Submitted on: 20.06.2016

Data In Libraries: The Big Picture
August 10, 2016
University of Chicago Library

Using Linked Data for Structuring and Interlinking Music Catalogs

How Three Major French Cultural Institutions Finally Came to an Agreement

Marie Destandau
Philharmonie de Paris - Cité de la Musique, Paris, France.

Raphaël Troncy
EURECOM, Sophia Antipolis, France.

Konstantin Todorov
University of Montpellier / LIRMM, Montpellier, France.

Cécile Cecconi
Philharmonie de Paris - Cité de la Musique, Paris, France.

Martine Voisin
Radio France, Paris, France.

Isabelle Canno
Radio France, Paris, France.

Pierre Choffé
Bibliothèque Nationale de France, Paris, France.

Françoise Leresche
Bibliothèque Nationale de France, Paris, France.

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 music domain. In this project, we develop methods to describe, publish, connect and contextualize music catalogs on the web of data. Our focus is on the description of classical and traditional music works as well as their interpretations (events). In this paper, we highlight the difficulty for interlinking music catalogues due to the complexity and the heterogeneity of music data and we demonstrate how semantic web technologies enable to benefit from a common data model that takes advantage of similarities without losing the specificity of each database.

Keywords: FRBRoo, music works, semantic web, data conversion, data interlinking

Introduction

The DOREMUS project aims at bringing together the music catalogs of three major French cultural institutions: Bibliothèque Nationale de France, Radio France and Philharmonie de Paris. While they all contain music metadata, these catalogs exhibit key differences: they use different data models encoded in heterogeneous formats; they respond to specific editorial needs and adopt different viewpoints; they were built over long periods of time by numerous contributors, leading sometimes to potential inconsistency. However, they still share common objectives: collect the music-related memory produced in our society as precisely as possible, and make it available to the widest range of users in a great variety of use cases. Our challenge is to unify these catalogs without losing their richness nor their particularities.

We advocate the usage of linked data technologies, and in particular RDF, as the common data model for representing and interlinking music catalogs coming from various cultural institutions. More specifically, we rely on the FRBRoo model [1], and we propose several extensions composed of classes and properties specific to the music domain [2]. In this paper, we first provide a thorough inventory of the music catalogs with their salient characteristics. We distinguish the description of the entities - works, scores, recordings - that compose the core part of the knowledge base (Section 2) from a large set of controlled vocabularies that are used in music metadata (Section 3). We explain how we have converted those resources into RDF using the DOREMUS ontology for the data and SKOS for the controlled vocabularies. We present our interlinking strategy for connecting those datasets in the Section 4. We conclude and present some future work, and in particular, we briefly introduce the applications that will typically consume this rich set of metadata such as a recommender system or an exploratory search interface that enables to browse the music catalogs (Section 5).

Data Inventory and Conversion

Music Works, Performances, Publications and Recordings

FRBRoo defines a Work as "t*he product of an intellectual process of one or more persons*" [1]. It may be realized through one or several Expressions. For example, the set of signs carried by the original musical score of the Work "Moonlight Sonata", composed by Ludwig van Beethoven, is one of its Expressions.

An Expression class always come with an Expression Creation event. Properties such as the composer name, or the date of creation are attached to the Expression Creation event, while properties such as the key, the mode or the distribution are attached to the Expression itself.

	BnF	**Philharmonie Médiathèque**	**Philharmonie Events**	**Radio France Discotèque**	**Radio France Documentation musicale**	**Radio France Documentation sonore**
	XML / INTERMARC	XML / UNIMARC	XML	XML	XML	XML
uniform titles (UT) & musical work entries	135 940	6 846			62 550	
scores	89 184	30 319			9 154	
books		21 035				
recordings (audio & visual)	156 159	11 049	2 717	340 609	7 700	1 800

Table 1 - Data type collected from the BnF, Philharmonie de Paris and Radio France

Table 1 lists different databases provided by the DOREMUS partners. Some of the entries, e.g. uniform titles or musical work entries, correspond to the level of the Work. Thus, the TUM for Beethoven's Sonate au clair de lune `Beethoven, Ludwig van (1770-1827)[Sonates. Piano. Op. 27, no 2. Do dièse mineur]` generates the creation of a Work and its main Expression. TUMs and work entries also enable the creation of derivation links between works. For instance, the rules for interpreting Philharmonie's work entries are the following:

- If the entry is linked to an authority control with the code 100 (author of the original work), 233 (composer of the adapted work) or 236 (composer of the original work), then two distinct FRBRoo Complex Work classes are created (one for the original

work and one for the adaptation), the two being related using a derivation relation. The properties coming from the database entry are attached to the derived work. Thus, in the original database, `Réminiscences de Simon Boccanegra` de `Verdi` is linked to Verdi authority through the code 100 (`700 $30046861 $aVerdi $bGiuseppe $f1813-1901 $4100`), so a Work is created for the original Simon Boccanegra by Verdi, and another one for the adaptation by Franz Liszt.

