IFLA WLIC 2016 COLUMBUS

Submitted on: 16.10.2017

Data in libraries: the big picture
10 August 2016
University of Chicago, Regenstein location, Chicago, IL, USA

Data processing for digital libraries: the experience of the BnF with Europeana Sounds project

Anila Angjeli
Bibliothèque nationale de France, Metadata department, Paris, France.
anila.angjeli@bnf.fr ISNI 0000 0004 2755 4724

Bertrand Caron
Bibliothèque nationale de France, Metadata department, Paris, France.
bertrand.caron@bnf.fr ISNI 0000 0004 3238 8249

Emmanuelle Bermes
Bibliothèque nationale de France, Directorate of services and networks, Paris, France.
emmanuelle.bermes@bnf.fr ISNI 0000 0001 2355 8080

ABSTRACT:

As one of the 24 partners of the project Europeana Sounds, the BnF provides to Europeana metadata related to digitized sound recordings from its cultural heritage collection. This article describes the BnF experience with the Europeana Sounds project by taking as its point of departure the multiple challenges related to data transformation from base material, the MARC XML records that comprise also metadata related to digital objects, into the RDF based Europeana Data Model (EDM).
By way of this experience the BnF is exploring technical methods for exposing and sharing in the best possible way the richness and granularity of its metadata, as well sharing its areas of expertise, such as controlled vocabularies and best practices with persistent identifiers.
The array of issues addressed involves instructions that combine analysis of traditional bibliographic description, digital library metadata curation, critical look on the extant relationships between various data repositories within the BnF, current and future role of library catalogues, metadata production workflows, role of metadata quality, licencing issues, and more.
The article discusses the organization of work so that skills and responsibilities are shared in the most efficient way. It also discusses lessons learned and perspectives regarding the evolution of the profession of the metadata librarian, including the need for a better understanding of the potential for the metadata to be reused and reprocessed, and the necessity to develop modeling and technical skills.
Finally, in exploratory mode, the article highlights the possibilities offered by the Linked Open Data technologies for metadata reprocessing in innovative ways by taking down the data silos borders.

Keywords: Metadata processing, Digital library, EDM, Semantic Web technologies, Sound recordings

1 Introduction: the Europeana Sounds Project

Europeana Sounds is a three-year project, scheduled to run from 2014 to 2017, funded under the European Competitiveness and Innovation (CIP) framework programme.[1] The objective is to "increase the opportunities for access to and creative re-use of Europeana's audio and audio-related content and […] build a sustainable best practice network of stakeholders in the content value chain to aggregate, enrich and share a critical mass of audio that meets the needs of public audiences, the creative industries (notably publishers) and researchers".[2] The project comprises 24 partners from 12 European countries, ranging from Ireland to Greece and from Latvia to Italy. The partners represent national libraries, archives and research centres, universities, non-profit organisations as well as private companies interested in opening to the widest audience the European sound and music heritage. The partners form the Europeana Sounds Consortium.

The activities of the project are organised in **seven thematic work packages**: aggregation, enrichment & participation, licensing guidelines, channels development, technical infrastructure, dissemination & networking, project management & sustainability.[3]

2 Participation of the BnF in the project – Definition of the corpus and its characteristics

2.1 BnF digital collections of sound recordings

The BnF digital heritage sound recordings are part of the BnF digital library Gallica.[4] Online since 1997, Gallica is one of the world pioneers in the effort of opening the library heritage collections to a global audience. More than 3 million materials are currently available and this collection grows by thousands of new digitized contents each and every week.

The BnF Audiovisual Department is in charge of the sound recordings collection of the BnF. With more than one million items, this exceptional collection founded in 1911 is one of the oldest and of the most prestigious in the world. The collection is gradually digitized and partly made available online on Gallica.[5] The Audiovisual Department monitors the process of the digitization and online publication of the cultural materials under its care, including editorial enhancements on Gallica. The online contents amount to around 35,000 documents, 5,000 of which will feature on Europeana Sounds, representing only a selection of the wealth and of the diversity of the entire BnF sound recordings collections. The selection contains recordings from the Archives of Spoken Word (Archives de la Parole); recordings from the International Colonial Exhibition of 1931 in Paris; recordings of testimonials, famous

[1] A program of the European Commission http://ec.europa.eu/cip/

[2] From the reference document "Europeana Sounds Description of Work"

[3] A dedicated website "Europe's sound heritage at your fingertips" provides overall, up-to-date information on the project http://www.europeanasounds.eu/about

[4] http://gallica.bnf.fr/

[5] The entire collection is digitized for preservation purposes, and a selection of it is made available on Gallica. The selection is based on a variety of criteria such as thematic, conditions of use, etc. For more information see: http://gallica.bnf.fr/html/und/enregistrements-sonores/fonds-sonores

speeches and songs from the period of World War I; recordings of lectures from notorious personalities such as the lectures from the philosopher Gilles Deleuze; musical recordings; interviews with eminent figures of the phonographic publishing industry, and more. From the carriers perspective the BnF sound recordings collection includes wax cylinders, shellac, acetate, vinyl, tape, mp3 and CDs.

The descriptive metadata related to the BnF digital contents from Gallica are made available on the BnF OAI-PMH repository for digital contents[6] and are regularly harvested by Europeana.[7] The descriptive metadata for digitized sound contents are, consequently, part of this harvesting process and are displayed on the Europeana Collections portal. However, the metadata on the BnF OAI-PMH repository for digital collections are less granular and rather impoverished as compared to the original descriptive metadata created by the cataloguers. Furthermore, over the course of the transformation process performed by Europeana to make them fit into the aggregation data model used by the latter, these metadata are subject to additional contractions, reductions, and losses. Predictably enough, such successive contractions and reductions lessen considerably the potential for the metadata to be correctly mined, thereby boiling down the searchability of the collections themselves on the Europeana Collections portal.

2.2 Motivation of the BnF for participating in Europeana Sounds project

The Europeana Sounds project requires each of the partners to monitor the transformation of its own metadata to the Europeana Data Model (EDM)[8] – the data model Europeana uses for aggregating metadata from its numerous data providers[9].

The BnF viewed this constraint rather as an opportunity from a variety of perspectives.

- The first motivation was to experiment with the semantic expressiveness of EDM and to provide to Europeana rich and fine-grained metadata (including links to collections descriptions, controlled vocabularies for agents, phonographic labels, concepts, places, works…), as opposed to the DC simple formatted metadata harvested by Europeana from the BnF OAI-PMH repositories.
- In addition, transforming metadata directly from the source format to EDM would avoid the loss of information that occurs over the course of the current routine of successive transformations, as described above.
- Satisfactory metadata transformation requires a combination of good knowledge of source formats and metadata of the domain, more specifically, the characteristics of the selected collection, as well as knowledge of the relevant technologies and tools. This would also enable BnF to familiarize itself with EDM and better communicate with Europeana. To meet these requirements BnF metadata experts would be in charge of the data remodelling and of the entire data reprocessing; they would also work in close

[6] Identification of the BnF OAI-PMH repository for digital materials: http://oai.bnf.fr/oai2/OAIHandler?verb=Identify. This repository provides the descriptive metadata for BnF digital contents available on Gallica. Note that the descriptive metadata for contents from other BnF partners included in Gallica are not provided in this repository.

