology`).

For the description of music-specific concepts like the key, the genre or the MoP, we publish controlled vocabularies (using SKOS and MODS standards), realized and enriched by an editorial process that involved also librarians, in order to overcome multilingualism and alternative names issues. Some of these vocabularies were already available and in use by the community: in this case our contribution consists in their collection, conversion to SKOS (if needed) and alignment. As a result, we collected, implemented and published 17 controlled vocabularies belonging to 7 different categories (musical keys, types of derivation, modes, thematic catalogs, functions, musical genres and MoP) (Table 1) [4]. The vocabularies are all available in the DOREMUS triplestore server via its public SPARQL endpoint. Alternatively, they can be explored by a web browser starting from `http://data.doremus.org/vocabularies/`. Each vocabulary is licensed for free distribution, following a Creative Commons Attribution 4.0 license.

The categories of genres and MoPs contain each 6 different vocabularies, including well-established reference thesauri, as well as institution-specific lists. The vocabularies of these two categories have been aligned by establishing `skos:exactMatch` relations between their elements in a pairwise manner using an automatic ontology and thesaurus matching system and these alignments have been manually validated and enriched by the experts of the three institutions. This process has been assisted by a dedicated generic web-application for ontology matching and mapping validation, YAM++ *online* [5], developed in part for the purposes of the project (link available in Table 1).

Statistics. Currently, the DOREMUS dataset includes more than 16 million triples, which describe over 3 million distinct entities. The classes and properties used come mostly from the DOREMUS ontology, FRBRoo and CIDOC-CRM, counting in total 57 distinct classes and 120 distinct properties. Table 2 summarizes the number of entities for the most representative classes and reports details about the presence of specific information.

3 Resource Development and Reconstruction

The general workflow of DOREMUS is depicted in Fig. 1. The data from the three partner institutions is first converted to RDF following the DOREMUS model, resulting in three independent knowledge graphs (one per institution), which are then linked. After a manual validation of a set of uncertain links, a pivot graph is built containing identifiers of the union of all works found in the three graphs, together with identity links to the resources in each of the three institutional graphs. We detail on these stages of the workflow in this section.

class	bnf	philharmonie	itema3	total
Expression	135818	9005	8319	153142
- with casting detail	123219	4621	0	127840
- with key	19645	1973	0	21618
- with genre	128497	3820	8071	140388
- with composer	133371	7741	8231	149343
- with composition date	91566	5712	4856	102134
- with catalogue	20796	2908	0	23704
- with order number	11598	1612	0	13210
- with opus number	21836	1985	0	23821
Performance	15784	784	1531	18099
- with more than 1 performed work	0	713	1277	1990
Track	0	6538	18273	24811

Table 2: Number of entities of given classes for each dataset.

3.1 Data Conversion

The data collected from the BnF and the PP describing music works is represented in the UNI– and INTERMARC variants of the MARC format. A MARC file is a succession of fields, each carrying a 3-digit label, and subfields, delimited by the $ symbol (e.g., "50011$313908188$qSonates$rPiano$sOp.27, no 2$uDo mineur").[10]. We have developed an open source prototype, named `marc2rdf` to automatically convert UNI- or INTERMARC bibliographic records to RDF, implementing the DOREMUS model (link to the tool given in Table 1). The conversion process relies on explicit expert-defined transfer rules (or mappings), which provide the corresponding property path in the model as well as useful examples [6]. We have used the DOREMUS properties to name the extracted relations (e.g., `mus:U12_has_genre` is the property describing the genre of a work). Beyond being a documentation for the MARC records, these rules embed information on specific and distinct librarian practices in the formalization of the content (format of dates, syntax of textual fields, default values for missing information), making `marc2rdf` a robust generic converter for MARC files.

