Integrating library and cultural heritage data models: the BIBFRAME - EDM case

Sofia Zapounidou, Michalis Sfakakis, Christos Papatheodorou

Department of Archives, Library Science and Museology, Ionian University, 72 Ioannou Theotoki str., 49100 Corfu Greece

ABSTRACT

Libraries create and preserve bibliographic data using the MARC family of standards to encode and interchange them. Aggregation and exposure of these data into the Semantic Web universe is a key issue in libraries and is approached on the basis of library data conceptual models. Examining the way that data are represented in each data model, as well as possible mappings between different data models is an important step towards interoperability. This paper aims to contribute to the desired interoperability by attempting to map core classes and properties between two well known conceptual models, namely BIBFRAME and EDM. BIBFRAME aims to transform the widely used MARC data structure in libraries to the Linked Data context and EDM is the model developed and used in the Europeana Cultural Heritage aggregation portal.

Categories and Subject Descriptors

D.2.12 Interoperability

General Terms

Design, Standardization, Theory

Keywords

Conceptual models, linked data, interoperability, data integration, BIBFRAME, EDM

1. INTRODUCTION

Web scale discovery services, aggregation portals and linked data offer memory organizations, such as museums, libraries and archives, the opportunity to enhance the impact of their collections, as well as provide them new ways to fulfill their role as major contributors in research, teaching and learning. This paper focuses on libraries and investigates the integration of their data with third party services and their reuse in new contexts.

Libraries host a variety of materials and traditionally use many metadata formats. Aggregation or harvesting of these metadata presupposes that library metadata are interoperable. In the linked data environment there is the apparent need the metadata (i) to be expressed by common vocabularies and (ii) their semantics to be harmonized with shared and commonly accepted conceptual models. There is a number of initiatives regarding the publication of library data as Linked Data. Each initiative developed its own interpretation of how the library data may be integrated into the semantic web, providing its own conceptual model. The most known of them are FRBR [8], FRBRoo [2] and BIBFRAME [12]. However, these different views cause interoperability problems and prevent data integration and/or aggregation.

In the cultural heritage domain there have been developed aggregation services that collect from libraries and other memory institutions metadata about cultural heritage objects with the aim to provide advanced research support services. There are domain-specific aggregation services, as well as national and transnational ones. The most well known are the panEuropean aggregation portal of Europeana (http://www.europeana.eu/) and the Digital Public Library of America – DPLA (http://dp.la/). Both Europeana and DPLA have developed data models, namely Europeana Data model (EDM) [9] and DPLA Metadata Application Profile - DPLA MAP [5], to enable proper harvesting of metadata from a variety of data providers.

Interoperability of library data for successful integration in third-party systems or aggregation by third-party services is a major research issue. This paper aims to contribute to interoperability of library data by examining how BIBFRAME [12] data could be integrated in the Europeana aggregation portal. BIBFRAME is a new library data model currently being developed by the Library of Congress with the aim to “translate MARC 21 to a Linked Data (LD) model” [11, 12]. The Europeana portal aggregates digitized Cultural Heritage Objects (CHOs) by European Libraries and other cultural institutions. These CHOs are described with the Europeana Data Model (EDM) [9].

In the next section the BIBFRAME and EDM conceptual models are briefly presented, while section 3 describes the methodology followed for the proposed mapping and provides a test case, consisting of seven library records, which demonstrates the complexity of linking library data. Section 4 presents the proposed mapping between the two models and Section 5 discusses and concludes the derived results.

2. BACKGROUND

Libraries use record-based descriptions about library objects. These record-based descriptions are used to find and access the described physical library objects. Even though they focus on the item at hand, they provide information, either implicitly or explicitly, regarding both the intellectual content (work) contained in the physical library object and the library object itself. Moreover, other bibliographic details related to the production process such as publisher and edition, handling of the item such as reproduction, as well as relationships between and among different bibliographic entities (contributors, intellectual works, subjects, etc) [1, 2] are also included in the descriptions. Until recently, all this information is encoded and exchanged according to the MARC family of standards [11].

BIBFRAME as transition model from the currently used MARC records to the linked data model does not adopt current bibliographic records' flat structure and uses separate entities (classes) and properties to describe library objects, their characteristics and the relationships between them.