- If the entry is not linked to any authority control with the code 100, 233 or 236, then a single FRBRoo Complex Work is created, together with its representative Expression triad, and all properties from the entry are attached to this expression.

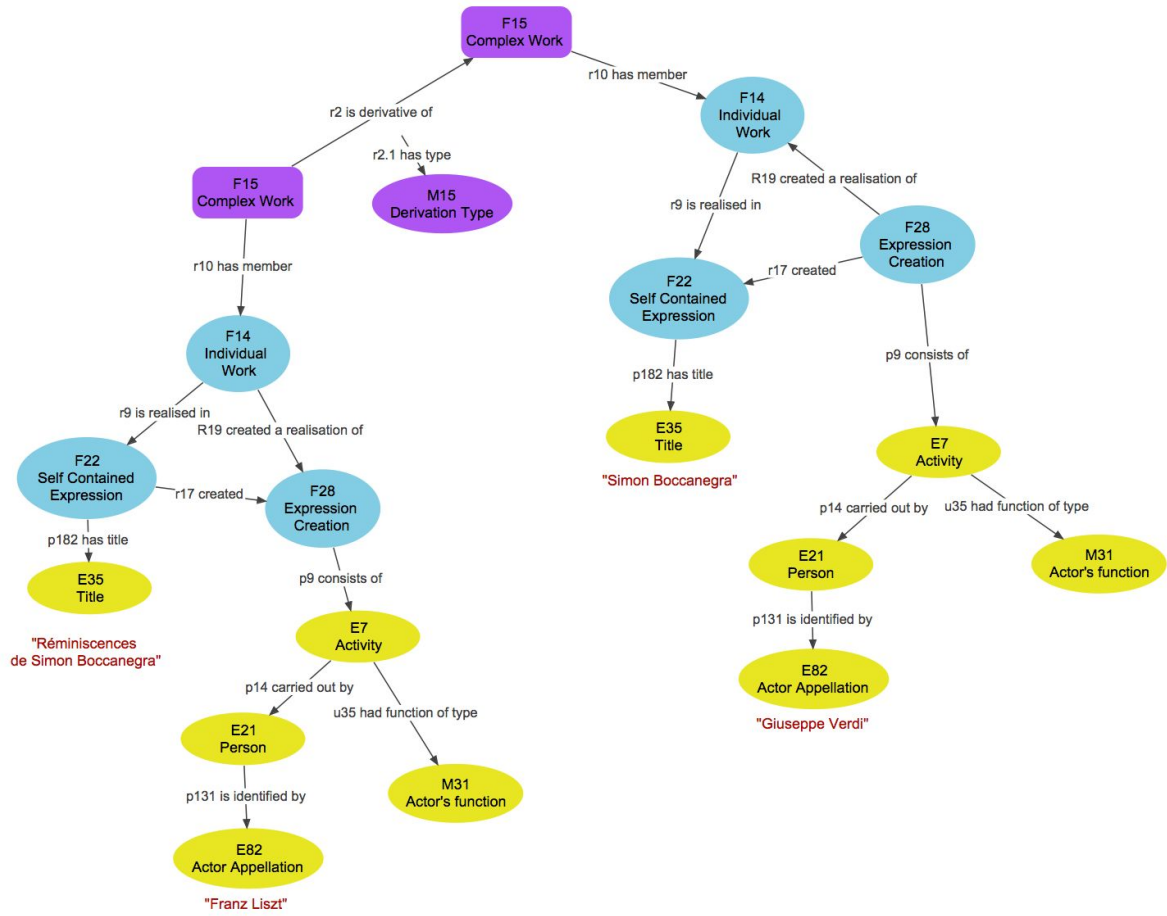

Fig 1 - Relationship between the work Simon Boccanegra by Giuseppe Verdi and its adaptation Réminiscences de Simon Boccanegra by Franz Liszt

Other entries such as score descriptions correspond to the level of the Expression of a work. The publication of the score itself is another Work (a Publication Work) that incorporates the Expression describing the content of the score. Therefore, an Expression and its creation event are materialized: when the score entry contains a link to a work entry, this Expression is linked to the FRBRoo Work resulting from the previous process. The Expression created from the score Six suites pour violoncelle (`200 Six suites pour violoncelle , BWV 1007/12 $bMusique imprimée $dSix suites by J.S. Bach, arrangement for clarinet $dSeis series (!) de J.S. Bach, adaptadas al clarinete $dSech Suiten von J.S. Bach, für Klarinett bearbeitet $fJ.S. Bach $gadaptation par Ulysse`

Delécluse) is linked to the FRBRoo Works that were created from the several suites' work entries, referenced in the score entry:

```
500(|0)  $30769393  $aSuites  $rVioloncelle  $uSol  majeur
$warrangement
500(|0)  $30769398  $aSuites  $rVioloncelle  $uRé  mineur
$warrangement
500(|0)  $30769398  $aSuites  $rVioloncelle  $uRé  mineur
$warrangement
500(|0)  $30769400  $aSuites  $rVioloncelle  $uMi  bémol  majeur
$warrangement
500(|0)  $30769401  $aSuites  $rVioloncelle  $uUt  mineur
$warrangement
500(|0)  $30769402  $aSuites  $rVioloncelle  $uRé  majeur
$warrangement
```

In the case where the original score entry does not reference a work entry, a new FRBRoo Complex Work is created together with the Expression.