The converter is composed of different modules that work in succession (Fig. 2). 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. 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

[10] For detailed information, we refer to the documentation released by The International Federation of Library Associations and Institutions (IFLA)`http://www.ifla.org/publications/ifla-series-on-bibliographic-control-36`

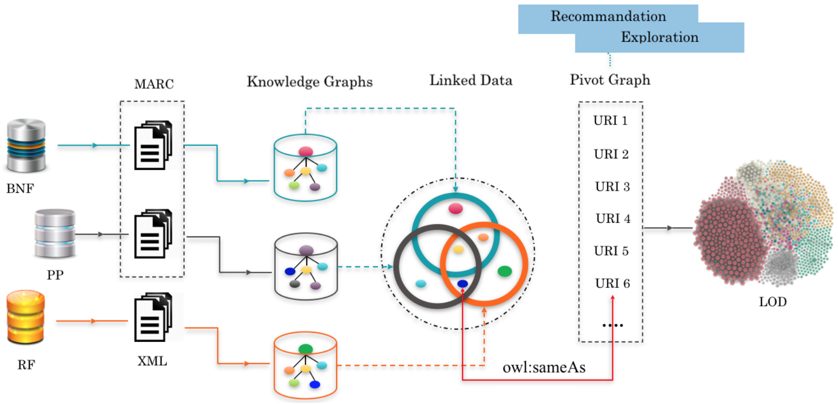


Fig. 1: The DOREMUS data lifecycle.

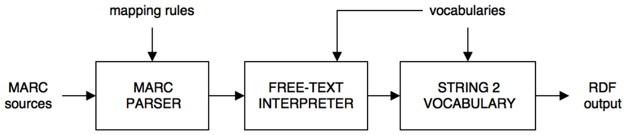


Fig. 2: The application flow of `marc2rdf`.

realized through empirically defined regular expressions, that are going to be supported by Named Entity Recognition techniques as 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 recognize and correct noise that is present in the MARC file: this is the case of musical keys declared as genre, or fields for the opus number that actually contain a catalog number and vice-versa.

The `marc2rdf` tool allows to reproduce deterministically the conversion process at any moment in time, providing the opportunity to seamlessly take into account possible updates of the ontology (e.g., the addition of a new property) and/or the data entries (a new record entering the catalog of one of the institutions), ensuring in that way the currentness and dynamics of the graphs.

The works from Radio France, described in XML, are managed by an *ad hoc* software that parses the input file, collects the required information, creates the RDF graph structure and runs the *string2vocabulary* module.

3.2 Data Linking

The three datasets that are currently subject to interlinking are highly heterogeneous: a given entity (e.g., a musical work) can be described quite differently across the three institutions. In addition to well-known data discrepancies such as lexical, semantic (polysemy, synonymy) and orthographic mismatches of string literals, the use of acronyms and abbreviations or differences in formats and types of numerical values, we have encountered several commonly occurring issues that are specific to our data. We outline some of them below.

— *Differences in coverage* and particularly lack of information in one of the graphs as compared to a richer description in another. In our case, the works coming from RF are systematically described by a considerably smaller set of attributes, than those found in the catalogs of the BnF and the PP (see Table 2).

— *Different depths in the graphs*, at which we find the value of interest—e.g., the birthplace of a composer can be directly assigned to the entity in one graph, or via a longer property chain in another.

— Presence of *comments in the form of free text* (given by the property `ecrm:P3_has_note`) that are difficult to compare, as well as presence of *institution-specific resource identifiers* (bibliographical records ID's) given under the same property name across different datasets, although not comparable.

— Presence of *blocks of highly similar in their descriptions, but yet distinct instances* in each of the graphs—e.g., the set of all piano sonatas by Beethoven, differing from one another in only one or two property values, which makes their disambiguation difficult and is likely to produce false positives.

In a first attempt to interconnect these graphs, we relied on state-of-the-art linking systems [7,8] that adopt a property-based philosophy where a set of attributes is selected in order to compare instances across two datasets based on (an aggregation of) similarity measures computed on their literal values. The results obtained proved to be not satisfactory.[11]. Consequently, we develop our own linking tool, named *Legato* [10]—a generic data linking system motivated by the DOREMUS use-case scenario and data linking challenges.