In particular, BIBFRAME is a model under development by the Library of Congress. Its main classes are: *Creative Work, Instance, Authority and Annotation* [12]. The class *Creative Work (*or simply *Work)* reflects the "conceptual essence of the cataloguing item" [12]. The class *Instance* reflects "an individual, material embodiment of the Work". The class *Authority* is used to identify *People, Places, and Organizations* involved in the creation or publication of a *Work*. For the expression of topics, BIBFRAME *Authority* simply works as a linking mechanism to LC Subject Headings published as linked data at the ID.LOC.GOV site. The class *Annotation* expresses comments made about a BIBFRAME *Work, Instance,* or *Authority*. Examples of BIBFRAME annotations are: library holdings, cover arts, sample texts, reviews, etc. (see Figure 1).

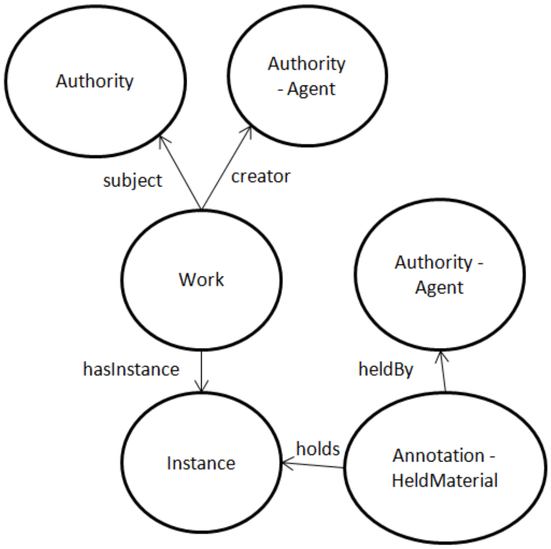

Figure 1: BIBFRAME model with Annotation for holding [8]

Europeana aggregates metadata about and enables access to born-digital or digitized cultural heritage content provided by European memory institutions. Descriptions over Europeana are made with Europeana Semantic Elements [7], a basic data model that uses Dublin Core's 15 elements and other 12 additional elements. The Europeana Data Model (EDM) [9] has been developed for the better semantic expression of the cultural heritage descriptions that Europeana data providers contribute. No community–driven standard was used as a basis for its development and the Semantic Web framework was taken into account [9].

EDM's scope is diverse than BIBFRAME's; thus different semantics and abstraction layers are used. For each provider, EDM distinguishes between real provided cultural heritage objects and their digital representations, and between provided cultural heritage objects and their descriptions. It is worth mentioning that Europeana collects only descriptions for objects having at least one web representation [9]. As depicted in figure 2, EDM provides three core classes, namely *edm:providedCHO* (for provided Cultural Heritage Object), *edm:webResource* (for the *edm:providedCHO* digital representations) and *ore:aggregation* (for the aggregation of the activities made by the provider of the *edm:providedCHO*).

The alignment of EDM to library metadata is a work in progress. The library metadata alignment report published in 2012 [1] mainly takes into consideration FRBR semantics [8], focuses on specific library materials (monographs, multi-volume works and serials), does not adopt current bibliographic records' flat structure and adheres to linked data principles. A key point for the development of the report was the separation of the item in hand (e.g. the book) from its edition which represents the entirety of all identical copies of the item in hand. Therefore abstract levels have been defined to "differentiate between:

- the description of the information (the entirety of all identical copies of a book) and the information carrier (the book in the shelf)
- the description of the real world object (the book) and its digital representation (a digital copy of this book)
- the description of the object described (the book) and the object describing it (the metadata)."

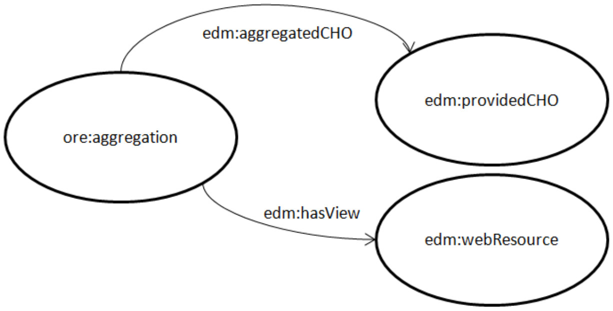

Figure 2: Europeana Data Model [9]

While the EDM library data alignment report [1] has considered the concepts of the FRBR model and the compliance of the report with the FRBR was recognized, in the framework of this report compliance with FRBR was not achieved and the introduced concept of 'edition' represents the union of the FRBR Work, Expression and Manifestation entities. According to the report the 'edition' level information of the resource is represented by the *edm:providedCHO* class, while the digital representation of the real world object is represented by the *edm:webResource* class. The *ore:aggregation* class links the description of the provided resource with its digital representations.