[7] Note that the BnF provides to Europeana only the descriptive metadata for the digitized objects. Persistent URIs to the digitized contents on Gallica are included in the metadata, thereby ensuring that the objects are viewed in their original institutional environment. The digitized contents themselves are not provided to Europeana.

[8] http://pro.europeana.eu/page/provide-data-edm

[9] Europeana aggregates metadata from more than 3,500 institutions whether directly or via intermediate aggregators.

collaboration with the curators of the sound recordings from the Audiovisual Department. This would guarantee the optimal quality of metadata supplied.

– Alongside the BnF, another French partner, the French National Centre for Scientific Research (CNRS), participates in Europeana Sounds with content from the domain of ethnomusicology. In the meantime, in France, the BnF and the CNRS collaborate (and also with the quai Branly museum – Jacques Chirac) for the creation of a common platform that aggregates ethnomusicology-related contents held in these institutions. This platform will use EDM as the aggregation data model. This two-fold collaboration is an excellent opportunity to reinforce exchanges between the BnF and the CNRS teams and create synergy between both projects.

Expected results include significant improvement of the discoverability and visibility of the BnF sound heritage collection on the Europeana Collection portal; increase of traffic on Gallica; broad diffusion of the BnF controlled vocabularies in an environment that uses Semantic Web technologies, connecting BnF collections with those of the French partners and other European partners in a meaningful way, and more.

3 Metadata terms of use: legal issues

3.1 Rich fine-grained metadata *versus* Simple DC: licence issues

From January 1st 2014, the BnF provides all descriptive metadata under the French State Open Licence[10], which authorises the reuse of metadata free of charge and for any purpose, provided the re-user acknowledges its source provenance. Europeana, on the other hand, requires that each of its data providers comply with the Europeana Data Exchange Agreement (DEA)[11] according to which, metadata submitted to Europeana will be published as open data under the terms of the Creative Commons Zero Public Domain Dedication (CC0). The BnF, as the historical founding member of Europeana and one of its fervent supporters, has signed the Europeana DEA and allows metadata from the BnF OAI-PMH repository to be harvested by Europeana under CC0, thereby making an exception to the BnF overall policy related to metadata terms of use.

The metadata on the BnF OAI-PMH repository are though formatted in simple Dublin Core. They are therefore far less granular than the original BnF General Catalogue metadata and do not include links to the controlled vocabularies. In the specific case of sound recordings, this substantially penalizes the discoverability and the visibility of the BnF sound collection on the Europeana Collections portal.

As already mentioned, the primary interest of the BnF Audiovisual Department for participating in Europeana Sounds project was to enhance the discoverability and visibility of the BnF sound heritage collection on the Europeana Collection portal. This could be achieved if and only if the provided metadata is fine-grained and contains all related links as described above. The issue was long discussed within the BnF. As the Europeana Sounds project would serve multiple purposes, including experimenting with the expressiveness of EDM as described in part 2.2 of this article, it was agreed that it would be beneficial for the BnF in many respects that the metadata that would be reprocessed for the project should not be

[10] BnF website page describing the application by the BnF of the French State Open License: http://www.bnf.fr/fr/professionnels/anx_recuperation_donnees/a.ouverture_donnees_bnf.html.
The licence itself is available from: https://www.etalab.gouv.fr/wp-content/uploads/2014/05/Open_Licence.pdf.

[11] http://pro.europeana.eu/page/the-data-exchange-agreement All metadata submitted to Europeana will be published as open data under the terms of the Creative Commons Zero Public Domain Dedication (CC0). http://pro.europeana.eu/files/Europeana_Professional/DEA/Data%20Exchange%20Agreement.pdf

extracted from the OAI-PMH repository but directly from the source BnF General Catalogue. These metadata would be provided to Europeana under CC0 for experimental purposes.

3.2 Rights statements issues for digital contents: Public Domain *versus* Gallica specific conditions of reuse

While the BnF digitized sound recordings selected for the Europeana Sounds project are mostly no more protected by intellectual property rights, their reuse is subject to conditions specific to Gallica.[12] Indeed, if there are no more copyrights or related rights on these documents, they are submitted to other kinds of rights, which create limitations to their access and reuse.[13] Until recently the Europeana Terms for User Contribution[14] relied exclusively on Creative Commons licenses and did not provide any appropriate rights statement that could enable the cultural institutions to adequately communicate the conditions of reuse of their contents. Consequently, all BnF contents are currently given *de facto* and inappropriately the Public Domain Mark licence on the Europeana Collections portal.

This issue was discussed within the Europeana Sounds project work package devoted to licensing guidelines. Concomitantly, Europeana, the Digital Public of America[15], Creative Commons joined their efforts to provide cultural institutions with simple and standardized terms to communicate the copyright and re-use status of their digital objects to the public. These endeavours led to the creation of RightsStatements.org which provides 12 standardized, internationally interoperable rights statements[16], covering the variety of cases of in copyright, no copyright and with unclear copyright contents. The rights statements have been designed both for human users and machines and each rights statement is located at a unique URI. The BnF is looking forward to the implementation of these rights statements by Europeana, in particular of the statement "No Copyright – Other Known Legal Restrictions (NC-OKLR)"[17] which, actually, is the only rights statement that corresponds to the reality of the legal status of its contents.

4 Input data

As shown in the diagram below, the base metadata to be reprocessed and supplied to Europeana Sounds come from the bibliographic records for sound recordings created in the BnF General Catalogue. These metadata are expressed in an in-house MARC format called INTERMARC. In addition to being highly granular, a prominent feature of INTERMARC is that it allows for dynamic links among the bibliographic records and the related authority records that are part of controlled vocabularies (for agents, concepts, places, dates, phonographic labels, uniform titles for musical and other types of works), as well as the descriptive records for archival *fonds*, editorial sets and collections. Links between records are managed by local system identifiers. These identifiers are the basis for the creation of

[12] Conditions for the use of Gallica's contents: http://gallica.bnf.fr/html/und/conditions-use-gallicas-contents

[13] In this specific case, the limitation of the access provides from the PSI Directive (Directive 2003/98/CE of the 17 november 2003, modified by the directive 2013/37/UE of the 26 june 2013) and the french law n°2015-1779 of the 27 december 2015, which permits public institutions to limit access to public informations.