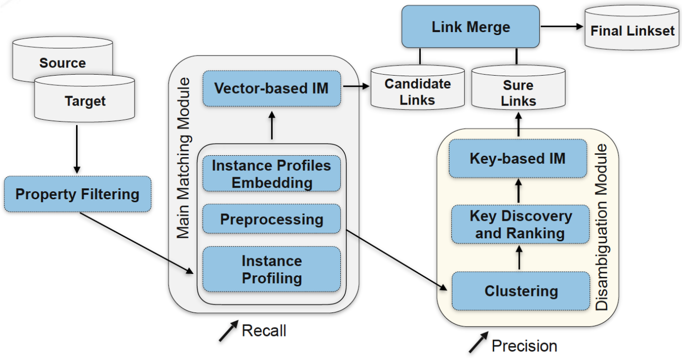


Fig. 3: Processing pipeline of *Legato*.

[11] Evaluation results of *Legato* on OAEI benchmarks can be found at `https://github.com/DOREMUS-ANR/legato/blob/master/Legato-Results.png`. Data and configuration files for SILK are available at `https://github.com/manoach/SILK-Evaluation`. Note that SILK is configured by using the best keys selected by the algorithm in [9].

Legato is designed to match entities from highly heterogeneous graphs, effectively disambiguating highly similar yet distinct resources. Fig. 3 shows the generic workflow of the system. The data cleaning module ensures to only keep properties that are comparable across the datasets (hence, comments in the form of free text, as well as institution-specific instance identifiers are removed). The instance profiling module represents instances by a subgraph corresponding to the union of the Concise Bounded Descriptions (CBD)[12] of each resource and its direct neighbors. In that, contrarily to SILK or Limes, Legato (in its default version) does not compare property values, but considers all extractable literal values as a bag-of-words. This representation addresses in its mechanism a number of data heterogeneities without requiring user input, in particular, the description differences and property depths discrepancies outlined above. The literals of these subgraphs are then used to project each instance in a vector space and the matching consists in comparing the resulting vectors. A deliberately low threshold is used for the vector similarity in order to ensure high recall. Then, highly similar instances are grouped together by the help of a standard hierarchical clustering algorithm [11]. An RDF key discovery [12] and a key ranking [9] algorithms are applied on each pair of similar clusters (identified by comparing cluster centroids) across the two graphs, in order to identify the set of properties that best allow to discriminate between the resources contained in each cluster. A new linkset (called "sure links") results from this process and is then compared to the links produced at the matching step (called "candidate links") in order to eliminate errors and increase precision, leading to the production of the final linkset. The outcome of *Legato* is presented in the EDOAL format,[13] allowing to keep track of the associated confidence scores, or as `owl:sameAs` triples. We provide an open source implementation of the system together with a simple user interface (see Table 1 for a link).

Given our knowledge of the DOREMUS data, we have customized the linking process for the purposes of the project in two respects. (1) The linking workflow begins by searching for values of the composer name and catalog number properties, because the set of these two properties has been identified as a key by our experts. If values for these two properties are found for a given pair of instances, they are directly used for the comparison and *Legato* is executed on the remaining instances only. Note, however, that these properties do not have values for a very large number of works and in particular, no entry of `itema3` has a catalog number (cf. Table 2). (2) In order to speed-up the execution of *Legato*, we have partitioned the datasets per composer and linked pairs of subsets across two graphs that gather works by the same composer.

To evaluate *Legato*, we have constructed benchmarks of music works from the BnF and the PP, by asking the librarian experts to manually select pairs of identical resources from the their respective catalogs. We have ensured that our benchmarks are representative and provide a fair account of the heterogeneity issues outlined above. This results in the generation of two benchmarks that

[12] `https://www.w3.org/Submission/CBD/`

[13] `http://alignapi.gforge.inria.fr/edoal.html`

have been released by the Instance Matching track of OAEI 2016 and 2017 (cf. Table 1). *Legato* has participated to the 2017 edition of the campaign, ranking first on the DOREMUS-FPT task. NjuLink surpasses *Legato* by 0.025 points (F1) on the DOREMUS-HT track, but performs worse by 0.044 points on the FPT track. Our data exhibits characteristics of the two, therefore we decided to go for *Legato*, in addition to the customizability argument given above.