The issue of transforming BIBFRAME data into EDM respecting the above framework is a key issue in examining interoperability between the two models.

3. METHODOLOGY - REQUIREMENTS

The methodology adopted in this work is a combination of the ones used in the Europeana Libraries project [1] for the alignment of library metadata with the Europeana Data Model and the EDM – FRBRoo Application Profile Task Force [6]:

1. Selection of specific type(s) of library material
2. Definition of requirements for a BIBFRAME – EDM profile
3. Selection of a real test case and bibliographic records
4. Representation of the test case in BIBFRAME
5. Attempt for a BIBFRAME – EDM profile following a path-oriented approach [10, 13]
6. Transformation of BIBFRAME representation in EDM following the library data alignment report [1]

Regarding the types of the library material, since library collections consist mainly of monographs, this paper focuses on monographs and multivolume works. Thus we define the following requirements for the BIBFRAME – EDM profile:

- Europeana is an important aggregator in the cultural domain. The European Library is the domain aggregator for libraries to Europeana.
- The selected EDM classes and properties will be used according to the Europeana Data Model for Libraries definitions [1].
- BIBFRAME is a linked data model. Therefore the BIBFRAME-EDM profile shall use Resource Description Framework (RDF) syntax and shall support the use of URIs.
- The BIBFRAME-EDM harmonization profile shall be flexible enough to enable meaningful representations for other types of library material.

The test case selected is Cervantes' "Don Quixote" because it provides the ability to build complex representations in BIBFRAME and to test how well these complex representations may be expressed by EDM. "Don Quixote" consists of two separate works: the first one was published in 1605 with the title "El ingenioso hidalgo don Quixote de la Mancha" and the second one was published in 1615 with the title "Segunda parte del ingenioso cauallero don Quixote de la Mancha". These two parts have been both published and translated afterwards as independent volumes, as well as in a single volume. Moreover, there are many reproductions to other materials than the original publications, as well as other works based on many variations of the original work. Seven bibliographic records from the National Library of Spain and the Library of Congress that describe (i) the first editions of the two Don Quixote's parts (denoted as 1 and 2 respectively in BIBFRAME and BIFRAME-EDM mapping representations of Figures 3 and 5), (ii) the first edition that incorporated both parts (denoted as 3), (iii) a French translation of both parts (denoted as 4), (iv) an English translation that was based on the former French one (denoted as 5), (v) an annotated edition of both parts by the Cervantes Institute (denoted as 6) and (vi) a CD-ROM (denoted as 7) that compiled the annotated edition's text with a linguistic database developed on this content. The linguistic database is denoted as 8 in BIBFRAME and EDM representations that follow. The CD-ROM is a born digital object, while the English translation is a physical object that has been digitized.

In BIBFRAME every record from our sample corresponds to a bf:Work class linked with the respective embodiment Instance (or subclass of Instance). Thus, eight individual works are generated to represent the intellectual content of the two independent volumes, the single volume publication for both parts, its translations in English and in French, the Cervantes