[14] Europeana Terms for User Contribution : http://www.europeana.eu/portal/rights/contributions.html developed within the Europeana Licensing Framework http://pro.europeana.eu/page/europeana-licensing-framework

[15] https://dp.la/

[16] http://rightsstatements.org/page/1.0/?language=en

[17] http://rightsstatements.org/page/NoC-OKLR/1.0/?language=en

persistent ARK URIs that enable the unique identification of each entity description on the Web (whether a bibliographic description of a cultural heritage object, a *fonds*, editorial set and collection description, or an authority record).

The input data for Europeana Sounds project is a catalogue extraction in XML expression of the INTERMARC format, called internally InterXMarc, gathering information from the full bibliographic record itself, from the description of the various holdings, including digital and analogue items (call number, technical characteristics like file format, rights, etc.), as well as from the related authority records and records of *fonds*, editorial sets and collections. Though this format was originally created to address the internal needs of the IT department, it proved to be the best source data for metadata mappings and analysis and is likely to be used extensively for other projects.

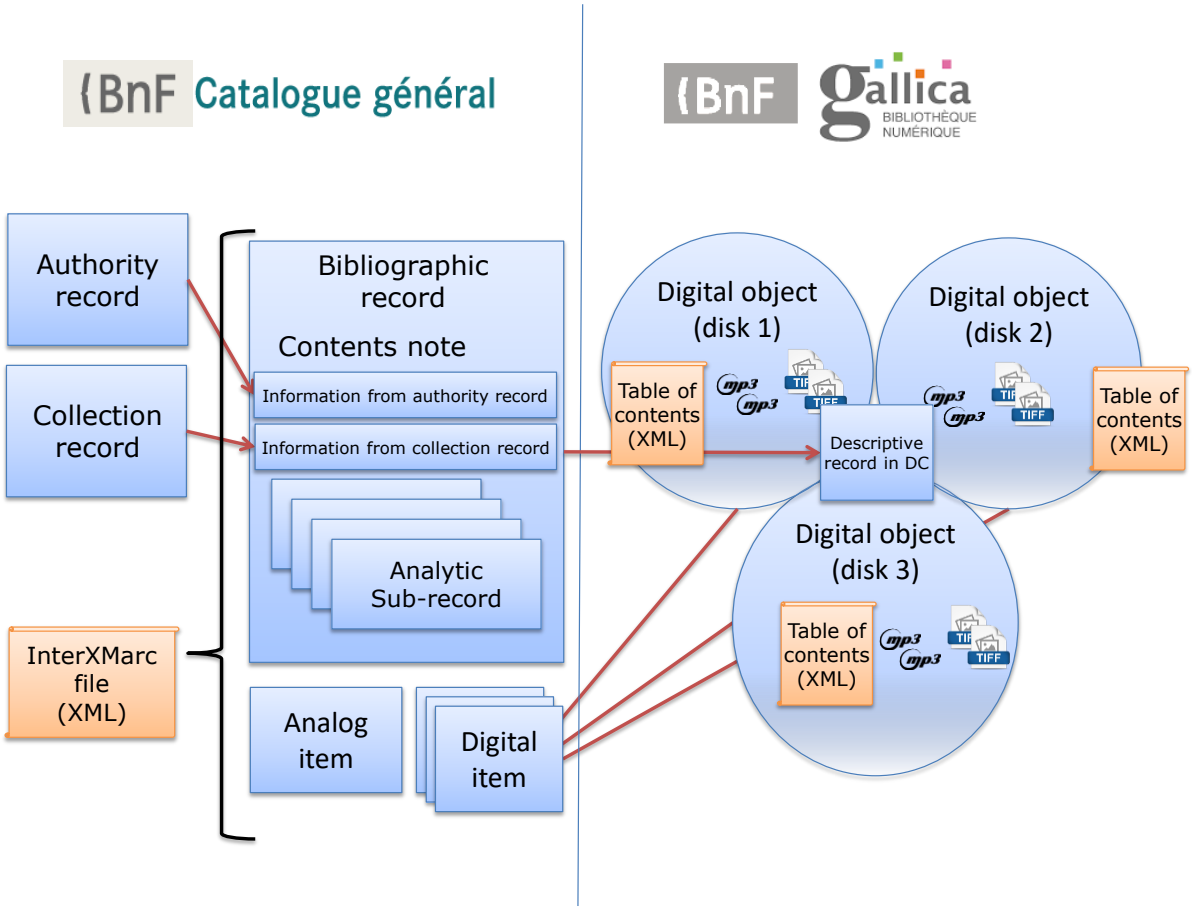

Figure 1. Relations between the BnF Catalogue environment and the BnF digital library Gallica environment

Each cultural object, whether consisting of one or several physical carriers (e.g. an album containing two or more shellacs), is described in one single bibliographic record. While the analogue item representing a cultural object in the BnF Catalogue encompasses all physical discrete carriers (volumes, issues, disks, etc.), its digitization creates as many digital items as the number of carriers of the analogue item.

The internal structure of each object is described by the cataloguer via different and sometimes competing mechanisms. The catalogue records describe the cultural objects from the intellectual perspective. The intellectual components of the object (songs, classical music works, etc.) are listed, alternatively and for historical reasons, either in a contents note, or described individually in separate analytic sub-records that provide detailed, structured

information on each contained recorded work, (e.g., performers for each song, recording information for each song, uniform titles for classical music pieces, etc.). Unfortunately, due to cataloguing practices, in both cases, the physical structure of the object is not reflected in the bibliographic record, thereby making impossible to associate each intellectual component to its own carrier and, consequently, to the digital files which represent them.

The BnF's service providers in charge of digitization also provide additional descriptive and structural metadata related to both the physical and the intellectual composition of the object (table of contents identifying carriers, sides, tracks, etc.). These metadata are stored in files that are part of the specific deliverables in the digitization process and are stored in the digital repository. This is the only place where each intellectual component of a cultural object is accurately connected to its carrier (disk and side of the disk). Regretfully, the cataloguing chain and the digitization chain follow different paths. Though the presence of a record in the catalogue is the condition that triggers the digitization of an object, the additional information created over the digitization process is not fed back into the catalogue. The only connection is made by way of the ARK identifiers (the ARK identifier of the catalogue record is included in the DC descriptive record for the digital item and, conversely, the ARK identifier of the digitized item is included in the catalogue record at the level of the information regarding the holding). Depending whether the analogue object is composed of one or many physical carriers, this connection is one-to-one or one-to-many.

5 Processing data: challenges & issues

5.1 Workflow and tools

In order to provide to Europeana metadata transformed in EDM, developing the mapping and processing data was the core action of the BnF project team. For the accomplishment of this task, the Europeana Sounds Consortium was supplied with MINT[18] (Metadata INTeroperability), a tool developed by the National Technical University of Athens (NTUA). MINT is a web-based platform for ingestion of CSV or XML files and mapping from standard or proprietary metadata records to EDM. It provides a visual mapping editor and a results preview of the defined mappings as applied to the ingested set of records. During the development phase, validated sets of records in EDM were published on the Europeana portal every two weeks. This process allowed for iterative refinement of mappings from the historical bibliographic format INTERMARC to the RDF-based EDM.