3.3 Link Validation and Pivot Graph Construction

As a result of the pair-wise alignments of the three graphs, we end up with three sets of links. We exploit the topology of the connectivity of the entities of the three graphs in order to define subsets of links to provide to data experts for manual validation, aiming to ensure a final set of links of high quality. We identify four connectivity patterns, shown in Fig. 4, according to which we classify the produced links to three categories: *certain links*, *invalid links* and *validation candidates*. The classification of a link as "certain" depends both on its confidence value and on the connectivity pattern in which it falls. The certain links are retained and included in the pivot graph constructed from our data (see below). If a link is approved by an expert during the user validation process, its confidence value is set to 1, which automatically classifies it as "certain", else it is declared as "invalid". We consider the following link patterns.

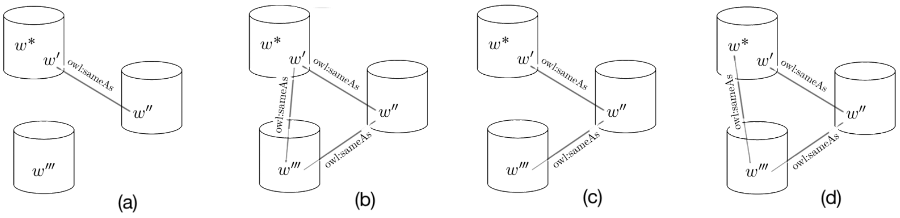


Fig. 4: Links across the three graphs: four connectivity patterns.

(1) Single link. This is the case when two works are connected via an identity link across their corresponding graphs as a result of the automatic linking process (Fig. 4(a)). According to the confidence value of the link, it is either classified as certain, passed over to the experts for validation, or discarded.

(2) Triangle. This is the case when three works from the three graphs are linked via three `owl:sameAs` relations. In this case, the three links are considered as certain and the expert is not solicited (Fig. 4(b)).

(3) Missing link. This is the case when an instance w' from one graph is linked to an instance w'' from the second graph, which in turn is linked to an instance w''' from the third graph, but no link has been created between w''' and w' (Fig. 4(c)). Instead of inferring that link, independently on the links confidence values, we pass the two link candidates $< w'$, `owl:sameAs`, $w'' >$,

$< w''$, `owl:sameAs` $w''' >$ to the experts for validation. If the validation process results in classifying these two links as certain, the link $< w'''$, `owl:sameAs`, $w' >$ is inferred and classified as certain. Note that the third link inference mechanism is activated only in case we have two certain links.

(4) Conflict. This is the case when an instance w' from one graph is linked to an instance w'' from the second graph, which in turn is linked to an instance w''' from the third graph, and w''' is linked to an instance $w*$ from the first graph, where $w' \neq w*$ (Fig. 4(d)). All three links are passed to the experts for validation. This necessarily leads to invalidating at least one of the three links, in which we fall into one of the three cases described above.

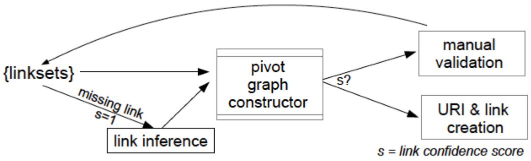


Fig. 5: Link validation and pivot graph construction workflow.