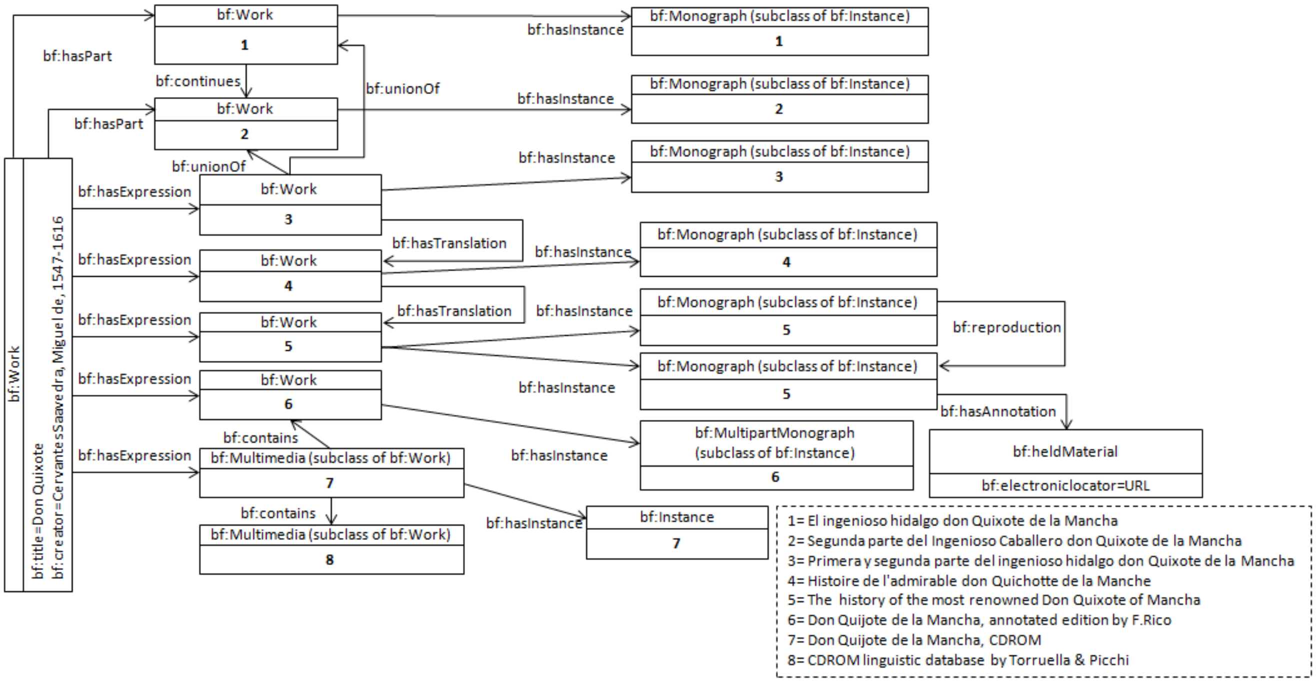
bf:Work
1
bf:hasPart
bf:continues
bf:unionOf
bf:hasInstance
bf:Monograph (subclass of bf:Instance)
1
bf:Work
2
bf:hasPart
bf:hasInstance
bf:Monograph (subclass of bf:Instance)
2
bf:unionOf
bf:hasInstance
bf:Monograph (subclass of bf:Instance)
3
bf:Work
bf:title=Don Quixote
bf:creator=Cervantes Saavedra, Miguel de, 1547-1616
bf:hasExpression
bf:Work
3
bf:hasTranslation
bf:hasInstance
bf:Monograph (subclass of bf:Instance)
4
bf:hasExpression
bf:Work
4
bf:hasTranslation
bf:hasInstance
bf:Monograph (subclass of bf:Instance)
5
bf:reproduction
bf:hasExpression
bf:Work
5
bf:Monograph (subclass of bf:Instance)
5
bf:hasExpression
bf:Work
6
bf:hasInstance
bf:hasAnnotation
bf:contains
bf:hasExpression
bf:Multimedia (subclass of bf:Work)
7
bf:hasInstance
bf:MultipartMonograph (subclass of bf:Instance)
6
bf:heldMaterial
bf:electroniclocator=URL
bf:contains
bf:Multimedia (subclass of bf:Work)
8
bf:hasInstance
bf:Instance
7
1= El ingenioso hidalgo don Quixote de la Mancha
2= Segunda parte del Ingenioso Caballero don Quixote de la Mancha
3= Primera y segunda parte del ingenioso hidalgo don Quixote de la Mancha
4= Histoire de l'admirable don Quichotte de la Manche
5= The history of the most renowned Don Quixote of Mancha
6= Don Quijote de la Mancha, annotated edition by F.Rico
7= Don Quijote de la Mancha, CDROM
8= CDROM linguistic database by Torruella & Picchi

Figure 3: BIBFRAME representation

Institute's annotated edition, the linguistic database derived from the Cervantes' annotated edition and the CD-ROM containing (bf:contains) both the Cervantes' edition with the linguistic database (Figure 3). The relationships between the parts are implemented using the partOf relation of the Work class. Moreover, an additional work class representing the dominant concept originally conceived by Cervantes is defined with its partOf relations. All instances, except the CD-ROM, are digitized and may be openly accessed online. For clarity reasons information about holdings is not given for each instance. Item-level information is only given indicatively for the English translation of Don Quixote.