Difference of granularity and structure

The DOREMUS model can sometimes represent more detailed information than the individual original database schemas. When possible, the original data is transformed to populate FRBRoo classes. The rules for the analysis are defined in mapping tables. For instance, we take advantage of the fact that loose textual notes are often written in a systematic way, using punctuation to give the illusion of a structure. The rule for extracting the description of the first execution of a work from the UNIMARC field 919$a[1] in the Philharmonie's database is: Remove the following chains of characters and the associated sentence, if they appear at the beginning of a sentence : "Editeur", "Première édition", "1ère édition", "Publication", "Première publication", "1ere publication", "Création française". So if such a note is: "Créé à Berlin, le 5 septembre 2003, par l'Orchestre Philharmonique de Berlin, sous la direction de Simon Rattle, avec Dawn Upshaw (soprano). Création française (avec l'ajout de Gong 2) le 16 septembre 2004 par Barbara Hannigan et l'Orchestre National de France sous la direction de Kurt Masur", then the conversion algorithm will keep "Créé à Berlin, le 5 septembre 2003, par l'Orchestre Philharmonique de Berlin, sous la direction de Simon Rattle, avec Dawn Upshaw (soprano).".

Authorities and controlled vocabularies

The description of music records refer to common entities that one can either find in general encyclopedia (e.g. persons, corporate bodies) or that are specific to the music domain (e.g. functions of persons, keys). Using a list of controlled values for a specific entity enables to foster automatic processing of the data, such as search, interconnection or translation. Table 2 shows the inventory of authority records that we already use or plan to use when converting the data. Some are in-house lists that have been shaped over time by our institutions, while

[1] This is the note field related to the creation of the work

others are standard lists coming from intermarc, unimarc or simply vocabularies published by other institutions.

	BnF	Philharmonie	Radio France Discothèque	Radio France Documentation musicale	Radio France Documentation sonore
Persons & Corporate bodies	* 56 881	* 93 733	459 938	98 490	93 000
Functions	** 504	** 125	123		81
Keys	** 30			32	
Instruments (and voices, IAML)	** 109 *mapping IAML* *** 850 *RAMEAU*	*** 2480 *MIMO* *** 647 *Hornbostel & Sachs*	2 099		314
Modes	** 13				
Range			75		11
Musical genres	**** 859 *IAML* *** 513 *RAMEAU*	** 179	625		40
Catalogs of works	**** 154				
Ethnic groups	*** (+) 3 500 *RAMEAU*		419	296	288
Geographical places		* 2 146	1 888	3 159	3 112
Historical periods		* 39	131	9	
Topical descriptors		* 4 212	5 413	5 510	3 400

Format	* XML / intermarc ** key - value / intermarc *** SKOS / XML **** XLS	* XML / intermarc ** key - value / unimarc *** SKOS / XML	XML	XML	XML

Table 2 - Lists of controlled vocabularies collected from the BnF, Philharmonie de Paris and Radio France databases. The number of stars indicate the format used to represent those controlled vocabularies.

On the difficulty of adopting pivot controlled vocabularies

Whenever possible, we have tried to agree on a common pivot vocabulary, the simplest scenario being to reuse an existing vocabulary, already published on the web of data, used by a community, and covering the needs of our three institutions. For example, Geonames[2] meets all those requirements for geographical places. This is by far the preferred option, since it ensures the vocabulary will be maintained and enriched over time.

Another scenario emerges when there already exists several lists of terms, more or less structured and used by some communities, but varying in terms of granularity and coverage, each with its specificities and richness. For example, for voices and musical instruments, Table 3 shows how the description of the harp instrument family differ from a vocabulary to another. In this case, choosing a single list is really difficult, and merging them into a super authority would not make sense, since they really express different viewpoints. We advocate a method where we first convert those thesauri into a common format, namely SKOS/RDF, and then find alignments between them. The set of vocabularies together with their relationships is then considered as our pivot.

Some of the ad-hoc lists, shaped over time in our catalogues, can be published as part of an ensemble, when we regard them as consistent enough to be of interest. Thus, Radio France's ad-hoc list will be one of the musical instruments ensemble.