[18] http://mint.image.ece.ntua.gr/redmine/projects/mint

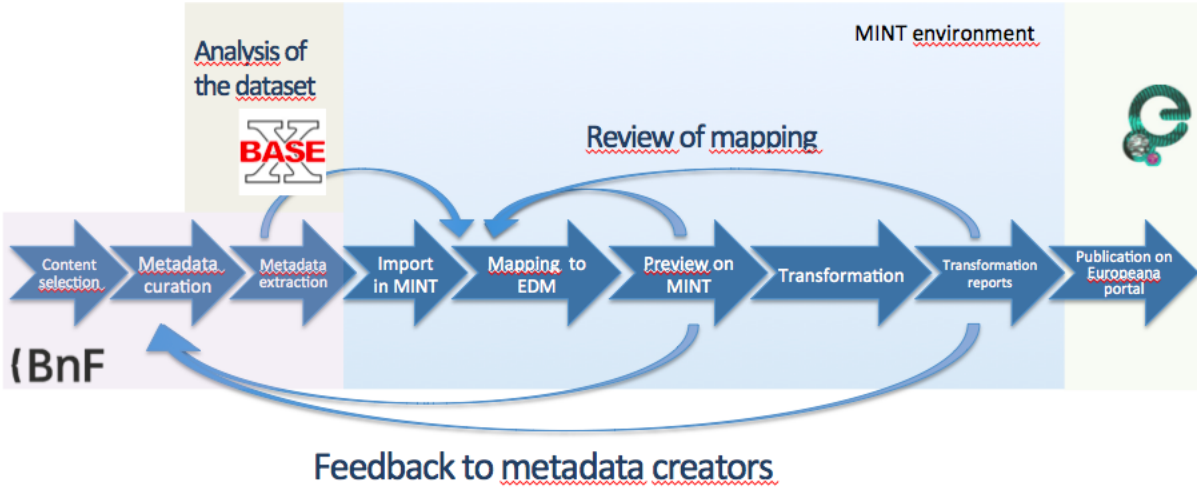

Figure 2. Iterative workflow for metadata provision, processing and publication

MINT relies on XSL technologies to perform the transformation: mappings developed by using the visual editor are transcribed in the XSL programming language. Applying this XSL stylesheet to the input data creates EDM records, which are validated using XML schemas and additional rules expressed in Schematron.[19] Furthermore, it allows the user to download the XSLT file and enrich it with custom functions.

The set of issues raised by the process usefully pointed out internal and external problems, namely modelling decisions on the implementation of EDM in order to express correctly the hierarchical arrangement of intellectual and physical components of the sound recordings, complexity of the BnF input data, as well as cataloguing and digitization practices which could be reconsidered.

5.2 Target model - EDM

As the EDM Primer indicates, "EDM is not built on any particular community standard but rather adopts an open, cross-domain Semantic Web-based framework that can accommodate the range and richness of particular community standards." [20] It re-uses, as much as possible, existing classes and properties from other vocabularies (OAI-ORE, Dublin Core, SKOS, EBU core, etc.). The base model rests upon three main classes: the Aggregation (ore:Aggregation) – the set of resources about a cultural object delivered to Europeana by one provider, the Cultural Heritage Object (hereafter referred to as CHO in this article) (edm:ProvidedCHO) and the Web resource (edm:WebResource) – a digital representation of the CHO. Additionally the model proposes Contextual classes for entities such as Agents, Subjects, etc. This topic will be addressed in more detail below.

[19] http://www.schematron.com/

[20] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf

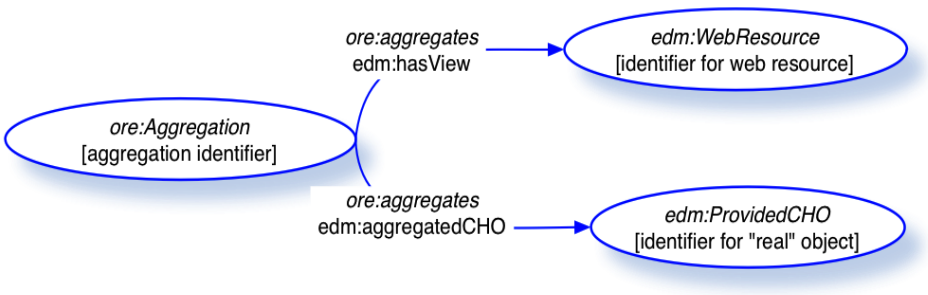

Figure 3. Core EDM classes

The first question the BnF had to solve before developing the mapping was to define the instances of these three main classes. To decide upon the identity of the provided CHO, the Europeana Sounds partners were invited to ask themselves two questions:

"Do I have enough data to represent an artifact as a cultural heritage object?

What would be useful to the end user? How do I want my collection to be best discovered?"[21]

Responses to these questions would provide guidance for accurate modelling choices.

5.3 Defining core EDM classes for the BnF dataset

The initial BnF dataset contained records describing objects composed of one or several digitized vinyls, mostly with two sides, and one or several works recorded on each of the sides. The BnF considered the object described by each bibliographic record in its Catalogue as the provided CHO. As EDM allows for a hierarchical description of CHOs with embedded CHOs corresponding to the physical parts or intellectual components of the top-level CHO, a maximalist approach could have proposed a complex hierarchical description of both kinds of components.

Nevertheless, unlike METS, EDM does not provide a mechanism for expressing different "structural maps" when the physical and the logical structure of the object are not matching – classical music works recorded in multiple tracks, spanning over several sides or disks for example – and pointing to each file from one division of the structure. Describing the internal structure of a CHO with nested CHOs had also consequences on the user experience, as the Europeana portal is currently showing different CHOs as separate results. A three-disk album, in this case, would generate four results in Europeana, one for the album as a whole and one for each of its disks, with dcterms:hasPart and edm:isNextInSequence relationships between them. As no particularly relevant descriptive metadata could be attached to a CHO representing a disk, this option was not chosen. The other option would have been to create lower-level CHOs for every work contained in the top-level CHO. This option would have increased the discoverability of sound recordings contained in albums. For example, a user interested in older versions of national anthems could access directly the recordings of "La Marseillaise" and "La Brabançonne" without having to browse through disks and tracks. Unfortunately, as described above, such information was sometimes supplied in a structured way, sometimes in a textual contents note, and above all, no information in the source dataset could bind the low-level CHO to its digital representation in the corresponding Web resources. Tables of content produced over the process of digitization

[21] http://fr.slideshare.net/CecileDevarenne/edm-for-sounds

by the BnF service providers could not be used either, as they are stored separately from the Catalogue descriptive information. This gap between bibliographic description and digital representation led the BnF to a fallback modelling choice: creating no intermediary CHOs for each and every separate recording contained in the cultural object, such as music pieces, for example.