Pivot Graph Construction. We construct a referential pivot graph of music works that is the mathematical union of the three sets of works from the three partner institutions. As an editorial decision, a novel URI is created for every entity in that graph, following the URI creation pattern described previously, together with a `owl:sameAs` link to the URIs identifying this entity in each of the three input graphs (at least one such URI exists). For example, if a given expression is described in both the BnF and the PP graphs, the pivot graph will contain the following two triples: <PIVOT_URI> `owl:sameAs` <BNF_URI>, <PP_URI>. If the work exists in one single graph only (e.g., the one of BnF), one single triple will be declared: <PIVOT_URI> `owl:sameAs` <BNF_URI>. To reconstitute these links, we rely on the linksets produced in the data linking phase and on the manual link validation task. As explained above, as a result of these processes, we end up with three sets of "certain" links. Only links from that category will appear in the pivot graph. As Fig. 5 shows, the process of pivot graph construction and that of the manual validation of links are tangled up in a single workflow. The code of the algorithm for (re-)generating the pivot graph is released as open source (cf. Table 1).

Currently, the manual validation process is in progress. Therefore, the published pivot graph contains the links that have been identified automatically (i.e., corresponding to the patterns in Figs. 4(a) and 4(b)) by using the non-conservative thresholds of *Legato* tuned by the help of our benchmarks (0.2 for `bnf-philharmonie` and 0.5 for the other two pairs of datasets). The graph contains also the links of all unique works (those that have no matches found in any of the two other bases) to their original URIs. The results of the automatic link discovery process on the three bases together with the resulting pivot graph in its current shape are available at `https://github.com/DOREMUS-ANR/knowledge-base/tree/master/linked-data`.

Links statistics. We have currently a total of 7495 links created automatically across the three graphs, among which we have 2520 links of type *single link*, 396 links of type *triangle*, 3378 links of type *missing link* and 261 links of type *conflict* (as labeled in Fig. 4), plus additional 940 links of type 1:many that are currently subject to post-processing. Updates in the datasets will not decrease the number of links since the source databases monotonically grow.

4 The DOREMUS Resource In Use

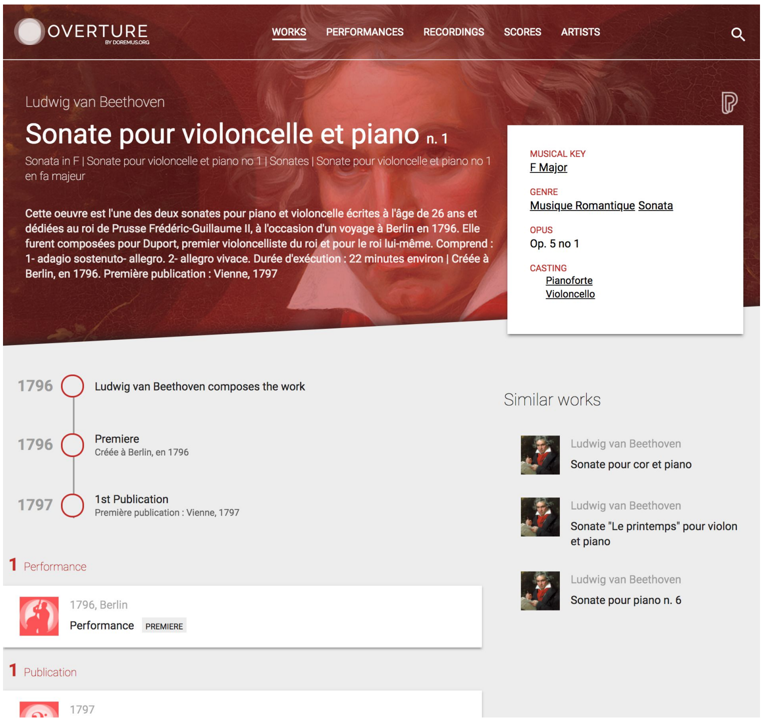
We proceed to discuss aspects related to the use of the resources, starting with their exploration and search.

`Overture`: *an Exploratory Search Engine.* We develop OVERTURE (Ontology-driVen Exploration and Recommendation of mUsical REcords), a prototype of an exploratory search engine for DOREMUS data, available at `http://overture.doremus.org`. The application makes requests directly to the SPARQL endpoint and provides information in a web user interface (UI).