	RAMEAU	**Hornbostel & Sachs**	**MIMO**	**IAML**
Terms	- Harpes (famille d'instruments) • Claviharpe • Harpes à chevalet • Bolon • Kora (instrument de musique)	- Harps 322 • 322.1 Open harps – The harp has no pillar. • 322.11 Arched harps. • 322.12 Angular harps. • 322.2 Frame harps – The harp has a pillar • 322.1 Open harps • 322.11 Arched harps.	- Harps • Arched harp • Ngombi • Saung-gauk • Dital harp • Ardin • C'angi	Harps can be found among dozens of other instruments in the following categories : - A/4. Strings, bowed - A/5. Strings, plucked - A/9. Miscellaneous,

[2] http://www.geonames.org/

	• Soron • Harpes angulaires • Ardin • Changi • Hares arquées • Arpa • Bolon • Harpe • Harpe celtique • Kinde	• 322.111 Arched harps - Wachsmann type 1 • 322.112 Arched harps - Wachsmann type 2 • 322.113 Arched harps - Wachsmann type 3 • 322.114 Arched harps sounded by the bare fingers • 322.12 Angular harps. • 322.2 Frame harps • 322.21 Frame harps without tuning action. • 322.211 Diatonic frame harps without tuning action. • 322.212 Chromatic frame harps without tuning action. • 322.212.1 Chromatic frame harps without tuning action, with the strings in one plane. • 322.212.2 Chromatic frame harps without tuning action, with the strings in two planes. • 322.212.3 Chromatic frame harps without tuning action, with the strings in two or more parallel planes. • 322.22 Frame harps with tuning action. • 322.221 Frame harps with manual action. • 322.221-5 Frame harps with manual action sounded by the bare fingers • 322.222 Frame harps with pedal action. • 322.222-5 Frame harps with pedal action sounded by the bare fingers	• Chromatic harp • Clarsach • Irish harp • Ennanga • Harp • Hook harp • Kin • Kinde • Kundi • Nedomu • Ombi • Para • Spike harp • Bolo bogo • Bolon • Chang • Diatonic harp • Domo • Ground harp • Kwarnda • Waji • Wa-konghu	other, unspecified instruments
Languages used	French	English	Catalan, German, English, French, Italian,	Arabic, Basque, Catalan, Croatian, Czech, Dutch, English, French, German, Greek,

			Dutch, Polish, Swedish, Chinese (all terms translated in all languages)	Hebrew, Hungarian, Italian, Japanese, Latin, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish (languages depend on the terms)

Table 3 - The "harps" instruments families in MIMO, Hornbostel & Sachs, RAMEAU and IAML thesauri

There are also cases where no existing list describes an entity according to our needs. When the number of terms to be considered is relatively small, an editorial work is carried out to propose a single new authority which represents the shared viewpoint of the cultural institutions. This is what we have done for musical keys or types of derivation between works. However, when the work is too big to be accomplished within the frame of this project, as for rhythmic patterns, we plan the description of the entity in the model, but do not publish a corresponding vocabulary.

Conversion to RDF

We use Open Refine to convert the original controlled vocabularies into SKOS/RDF, since often, they were available as spreadsheet formatted information. The conversion involves the creation of a URI for identifying the `skos:Concept` following a strict URL design policy. Simple properties such as labels (preferred and alternative ones in different languages), descriptions and usage notes are attached to each concept. Sometimes, the structure of the thesauri has to be made explicit. For example, in the IAML thesaurus of music instruments, the code identifying each term conveys directly its position in the hierarchy that we need to make explicit using `skos:broader` and `skos:narrower` relationships. Finally, at the border between a controlled vocabulary and a small specific dataset, the "Catalogs of works" are lists of works by a specific composer, commonly referred to in order to identify a work without ambiguity. We use the MODS vocabulary rather than SKOS or even DCAT to describe those resources. Hence, the information described in the Table 4 will be converted in RDF as depicted in the Figure 2.

Catalogues thématiques et catalogues d'œuvre	**Édition**	**Catégorie**	**Indice**	**Institution utilisant cette référence**	**Citation abrégée (Autorités de BnF catalogue général)**	**Lien à la notice bibliographique dans BnF catalogue général**

SAINT-SAENS, CAMILLE (1835-1921) Ratner, Sabina Teller. – Camille Saint-Saëns : 1835-1921 : a thematic catalogue of his complete works. – Volume I : The instrumental music ; Volume II : The dramatic works. – Oxford : Oxford university press, 2002-2012	2002-2012	Catalogue thématique	R	BnF	Ratner, Saint-Saëns	http://catalogue.bnf.fr/ark:/12148/cb388551727/PUBLIC http://catalogue.bnf.fr/ark:/12148/cb426686040/PUBLIC