Every ore:Aggregation binding together the description of an edm:ProvidedCHO and the edm:WebResource(s) representing it needs also to contain "the URL of a web view of the object" (via the property "edm:isShownBy") in order to provide the user with a direct access on the Europeana portal, and/or "the URL of a Web view of the object in full information context", i.e. a pointer to the object within its own institutional context (via the property "edm:isShownAt"). Each of these elements must appear at most once. The BnF preferred to indicate with edm:isShownBy the first sound WebResource, as, unlike the misinterpretation the BnF team made at first, edm:isShownBy does not aim at giving a full overview of the CHO but, in the case of a CHO represented by a series of Web Resources, only the most significant or, failing that, the first of a series of WebResources attached to the same CHO. This question was even more critical for the Web Resource which the edm:isShownAt property points to, as in the BnF's digital library, Gallica, the CHO has no unique landing page for all digital items in the case of an album constituted of several disks. As the property must appear only once for each aggregation, the dilemma was, to give access to the first digital item on Gallica or to the Catalogue record on the web, which gathers all descriptive information and access to all digital items of the cultural object. The digital library was chosen as the access point for the edm:isShownAt property instead of the Catalogue as it favoured a direct, though incomplete, access from Europeana results list to digital content on Gallica, avoiding thus an additional and possibly discouraging step between the user and its final goal, that is, the digital content itself.

One of the conclusions of this modelling step was that EDM classes are designed for a specific usage of metadata aggregation and display which are not described in the EDM documentation on purpose, as they can change over time. Only after having made the first transformation and having looked at the result on the Europeana portal could we deduce the meaning of such EDM properties. Note that some clarification on the EDM elements is likely to come from the Europeana Data Quality Committee.[22]

5.4 Generating EDM Contextual Classes

As described in 2.2, among the motivations for participating in Europeana Sounds project was the desire to expose and share one of the BnF's areas of expertise, that is, controlled vocabularies about agents, phonographic labels, concepts, places, works, etc. These naturally fall under the scope of EDM Contextual Classes. EDM defines the following contextual classes: edm:Agent, edm:Concept, edm:Event, edm:Place, edm:TimeSpan, edm:PhysicalThing.

The Audiovisual Department is particularly interested to highlight the BnF records that describe editorial sets, archival *fonds*, or collections as, for example, the record that describes the *fonds* of sound recordings made at the Sorbonne University between 1911 and 1914[23], which itself belongs to the large collection of Archives of Spoken Word. Recently the class edm:Collection has been introduced in EDM and the BnF will take advantage of this class for

[22] A newly created standing committee defined as a Europeana Network and EuropeanaTech Working Group, http://pro.europeana.eu/page/data-quality-committee

[23] See the catalogue record for this *fonds* at http://catalogue.bnf.fr/ark:/12148/cb412948151

this purpose. This will enable to group together in a meaningful way records that otherwise would be scattered in the large Europeana dataset without connection to one another.

As described in part 4 of the present article, INTERMARC allows for dynamic links via identifiers between, on one hand, the bibliographic records for cultural objects and, on the other, the authority records or *fonds*, collections and editorial sets records. Moreover, as ARK URIs are implemented for the identification of each and every record, the necessary conditions are fulfilled for the creation of EDM contextual classes as needed.

While it was obvious that the categories Person and Corporate body would fall under the scope of the class edm:Agent, modelling options were less obvious for other categories of records. In agreement with the Audiovisual Department, phonographic labels are modelled as edm:Concept, and so are the Uniform Titles [24] as well. Another decision related to distinguishing among the different entities that are represented by the subject headings (in 6XX MARC fields) and to map each of them with the appropriate contextual class. In this way, Topical subject headings are modelled as edm:Concept, personal and corporate names as edm:Agent, and geographical names as edm:Place.

In the input data, each InterXMarc expression of the full bibliographic record describing a cultural object is enhanced with information extracted from the authority records and from the records for *fonds*, collections or editorial sets related to the cultural object. For each and every linked record, the InterXMarc enhanced record includes the identifier for the authority record (or the *fonds*, editorial set or collection record), the authorized access point for the given entity, and all the variant access points (other forms of the name). In addition, for persons or corporate bodies, the role of the agent with respect to the cultural object is given in a coded form, as well as the ISNI identifier, whenever it is present in the original catalogue record. These metadata enable at least the creation of EDM contextual classes with ARK identifiers as pointers to the source records and a minimum of properties such as skos:prefLabel, dc:identifier and, as appropriate, rda:Gr2dateOfBirth, rda:Gr2dateOfDeath. Although EDM provides many other properties that could accommodate extensive information on these entities, there are reasons for which the BnF has not provided more information. Within the framework of the Europeana Sounds project, the InterXMarc enhanced records that serve as input data contain only a limited amount of metadata extracted from the related authority records. Notwithstanding, had the input data contained more information from the authority records, the following question would have however arisen: is it good practice to deliver to Europeana full authoritative information on contextual entities, thus replicating the BnF authority data information in the Europeana dataset? After discussion, it was deemed as better practice, in the context of Linked Open Data, to provide, for each of the instances form any of the controlled vocabularies, the persistent URIs accompanied by only minimal identification information.

[24]This decision was taken relying on the conclusions of "Final Report on EDM – FRBRoo Application Profile Task Force": http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/EDM_FRBRoo//TaskfoApplication%20Profile%20EDM-FRBRoo.pdf, reflected also in the "EDM profile for Sound" http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/EDMSound//TF_Report_EDM_Profile_Sound_301214.pdf