At the top of the UI, the menu bar allows the user to navigate between the main concepts of the DOREMUS model: expression, performance, score, recording, artist. Fig. 6 represents Beethoven's *Sonata for piano and cello n.1* as seen in OVERTURE. 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 present 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 in the story of the work (the composition, the premiere, the first publication). Other performances and publications can be represented below.

The richness of the DOREMUS model offers to the end-user the chance to perform a detailed advanced search. All expressions (works) are searchable by facets, that include the title and the composer, but also keys, genres, detailed castings, making it possible to select very precise subsets of data, like all sonatas (genre) that involve a clarinet and a piano (MoPs). The hierarchical properties in the controlled vocabulary allow the smart retrieval not only of the entity that match exactly the chosen value (i.e. *Strings*), but also any of its narrower concepts (i.e. *violin*, *cello*, etc.).

A work-in-progress recommendation system is also implemented in `Overture` in order to suggest to the final user different works to discover. The recommended works have similar properties to the current one, like the genre, the composer and the foreseen instruments. The recommendation is realized by computing knowledge graph embeddings using *node2vec* [13] on the DOREMUS knowledge graph and selecting the closer works using the euclidean distance [14].



Ludwig van Beethoven

Sonate pour violoncelle et piano n. 1

Sonata in F | Sonate pour violoncelle et piano no 1 | Sonates | Sonate pour violoncelle et piano no 1 en fa majeur

Cette oeuvre est l'une des deux sonates pour piano et violoncelle écrites à l'âge de 26 ans et dédiées au roi de Prusse Frédéric-Guillaume II, à l'occasion d'un voyage à Berlin en 1796. Elle furent composées pour Duport, premier violoncelliste du roi et pour le roi lui-même. Comprend : 1- adagio sostenuto- allegro. 2- allegro vivace. Durée d'exécution : 22 minutes environ | Créée à Berlin, en 1796. Première publication : Vienne, 1797

MUSICAL KEY
F Major

GENRE
Musique Romantique Sonata

OPUS
Op. 5 no 1

CASTING
Pianoforte
Violoncello

1796 Ludwig van Beethoven composes the work

1796 Premiere
Créée à Berlin, en 1796

1797 1st Publication
Première publication : Vienne, 1797

1 Performance

1796, Berlin
Performance PREMIERE

1 Publication

1797

Similar works

Ludwig van Beethoven
Sonate pour cor et piano

Ludwig van Beethoven
Sonate "Le printemps" pour violon et piano

Ludwig van Beethoven
Sonate pour piano n. 6

Fig. 6: The detail of an expression in OVERTURE

Other client applications that also make use of the DOREMUS dataset include CityMus [15], a mobile application that generates Spotify playlist composed of DOREMUS tracks based on the surrounding important buildings of a geo-localized user in a city. More precisely, interesting paths in the DBpedia knowledge base between POIs and composer are sought and shown to the end user in order to explain the recommendation. We also develop a chatbot that is capable of answering trivia questions in classical music.[14]

Current Users and Impact. The DOREMUS resource is currently used by librarians internally within each partner institution and across the three institutions, allowing for the fast retrieval of results for complex queries (see Table 1 for a link to examples). Thanks to the exploratory search engine, the DOREMUS data is open for access to a wide community of musicians, music theorists, connoisseurs and amateurs, who do not need to have any technical expertise in order to query the RDF graphs. The controlled vocabularies and the DOREMUS ontology are also being endorsed by IFLA, as a de-facto standard for this community. The French National Library, per its conservation mission, guarantees that the DOREMUS resources will always be accessible and maintained.

[14] `https://chatbot.doremus.org`

Our goal is also to use the resources for both pedagogical and editorial purposes. The recommendation system that is currently under development will assist the creation of playlists for radios, allowing to group works together by very specific criteria, or to uncover rare works and provide insights about possible relations between composers, genres, events, etc.