Table 4 - Example of a Work catalog entry in the list of thematic catalogs.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix mods: <http://www.loc.gov/standards/mods/modsrdf/v1/#> .
<http//data.doremus.org/workcatalog/UUID>
      mods:titlePrincipal@en "Camille Saint-Saëns : 1835-1921 : a
thematic catalogue of his complete works" .
      mods:title@fr "Ratner, Saint-Saëns" .
      mods:issuance@fr "Catalogue thématique. Volume I : The
instrumental music ; Volume II : The dramatic works." .
      mods:subjectName "SAINT-SAENS, CAMILLE (1835-1921)" .
      mods:name "Ratner, Sabina Teller" .
      mods:identifier:indice "R" .
      mods:publisher "Oxford university press (Oxford)" .
      mods:languageOfResource "en" .
      mods:dateOfCopyright "2002-2012" .
      owl:sameAs                                                    :
<http://catalogue.bnf.fr/ark:/12148/cb388551727/PUBLIC>,
<http://catalogue.bnf.fr/ark:/12148/cb426686040/PUBLIC> ;
```

Figure 2 - The result of the RDF transformation on the data depicted in Table 4

Visualization

Though the structure of a thesaurus is relatively simple, it's not easy to have a clear insight of its shape and specificities by looking at RDF description. We have developped a tool for visualizing SKOS controlled vocabularies that enables to inspect and assess the quality of the data. It shows the detail of a concept, as well as its position in the hierarchies, and links to other vocabularies.

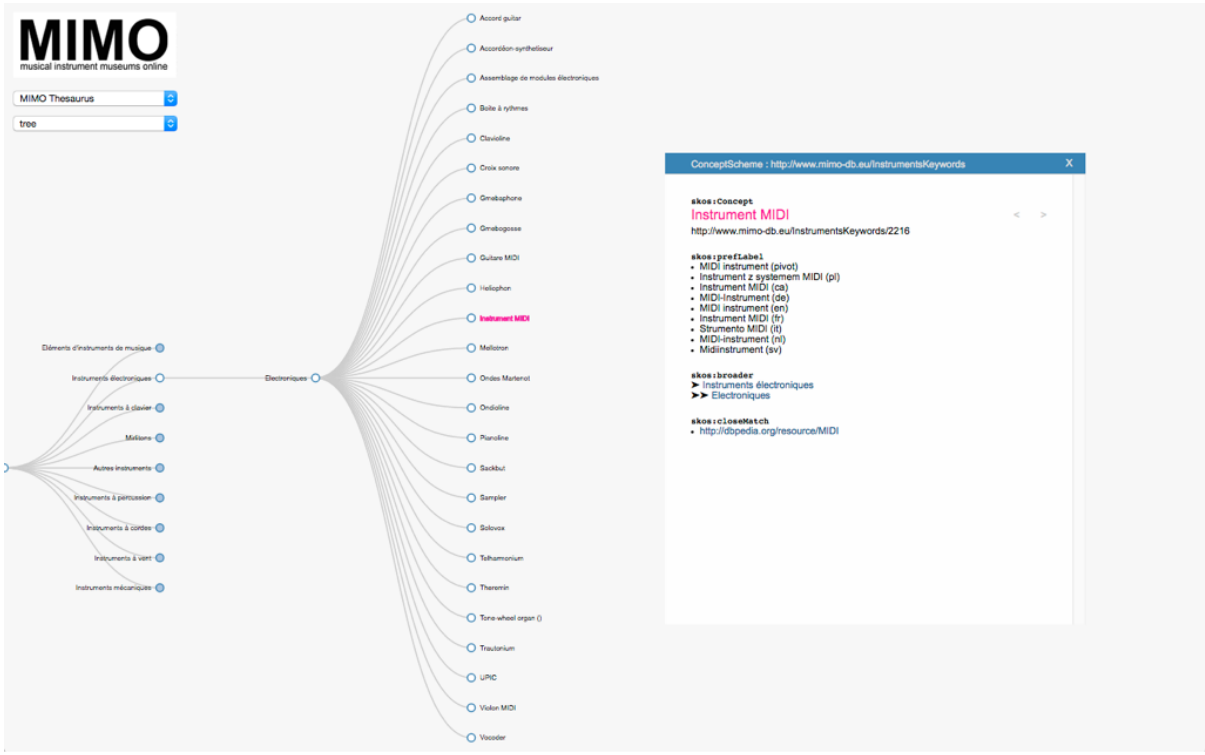

Figure 3 - VIZSKOS, a thesaurus visualization tool

Interlinking the data

Both controlled vocabularies and datasets, once being converted in RDF, need to be interlinked. The DOREMUS controlled vocabularies are pairwise aligned, as well as interlinked with popular knowledge graphs such as DBpedia and Wikidata. We use the SKOS mapping relationships (e.g. `skos:exactMatch`, `skos:narrowMatch`) for materializing those alignments.

Data interlinking is defined as the process of establishing relations between resources across datasets. Most commonly, one is interested in the discovery of exact equivalence relations expressed by the `owl:sameAs` statement.

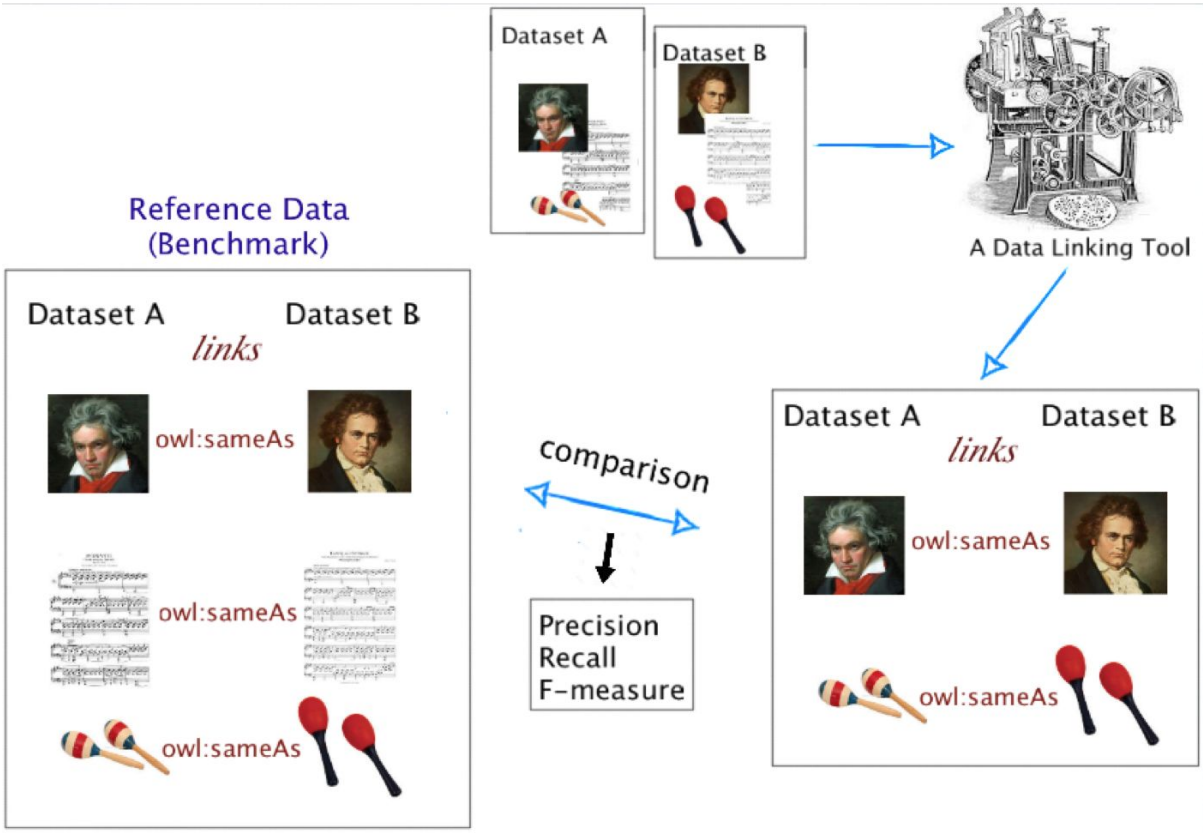

Figure 4 - development and evaluation of interlinking tools