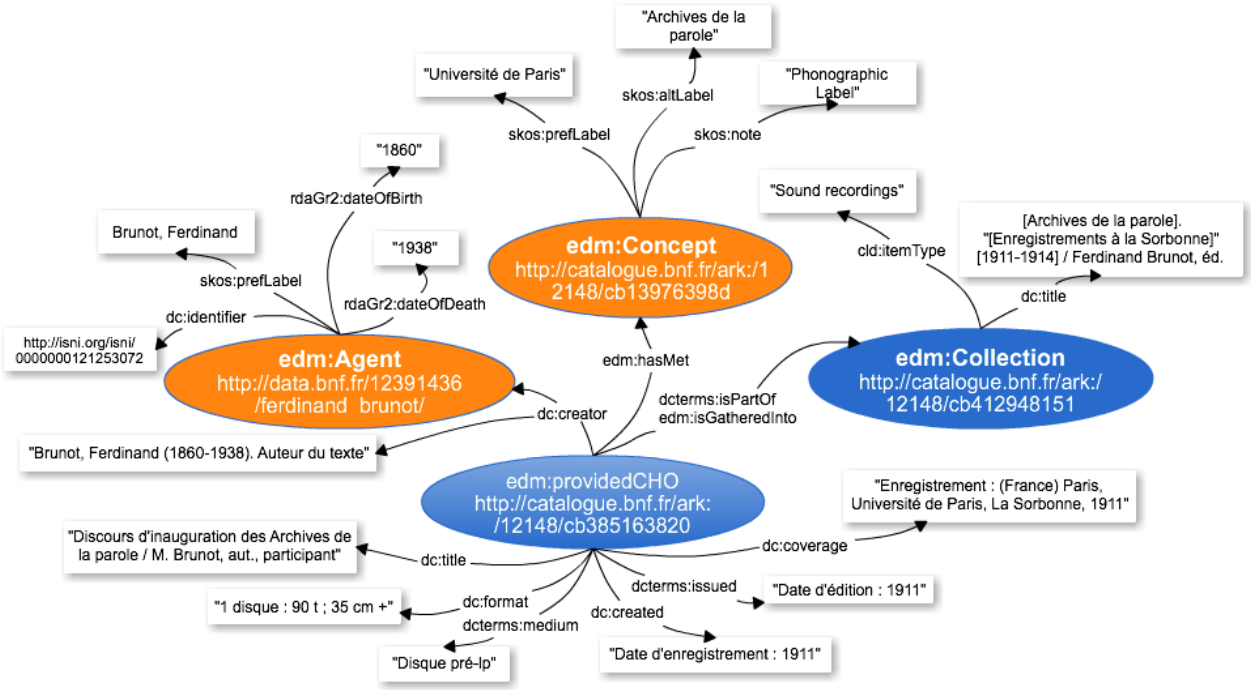

Figure 4. Modelling contextual classes

One issue that remains still unresolved in EDM is that of modelling the "Role" of an Agent associated to a cultural object (author, composer, performer, etc.). Currently EDM does not provide the mechanism for establishing a semantic relationship between an edm:ProvidedCHO, an edm:Agent and the Role this agent performs with respect to the cultural object. Discussions are underway within the Europeana Data Quality Committee, and developments are expected in the near future.

5.5 Choice of URIs for instances from controlled vocabularies

Regarding the URIs, there are still some open discussions within the project team about which ARK URIs to provide for any of the instances from the controlled vocabularies: the URIs for the authority records from the BnF general catalogue[25], or those from data.bnf.fr[26], the BnF Linked Open Data service. Data.bnf.fr is fundamentally entity oriented. For this it fully relies on and leverages the existing catalogue authority records for persons, corporate bodies, works, concepts (or themes), geographical entities, etc. [27] While catalogue authority records present the advantage of containing the most complete and up-to-date information, their URIs point to the html pages (the online publication of the BnF catalogue records). Data.bnf.fr URIs, on the other hand, identify entities and point to machine-processable data and are thus fit for Linked Open Data applications. As EDM is RDF-based, it would be preferable to provide URIs that identify machine-processable data rather than html pages. Yet, as authority records from the BnF catalogue should satisfy some quality conditions to be fed into data.bnf.fr, currently not all catalogue records are present in the data.bnf.fr dataset.

[25] Example of an URI for a catalogue authority record for a person: http://catalogue.bnf.fr/ark:/12148/cb123914365

[26] Example of an URI for the same person on the data.bnf.ft dataset http://data/bnf.fr/12391436/ferdinand_brunot/

[27] For information about data.bnf.fr and the Semantic Web see http://data.bnf.fr/en/semanticweb

Another reason for discussing the catalogue URIs *versus* data.bnf.fr URIs is also the, until recently, rather low frequency of refreshment of data.bnf.fr dataset with information from the source catalogue. This should be substantially improved in the very near future, the target objective being to reach a daily update. Consequently, for experimental purposes and to comply, as much as possible, with Linked Open Data best practices, the project team decided to explore the data.bnf.fr-URIs-path for the types of entities that are represented in the latter. For this, some side analysis is necessary to check whether all BnF catalogue authority records related to the input dataset are represented in data.bnf.fr, and to spot those that are not. A combination of tools will be used, including the aforementioned tool BaseX, as well as a tool called "Robot-Données"[28], developed by the BnF to make the catalogue, in turn, benefit from data processed for the needs of data.bnf.fr. The spotted records will be completed/corrected by the curators/cataloguers of the Audiovisual Department so that qualitative criteria are satisfied for the generation of data.bnf.fr entities. This exploratory work had multiple benefits: it was viewed both from the project team and from the data.bnf.fr team as an opportunity to upgrade more authority records from the catalogue and feed them into data.bnf.fr. It also presented an additional use case for the "Robot-Données".

5.6 Handling a highly-structured source metadata format

INTERMARC fine-grained structuration was another tricky issue. MARC-like structured formats imply a complex system of concatenation to map into less structured formats like EDM in order to make it readable by users. Therefore, in case of fields with hardly predictable combination of optional subfields, the mapping had to foresee the most commonly used sequence of subfields to generate ISBD punctuation. A thorough analysis of our dataset with XML databases management software such as BaseX[29] was therefore the starting point for the treatment of such fields.

One of the major problems encountered during the elaboration of the EDM transformation was that INTERMARC relies on the use of subfields in order to express information specific to a part of the resource. For example, the INTERMARC field for title proper may aggregate the title proper, other titles from the same author, and titles from different authors. Corresponding statements of responsibility are only related to the preceding title. The example below:

(INTERMARC)
245 1. $a La |Marseillaise **$d** Enregistrement sonore **$f** Rouget de l'Isle, comp. **$c** La Brabançonne **$f** F. Van Campenhout, comp. **$j** M. Jean Noté, baryton de l'Opéra de Paris **$j** acc. d'orch., sous la dir. de Mademoiselle Bryant

(InterXMarc)

```
<datafield tag="245" ind1="1" ind2=" ">
    <subfield code="a" Barre="3">La Marseillaise</subfield>
    <subfield code="d">Enregistrement sonore</subfield>
    <subfield code="f">Rouget de l'Isle, comp.</subfield>
    <subfield code="c">La Brabançonne</subfield>
    <subfield code="f">F. Van Campenhout, comp.</subfield>
    <subfield code="j">M. Jean Noté, baryton de l'Opéra de Paris</subfield>
```

[28] "Data Robot"
[29] http://basex.org/

<subfield code="j">acc. d’orch., sous la dir. de Mademoiselle Bryant</subfield>
</datafield>

produced the following result in EDM when mapped with MINT:

<dc:title xml:lang="fr">La Marseillaise ; . La Brabançonne : . , / M. Jean Noté, baryton de l'Opéra de Parisacc. d'orch., sous la dir. de Mademoiselle Bryant / Rouget de l'Isle, comp.F. Van Campenhout, comp. ; </dc:title>

Otherwise stated, in the very common case of several recordings in the same album, there was no way in MINT to associate each separate title with its corresponding responsibility statement(s) – “La Marseillaise” with Rouget de l’Isle and “La Brabançonne” with Theo Van Campenhout, in this case. As XSL is a declarative language rather than a functional programming language, it is particularly complex and unnatural to map element contents depending on their position. MINT, quite logically, could not provide such mechanisms. In order to address these specific issues, the BnF chose to entrust the mapping development in MINT, whenever possible, to a metadata librarian, whereas for specific needs uncovered by MINT, help from an XSLT expert was requested. Before publication, the XSL produced by the mapping elaborated in MINT was downloaded and enriched with bits of code, then uploaded again in MINT.