We contribute to the semantic web community at large by providing open source implementations of novel and generic tools for data linking and fusion. We foster the adoption of semantic web technologies via the publication of numerous pedagogical materials, aiming to guide and encourage other cultural institutions to reuse the DOREMUS model and vocabularies and reproduce our data production framework (see Table 1), as similar initiatives exist in other fields [16].

5 Related Work and Graphs

There has been a significant effort in the last years to open and publish data from the field of cultural heritage [17]. An overview of related projects is given in [1], where the authors provide an evaluation of the various initiatives with regard to the well-known five-stars open data rating, applied to the cultural domain.

Regarding the more specific problem of producing linked data out of library records, addressed by the DOREMUS project, a number of related initiatives have recently been introduced. We refer to the multiple contributions of the Europeana project,[15], unifying and making accessible the catalogs of numerous libraries, museums and archives across Europe. One of the early efforts in that respect is made by the Library of congress,[16] which has become a dataset of reference in the field. In the same spirit, related projects include the German National Library linked data service,[17] the British National Bibliography Linked Data Platform,[18] the open data project of the French National Library BnF[19] or, more recently, the Virtual Library Miguel de Cervantes project [18].

In the majority of the cases, data comes in a given MARC variant and has to be converted to RDF. In certain cases the migration process goes through an intermediate phase of translation to relational database [18], or data is being directly converted to RDF based on the standards of bibliographical description, such as FRBR. DOREMUS follows this line of work by implementing its own expert-defined mappings-based conversion mechanism, enriching FRBRoo with more than 40 classes and 100 properties. The resulting (DOREMUS) model fills the important gap between library content description and music metadata.

As compared to music-related datasets, we outline that the BBC open datasets have tracks only, the Dutch Library (part of Europeana) has only publications, CPDL[20] is specialized for chorus (with scores and midi), while DOREMUS is

[15] http://www.europeana.eu
[16] http://id.loc.gov
[17] http://www.dnb.de/EN/lds
[18] http://bnb.data.bl.uk/docs
[19] http://data.bnf.fr
[20] https://www.cpdl.org/

general and can glue these datasets. MusicBrainz [19], one of the most popular knowledge bases about music metadata, started a few years ago its process of exposing its data as semantic triples through the platform LinkedBrainz [20]. In contrast to DOREMUS, which follows a librarian structure, MusicBrainz follows a more commercial practice giving a central role to tracks, albums and artists (un-distinguishing the composer from the performer), at the expense of all the information connected to the work concept (genre, casting, key, etc).

6 Conclusion and Future Work

We have presented the DOREMUS resource—a collection of linked RDF datasets representing the catalogs of music works of three major French cultural institutions. The construction of this resource implies the implementation of a processing pipeline that allows for the conversion of the original data to RDF following the DOREMUS ontology, the development, SKOS-ification and alignment of a number of music-specific vocabularies and the interlinking of the datasets, which results in the construction of a reference pivot graph of musical works shared by or unique to the three institutions. This pipeline defines the data production paradigm of DOREMUS that is applicable to other music-library data—the described process is deterministic, extensible, reproducible and documented in numerous pedagogical materials published online. A number of tools acting at different layers of this pipeline have been introduced: the `marc2rdf` data converter, the *Legato* data linking system, the web-interface for SKOS thesauri alignment and mapping validation and enrichment YAM++ *on line*, as well as the exploratory search engine Overture. We have relied on existing tools where appropriate (matching strings to URI), but the heterogeneity of the input data and the specificity of the librarian practices made this impossible in many cases. In terms of datasets, DOREMUS currently has published (1) three RDF graphs of musical works coming from the BnF, the Philharmonie de Paris and Radio France, (2) a pivot graph currently containing the certain links established automatically between the graphs of musical works, together with the results of the pairwise linking of these graphs, (3) expert curated benchmarks for evaluation of data linking systems, (4) a rich set of music-specific SKOS vocabularies together with their alignments.