Self-duplicates

The RDF conversion process may lead to the creation of multiple identities for the same Work within the same dataset. There are two possible ways to handle this issue: i) one can check before creating a new Work if a similar one does not exist already, and if so, reuse the existing identifier; ii) alternatively, one can systematically create new URIs each time that a Work is referenced, and then, identify and reconcile the duplicates.

Different type of heterogeneity to overcome

The data conversion process results in a number of RDF datasets: one for the BnF, one for the Philharmonie de Paris and three for Radio France (Figure 5). There is an important overlap of the resources described in these datasets, especially in terms of musical works, but also in terms of events. Therefore, a central data management task within the project is the linking of the resources across these RDF graphs. The policy that has been adopted consists in establishing `owl:sameAs` links between the resources of each of the institutional RDF datasets and a global, pivot graph (the union of the unique resources across all datasets), as shown in Figure 5. To this end, we are currently exploring and adapting existing data linking and instance matching approaches and tools (OnaGui, LYAM++ [3], SILK [4]).

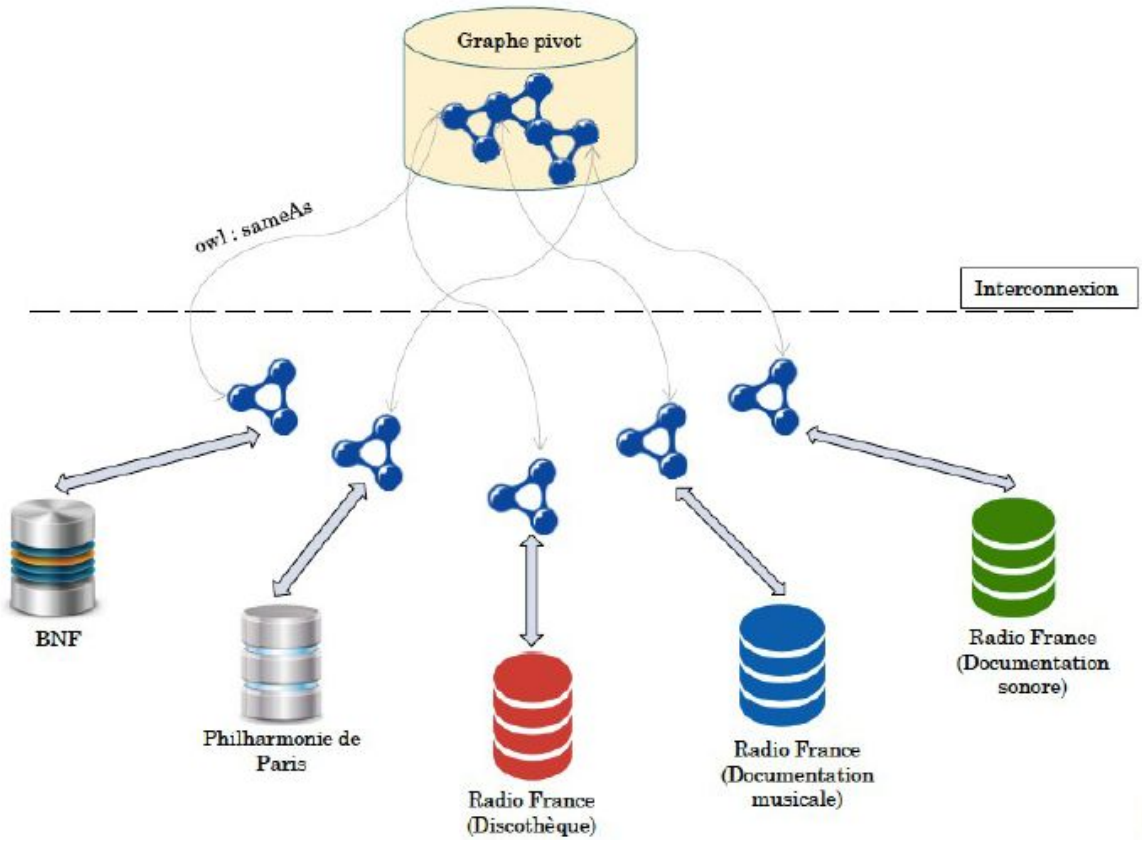

Figure 5 - Interlinking of the various datasets to a pivot graph that contains a unique identifier of all resources

Data heterogeneities

In order to train and evaluate the instance matching approaches that we develop, we have created a benchmark dataset containing pairs of equivalent cross-institution works (Figure 4). We have asked the experts to identify the most commonly encountered data heterogeneities (differences in the descriptions of equivalent resources) in the music field. In addition, we were able to identify a set of complementary heterogeneities by looking directly into the data and testing with data linking tools. This gave rise to the identification of nine categories of heterogeneities, altogether: H1: presence of letters and numbers in the title of the work; H2: terminological / orthographic differences; H3: the works do not have a catalog/opus