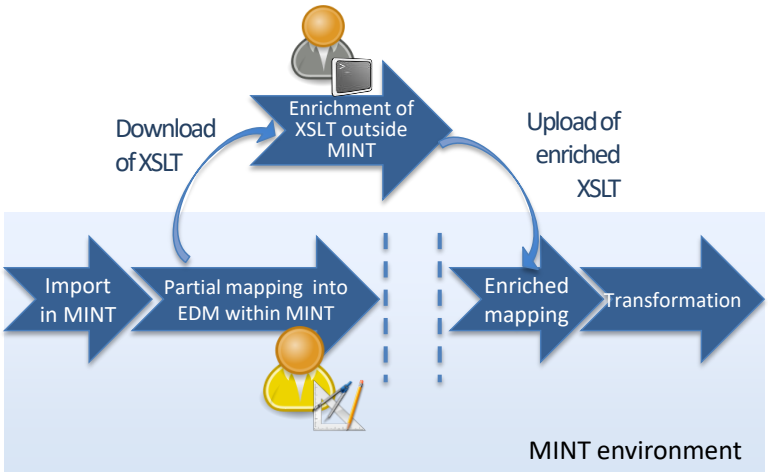

Figure 5. Mapping enrichment

The expected result below was obtained thanks to this method.

<dc:title>La Marseillaise / Rouget de l'Isle, comp.</dc:title>
<dc:title>La Brabançonne / F. Van Campenhout, comp. ; M. Jean Noté, baryton de l'Opéra de Paris ; acc. d'orch., sous la dir. de Mademoiselle Bryant</dc:title>

This complexity made also evident some flaws in the INTERMARC format structure and taught a useful lesson to agents responsible for its evolution.

5.7 Evolving cataloguing and digitization practices

Though the source dataset was not of large dimensions (5,000 records at most), the BnF had to face some heterogeneity in cataloguing and digitization practices, implying

troublesome mapping adaptations when the useful information was lacking. One example is the absence of some very basic technical metadata in the information about holdings, like the number of sound files generated by the digitization process. Actually the technical description of the digital representation is managed within applications for digitization services or digital preservation, that are separate from the catalogue and do not communicate with it. The initial dataset was mainly constituted of double-sided disks whose digitization produced, predictably enough, two digital sound files and two image files (reproducing disk labels) for each disk. But in the occasional case of single-sided disks, the BnF project team has to deduce the number of Web resources to be created from the textual analysis of descriptive uncontrolled fields.

The experience with Europeana Sounds data transformation was an opportunity to draw the attention of the departments that manage and catalogue collections on the importance of INTERMARC coded information when analysing or transforming metadata. Carefully informing the MARC coded fields and sub-fields is indeed a trying and time consuming activity in the day-to-day work. Consequently, such departments deplore the amount of time spent to create them but are unaware of their final use. The usefulness of coded information is not directly demonstrated to cataloguers and cannot be queried via the Catalogue production tool. Such information is, moreover, invisible to end-users. Hence, this aspect of the cataloguing is at times neglected, at times misinterpreted, without mentioning layers of quality in the catalogue due to changes and/or evolution of cataloguing practices over years. On the contrary, when analysing a metadata set to be corrected or transformed, coded information is a great asset if consistently and homogeneously filled in. Metadata become more predictable and can be successfully interpreted by automated processing methods.

As a matter of fact, coded information is critical to feed the mandatory EDM properties, such as genre of content. If all necessary coded information is consistently informed, metadata become more predictable and can be successfully interpreted by automated processing methods. Actually, the first steps of the experience for transformation to EDM showed that many records were rejected for lack of such information in the source data. Real time transformation validation reports provided by MINT enabled the metadata expert in charge of the transformation to give immediate feedback to the Audiovisual Department. Metadata was fixed by filling in correctly the coded information at the source and reprocessing was then made possible.

5.8 Experimenting other transformation methods

The experience with Europeana Sounds showed that the Catalogue descriptive metadata, though integrating holdings metadata about digital items, do not satisfy all needs of the target EDM. Metadata about Web resources are stored outside of the Catalogue. These will be vital if the Europeana Sounds experience should be extended to the entire set of BnF digitized materials that is characterized by a great variety of materials and formats. Reliable information about the number of Web Resources and their file format to be declared in EDM aggregations are stored in the BnF's digital repositories which will have to be queried to provide Europeana with this information.

Yet the BnF has developed two major tools for publishing data according to Linked Open Data principles: data.bnf.fr[30] for descriptive metadata coming from the Catalogue and SPAR[31] for digital data preserved by the BnF. Though the current scope of these tools is not

[30] http://data.bnf.fr/en/
[31] http://www.bnf.fr/fr/professionnels/spar_systeme_preservation_numerique.html

covering all metadata stored in the Catalogue, nor all digital data stored by the BnF, their scope is gradually extended and should be almost complete in the coming years. Both have a SPARQL endpoint to query RDF data. While access is public for data.bnf.fr[32], for SPAR access is restricted only to the BnF agents and can be operated only within the BnF environment.

A simple experiment has shown that a cross-query on the two SPARQL services can provide both descriptive and technical metadata, taking down the data silos borders. Indeed, a query brings back metadata coming originally from authority records (agents, topics, places, timespans, etc.) and from technical properties of the digital representation. Moreover, a CONSTRUCT query can easily create RDF/XML graphs that meet EDM requirements. No XSLT tool is thus needed as the CONSTRUCT SPARQL query type naturally creates new graphs with the element names provided in the query.