We are currently in the process of applying the data conversion and linking workflow to two additional databases from Radio France. Natural Language Processing techniques are being included in the conversion process in order to parse the numerous free-text fields. Overture will soon host all the links between the interlinked works, giving access at the same time to the joined knowledge and to the different information provenances. We have developed a web interface to assist the process of manual validation of links reducing the human effort, which is currently being deployed online. Alignments of our data to established datasets (in particular MusicBrainz) are currently being generated.

Acknowledgments

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

References

1. J. Marden, C. Li-Madeo, N. Whysel, and J. Edelstein, “Linked open data for cultural heritage: evolution of an information technology,” in *ICDC*, 2013.
2. M. Doerr, C. Bekiari, and P. LeBoeuf, “FRBRoo: a conceptual model for performing arts,” in *CIDOC Annual Conference*, pp. 6–18, 2008.
3. P. Choffé and F. Leresche, “DOREMUS: Connecting Sources, Enriching Catalogues and User Experience,” in *IFLA World Library and Information Congress*, 2016.
4. P. Lisena, K. Todorov, C. Cecconi, F. Leresche, I. Canno, F. Puyrenier, M. Voisin, and R. Troncy, “Controlled vocabularies for music metadata,” in *ISMIR*, 2018.
5. Z. Bellahsene, V. Emonet, D. Ngo, and K. Todorov, “YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation,” in *ESWC, Demo Track*, pp. 137–142, 2017.
6. P. Lisena, M. Achichi, E. Fernández, K. Todorov, and R. Troncy, “Exploring Linked Classical Music Catalogs with OVERTURE,” in *ISWC (Posters & Demos)*, 2016.
7. A. Jentzsch, R. Isele, and C. Bizer, “Silk-generating RDF links while publishing or consuming linked data,” in *ISWC*, 2010.
8. A. N. Ngomo and S. Auer, “LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data,” in *IJCAI*, pp. 2312–2317, 2011.
9. M. Achichi, M. Ben Ellefi, D. Symeonidou, and K. Todorov, “Automatic key selection for data linking,” in *EKAW*, pp. 3–18, 2016.
10. M. Achichi, Z. Bellahsene, and K. Todorov, “Legato: Results for OAEI 2017,” in *Ontology Matching*, 2017.
11. L. Rokach and O. Maimon, “Clustering methods,” in *The Data Mining and Knowledge Discovery Handbook*, pp. 321–352, 2005.
12. D. Symeonidou, V. Armant, N. Pernelle, and F. Saïs, “SAKey: Scalable Almost Key discovery in RDF data,” in *ISWC*, 2014.
13. A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *SIGKDD*, 2016.
14. P. Lisena and R. Troncy, “Combining music specific embeddings for computing artist similarity,” in *ISMIR, Late-Breaking Demo Track*, 2017.
15. P. Lisena, L. Canale, F. Ellena, and R. Troncy, “CityMus: Music Recommendation when Exploring a City,” in *ISWC, Poster Track*, 2017.
16. B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, “Methodological guidelines for publishing government linked data,” in *Linking government data*, pp. 27–49, Springer, 2011.
17. C. Dijkshoorn, L. Jongma, L. Aroyo, J. van Ossenbruggen, G. Schreiber, W. ter Weele, and J. Wielemaker, “The rijksmuseum collection as linked data,” *Semantic Web*, pp. 1–10, 2014.
18. G. Candela, P. Escobar, R. C. Carrasco, and M. Marco-Such, “Migration of a library catalogue into rda linked open data,” *Semantic Web*, pp. 1–11, 2017.
19. A. Swartz, “Musicbrainz: a semantic web service,” *IEEE Intelligent Systems*, vol. 17, no. 1, pp. 76–77, 2002.
20. K. Jacobson, S. Dixon, and M. Sandler, “Linkedbrainz: Providing the musicbrainz next generation schema as linked data,” in *Demo Session ISMIR*, 2010.