number; H4: works described in different catalogues; H5: the titles of the works are multilingual; H6: the description of a work contains diacritical characters; H7: different value distances: a property can be specified directly through a literal value while the same information can be given in a longer property chain including several triples; H8: the same information across two instances is described by different properties; H9: lack of description: an instance can be described by more properties in one dataset than in the other. Seven of these heterogeneities are depicted in the example in Figure 6.

We base our approach on the capacity of the linking tools to handle these 9 heterogeneities. The first results that have been obtained using SILK show that three of these nine heterogeneities appear to be more problematic than the other, namely H2 (orthographical

differences), H5 (multilingualism), H9 (lack of description). Our first observation is that more effort has to be directed towards the development of methods that allow to link correctly works that manifest these heterogeneity types.

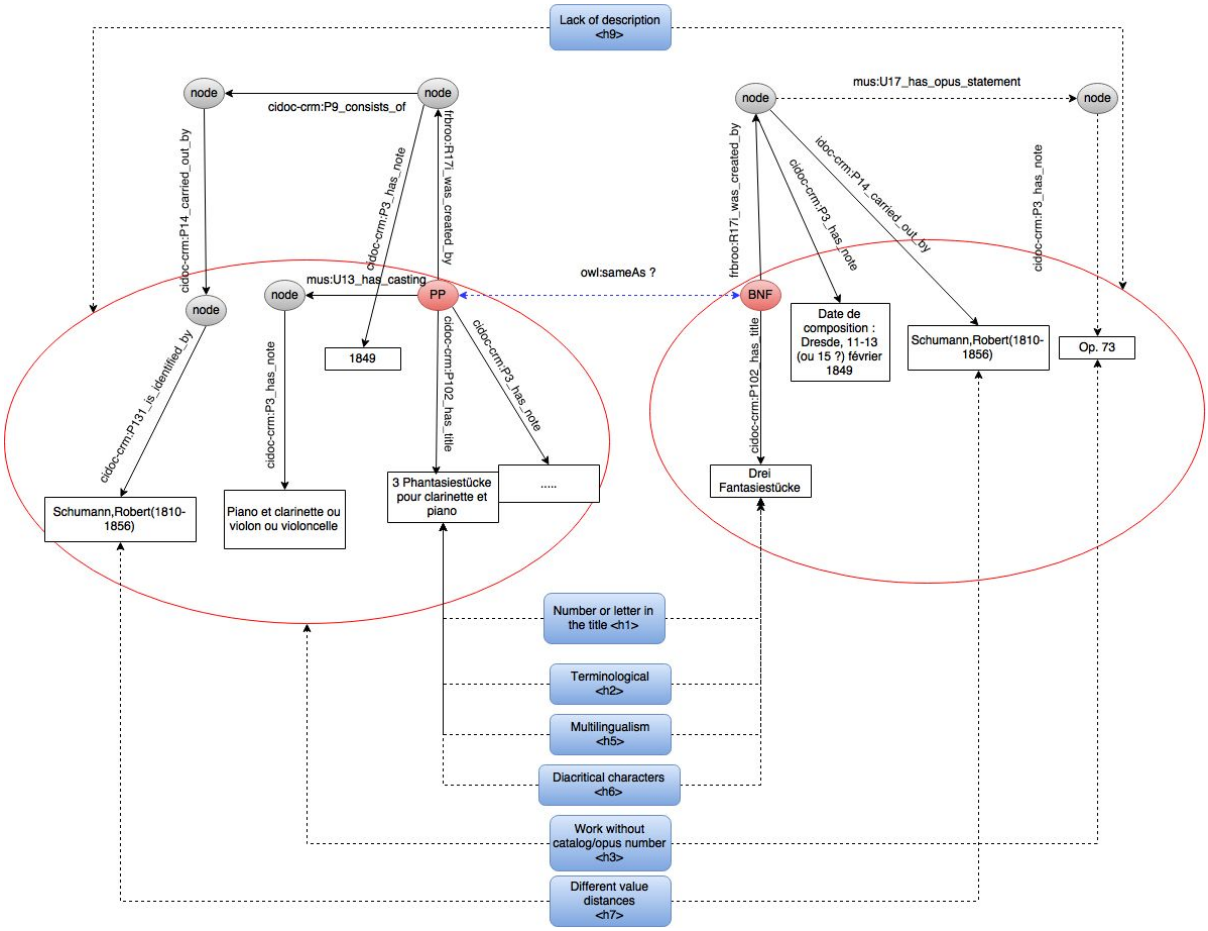

Figure 6 - Example of two heterogeneous works (BnF and PP that stands for Philharmonie de Paris).

Finally, we are also interested in mining more complex links expressing relations between elements of the model such as the interpretation of a work, the meronymy relation, the relation between a work and a number of events. Therefore, we aim to propose an approach that enables the automatic inference of these relationships using the established equivalence links.

Conclusion and Future Work

In this paper, we have presented how we represent and interlink music catalogs coming from three major cultural institutions. We rely on the FRBRoo model extended with new classes and properties that are part of the DOREMUS ontology. We ultimately publish the interlinked datasets following the linked data principles. Hence, the resulting RDF datasets have been loaded in a triple store which is publicly available for query at http://data.doremus.org/sparql.

OVERTURE: an exploratory search engine for music catalogs