For example, such a query in http://data.bnf.fr/sparql/

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX edm: <http://www.europeana.eu/schemas/edm/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ore: <http://www.openarchives.org/ore/terms/>
PREFIX sparcontext: <info:bnf/spar/context#>
PREFIX sparstructure: <info:bnf/spar/structure#>

CONSTRUCT {<http://data.bnf.fr/ark:/12148/cb30373095k#aggregation> a ore:Aggregation;
edm:isShownAt ?Gallica.
  <http://data.bnf.fr/ark:/12148/cb30373095k> a edm:ProvidedCHO; dc:title ?titre;
dcterms:extent ?description; dcterms:issued ?date; dc:publisher ?editeur .
?object a edm:WebResource.}
WHERE {
 {
 <http://data.bnf.fr/ark:/12148/cb30373095k> dcterms:title ?titre; dcterms:description
?description; dcterms:date ?date; dcterms:publisher ?editeur; rdarelationships:workManifested
?oeuvre; rdarelationships:electronicReproduction ?Gallica.
  BIND(IRI(STRAFTER(STR(?Gallica), "http://gallica.bnf.fr/")) AS ?docnum )
 }
        Service <http://srv-db-dmcons.spar.bnf.fr:8997/sparql>
  {
  ?docnum sparcontext:hasLastVersion ?v.
   ?v sparcontext:hasLastRelease ?r.
  ?r a sparstructure:group ; ore:aggregates ?object. ?object ore:isAggregatedBy ?stMap.
?stMap dc:type "physical".
 }
 }
```

creates the following RDF/XML result:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:ns4="http://www.europeana.eu/schemas/edm/" >
<rdf:Description rdf:about="http://data.bnf.fr/ark:/12148/cb30373095k">
 <rdf:type rdf:resource="http://www.europeana.eu/schemas/edm/ProvidedCHO" />
 <dcterms:extent>In-8° , 523 p., fig. et pl., portrait de Dumas</dcterms:extent>
```

[32] http://data.bnf.fr/sparql/

```
  <dcterms:issued>1846</dcterms:issued>
  <dc:title>Les Trois mousquetaires, par M. Alexandre Dumas</dc:title>
  <dc:publisher>Paris : J.-B. Fellens et L.-P. Dufour , 1846</dc:publisher>
 </rdf:Description>
 <rdf:Description rdf:about="ark:/12148/bpt6k61336787/f1.version0.release0">
  <rdf:type rdf:resource="http://www.europeana.eu/schemas/edm/WebResource" />
 </rdf:Description>
 <rdf:Description rdf:about="ark:/12148/bpt6k61336787/f10.version0.release0">
  <rdf:type rdf:resource="http://www.europeana.eu/schemas/edm/WebResource" />
 </rdf:Description>
       …
```

The experiment showed how metadata librarians could create EDM modeled metadata by only executing SPARQL cross-repositories queries. This is one among many use cases of using Linked Open Data technologies for publishing not only descriptive metadata, but also technical metadata about the digital representations. It reinforces the BnF's belief that the use of RDF technologies in its digital repository to allow for professional exploitation of technical and administrative metadata about the digital holdings opens new, unprecedented possibilities for a variety of projects of data reuse.

6 Leveraging skills, developing new skills, sharing work, interacting

A remarkable feature of the experience with Europeana Sounds is that the core work for metadata transformation is not performed by the IT developers but by metadata librarians. These include experts with good knowledge of the source data, format and cataloguing practices, and others who, in addition, have technical skills and experience with XSLT transformation technologies as well as with RDF.

The first category of metadata librarians had experience developing mapping specifications only from the intellectual perspective, which were then handed to IT technical staff for implementation. The user-friendly visual mapping facility within MINT, which uses a drag and drop function from the source format into EDM, introduced this category of professionals to automated processing of structured metadata. Other functionalities, such as conditional mapping and concatenation, further enhanced the technical understanding of metadata transformation and analysis and, thereby, improved the mapping challenge.

This experience was thus beneficial and educational in many respects. As both the intellectual and the technical aspect of the process are combined in one, the expert views in real time the data being transformed. In addition, the metadata validation facility provided by MINT gives the possibility to immediately analyse whether there are issues with the source metadata or with the mapping. This triggers two kinds of actions, fixing and refining the mapping (by the same metadata expert) and connecting with the Audiovisual Department curators and cataloguers to get the metadata fixed at the source in order to reimport them in MINT.

The second category of metadata librarians, due to their involvement in other projects in the library that require technical skills, such as digital preservation or developing Linked Data applications, dealt with enrichment of the mapping by using XSLT technologies at a higher level of complexity.

Interaction of both categories of metadata librarians was extremely beneficial as it created a new dynamics of collaboration and of sharing expertise.

The other immediate benefit of the multipart collaboration process is that all of the actors are getting a better understanding of the impact of metadata quality for future reuse.

This is also a way to increase awareness about the value of the cataloguing work. Descriptive metadata about the sound cultural heritage objects are initially created in the BnF General Catalog by cataloguers from the Sound section of the BnF Audiovisual Department. As with all metadata created by cataloguers in the library, from an overall perspective, these metadata constitute an invaluable asset, as they are the basis for multiple uses for different purposes, including publication of catalogues on the Web, metadata supply to library patrons and research projects, feeding descriptive data about the cultural heritage objects into the digital library applications, development of Linked Data services, and other future unpredictable uses and reuses. Creating metadata of high quality and as expressive as possible is *sine qua non* for libraries to meet the expectations of the future; it becomes a highly strategic mission.

Beyond requiring knowledge of the domain and of the collections described, creating high quality, reusable and machine processable descriptive metadata also requires the librarian's intellectual and technical skills. These traditionally include cataloguing mastery, good knowledge and consistent use of the encoding format used in production. Large institutions, traditionally, are characterized by a compartmentalization of work according to skills and expertise. Cataloguing is being performed by teams of librarians within departments that manage collections, instructions on the expected result is provided by metadata experts and automated processing is being performed by IT staff, without close collaboration among the ones and the others. The experience with the Europeana Sounds project showed that even with user-friendly tools such as MINT close collaboration between source data specialists and experts with data reprocessing techniques is indispensable.

7 Conclusion

Among other similar projects led by the BnF, participation to Europeana Sounds showed clearly that, even if different silos are used to store descriptive metadata and digitized or born-digital data, a bridge between those sources is necessary to provide high-quality descriptive and technical metadata. This bridge will become more and more important as institutions are exchanging metadata about a cultural object but also the corresponding digital data.

The rapid, ongoing evolution of the technological environment opens ever increasing, most possibly, unpredictable opportunities. This highlights the need for cataloguers to get aware that the act of creating high quality, expressive metadata is fundamental, as this is the starting point of a long journey for the metadata to be reused, reprocessed and circulate in a global environment.

Acknowledgments

The authors are grateful to all the Europeana Sounds project team at the BnF, who have provided valuable advice and contribution for this article. In the first place we would like to give special thanks to the curators of the Audiovisual Department, Pascal Cordereix (head of the sound recordings service) and Lionel Michaux (curator of historical sound recordings). Our thanks go to Laure Livet, the metadata librarian in charge of providing metadata to Europeana. We also wish to thank Anne-Claire Rebours, for her legal advice, and all our colleagues from the BnF Delegation for International Affairs, Marion Ansel, coordinator of the BnF Europeana Sounds project team, Axelle Bergeret-Cassagne and Elisabeth Freyre. And finally we thank the other colleagues from the Audiovisual Department, Audrey Viault, Cécile Kattnig and Marie-Pierre Bodez.