We are developing OVERTURE, an exploratory search engine that enables to browse the interlinked music catalogs and that facilitates serendipitous discovery. We are also developing powerful content based recommendation algorithms that uses the richness of the semantic graph that describes each Work and Expression. The application is available at http://data.doremus.org/overture.

Connectors of music works

We are currently working on the development of connectors for music works, allowing for the automatic interlinking of equivalent resources across datasets of different bibliographical agencies. Our first test with the SILK [4] and LIMES [5] tools confirm our initial hypothesis that new methods, specific to the music field, need to be proposed, in order to handle the high heterogeneity of these datasets. In particular, our first tests show that special attention has to be payed to multilingual descriptions of works, as well as to significant differences in lexical descriptions of music titles. Descriptive heterogeneities (level of detail, amount of information available) also appear to be hard to handle by the existing general purpose off-the-shelf linking tools. We plan to combine expert heuristics with an automatic heterogeneity centered data linking approach.

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] Martin Doerr, Patrick Le Boeuf and Chryssoula Bekiari. FRBRoo, a conceptual model for performing arts. In Annual Conference of CIDOC, Athens, Greece 2008.
[2] Manel Achichi, Rodolphe Bailly, Cécile Cecconi, Marie Destandau, Konstantin Todorov and Raphaël Troncy. DOREMUS: Doing Reusable Musical Data. In 14th International Semantic Web Conference (ISWC), Poster Track, CEUR Proceedings Vol. 1486, Bethlehem, PA, USA, 2015
[3] Abdel Nasser Tigrine, Zohra Bellahsene and Konstantin Todorov. Light-Weight Cross-Lingual Ontology Matching with LYAM++. In 14th International On the Move to Meaningful Internet Systems (OTM) Conference, pages 527-544, Rhodes, Greece, 2015.
[4] Julius Volz, Christian Bizer, Martin Gaedke and Georgi Kobilarov. Silk-A Link Discovery Framework for the Web of Data. In 2nd International Workshop on Linked Data On the Web (LDOW), Madrid, Spain, 2009.
[5] Axel-Cyrille Ngonga Ngomo and Sören Auer. LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain, 2011.