4. MAPPING BIBFRAME to EDM

Mapping two different conceptual models of dissimilar semantics is a not a straight forward issue. For accomplishing the mappings between the two models we thoroughly examined the properties identifying the members of every individual class, as well as the relationships between the classes, in both models. Then we manually compared these set or properties matching the classes with the most similar intentions. For clarity reasons we have chosen a path oriented approach [10, 13] with which (a) the semantics of conceptual models is expressed as paths having the form of a sequence of "domain class – property – range class" statements and (b) the paths of the source model are mapped to semantically equivalent paths of the target model.

Our mapping takes into account the report on the alignment of library metadata with the Europeana Data Model [1]. This report uses FRBR Group 1 entities [8], namely Work, Expression, Manifestation and Item, as point of reference, does not perform for the time being a one-to-one mapping to EDM classes; it states that as far as text resources are concerned "all information concerning the Manifestation, Expression and Work entities will be added to the ProvidedCHO" class [1]. The *WebResource* class which in EDM [9] is defined as "information resource that has at least one Web Representation and at least a URI", in the framework of the library alignment report is defined as the "digital representation of an item" [1]. In BIBFRAME information regarding FRBR Works and Expressions is given through the *Creative Work* class, while information regarding FRBR Manifestations is given through the *Instance* class. Holdings (Items in FRBR) are stated through the *Annotation* class.

According to the EDM library alignment report the *edm*:*providedCHO* class is at the 'edition' level and "all information concerning the Manifestation, Expression and Work entities will be added" to it [1]. Therefore, the corresponding instance from the path "Work –hasInstance – Instance" is mapped to a single *ProvidedCHO* instance, as shown in Figure 4 and selected properties from the bf:Work and bf:Instance could be mapped to similar *ProvidedCHO* properties. The existence of a library object that is in digital form and therefore is to be aggregated by Europeana is expressed by the following path "Work –hasInstance – Instance - hasAnnotation - heldMaterial - electronicLocator – URI". The same path declares the electronic location from which the either born-digital or digitized library object is available and therefore justifies an instantiation of the *edm*:*webResource* class, with id the URI from the BIBFRAME path. It is worth mentioning that not all instances of the class *bf*:*heldMaterial* may correspond to an *edm*:*webResource* instance due to the restriction of the latest that its instances must have at least one Web Representation and at least a URI. Therefore, only *bf*:*heldMaterial* class instances having a digital representation with a URI can be members of the *edm*:*webResource* class. As far as BIBFRAME Authority class and subclasses are concerned, mapping to EDM equivalent classes was a more straightforward issue.

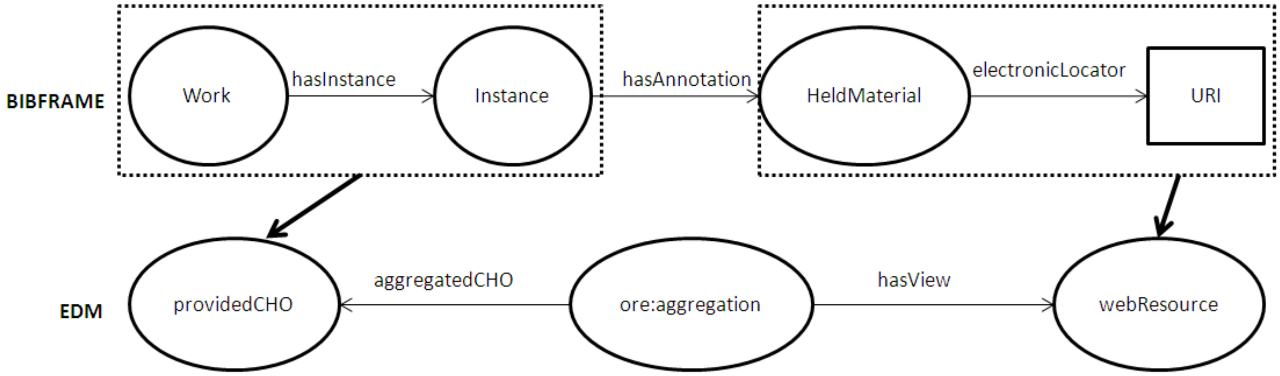

Figure 4: The mapping of the basic BIBFRAME path to core EDM classes

The mapping of BIBFRAME to EDM core properties is specified by the following path pairs. The validity of mappings requires that each BIBFRAME path ends with "HeldMaterial – electronicLocator - URI". In detail, the BIBFRAME path:

- "Work- bf:hasPart – Work" is mapped to the EDM path "ProvidedCHO - dcterms:hasPart – ProvidedCHO"
- "Work - bf:unionOf - Work" is mapped to the EDM path "ProvidedCHO - dcterms:hasPart - ProvidedCHO"
- "Work - bf:hasTranslation - Work" is mapped to the EDM path "ProvidedCHO - dcterms:hasVersion - ProvidedCHO"
- "Work - bf:contains - Work" is mapped to the EDM path "ProvidedCHO - dcterms:hasPart - ProvidedCHO"
- "Work - bf:hasExpression - Work" is mapped to the EDM path "ProvidedCHO - dcterms:hasVersion - ProvidedCHO"
- "Work - bf:continues - Work" is mapped to the EDM path "ProvidedCHO - Inverse of edm:isSuccessorOf - ProvidedCHO"

- “Work – hasInstance - Instance” triggers an instance of a single edm:providedCHO according to the EDM libraries metadata alignment report [1] (see figure 4). This single edm:providedCHO will have some properties that semantically refer to the BIBFRAME Work class and some others that semantically refer to the BIBFRAME Instance class.
- “Instance - bf:reproduction - Instance” is mapped to the EDM “ProvidedCHO – edm:isDerivativeOf - ProvidedCHO”.
- “Work - bf:subject - Authority” is mapped to the EDM path “ProvidedCHO - dc:subject - NonInformationResource”. It is reminded that subclasses of NonInformationResource are edm:agent, edm:place, edm:timeSpan and skos:concept.
- “Work - bf:creator - Agent” is mapped to the EDM path “ProvidedCHO - dc:creator – edm:agent”.

The EDM representation created according to our BIBFRAME – EDM mapping is presented in Figure 5. Following the EDM library alignment report’s [1] suggestions regarding multipart works, in this representation the dominant concept originally conceived by Cervantes and represented in BIBFRAME as *Work* with no *Instances* (see Figure 5) is expressed in EDM as a *ProvidedCHO* that has no *WebResource* of its own. Since proper representation in EDM requires a link to a Web Resource, the Don Quixote *ProvidedCHO* is linked to the *WebResource* of its first volume (denoted as 1 in Figure 5). It also must be noted that the CD-ROM and the linguistic database incorporated in it are not included in Figure 5, since the CD-ROM is not available online and there is no “Work –hasInstance – Instance - hasAnnotation - heldMaterial - electronicLocator – URI” BIBFRAME path describing it. Therefore instantiation of an *edm:webResource* class could not be justified.

5. DISCUSSION - CONCLUSIONS

The motivation of this paper was to examine how bibliographic data may be aggregated by third party services. We have focused on BIBFRAME source data and how they could be aggregated by the Europeana aggregator using the Europeana Data Model and the library data alignment report [1] in particular. Our investigation has showed that expression of the BIBFRAME conceptualization in the Europeana framework using EDM classes and properties is achievable without significant loss of semantics.

The process of mapping was a challenging one, since BIBFRAME and EDM models have different semantics and levels of conceptualization. BIBFRAME [12] defines the class *Work* for the expression of an intellectual work and the class *Instance* for the physical embodiment. The class *Annotation* serves to express items in hand. The EDM library data alignment report [1] suggests use of one class only; the *ProvidedCHO* class is considered to be at the edition level that includes information regarding the union of the intellectual work, its expression and its physical embodiment. The *WebResource* class accommodates item-related information only for born-digital or digitized items available online. Since non digital material is out of scope in Europeana, it was decided in our mapping that the existence of the following BIBFRAME path “Work – hasInstance - Instance - hasAnnotation - heldMaterial - electronicLocator – URI” justifies

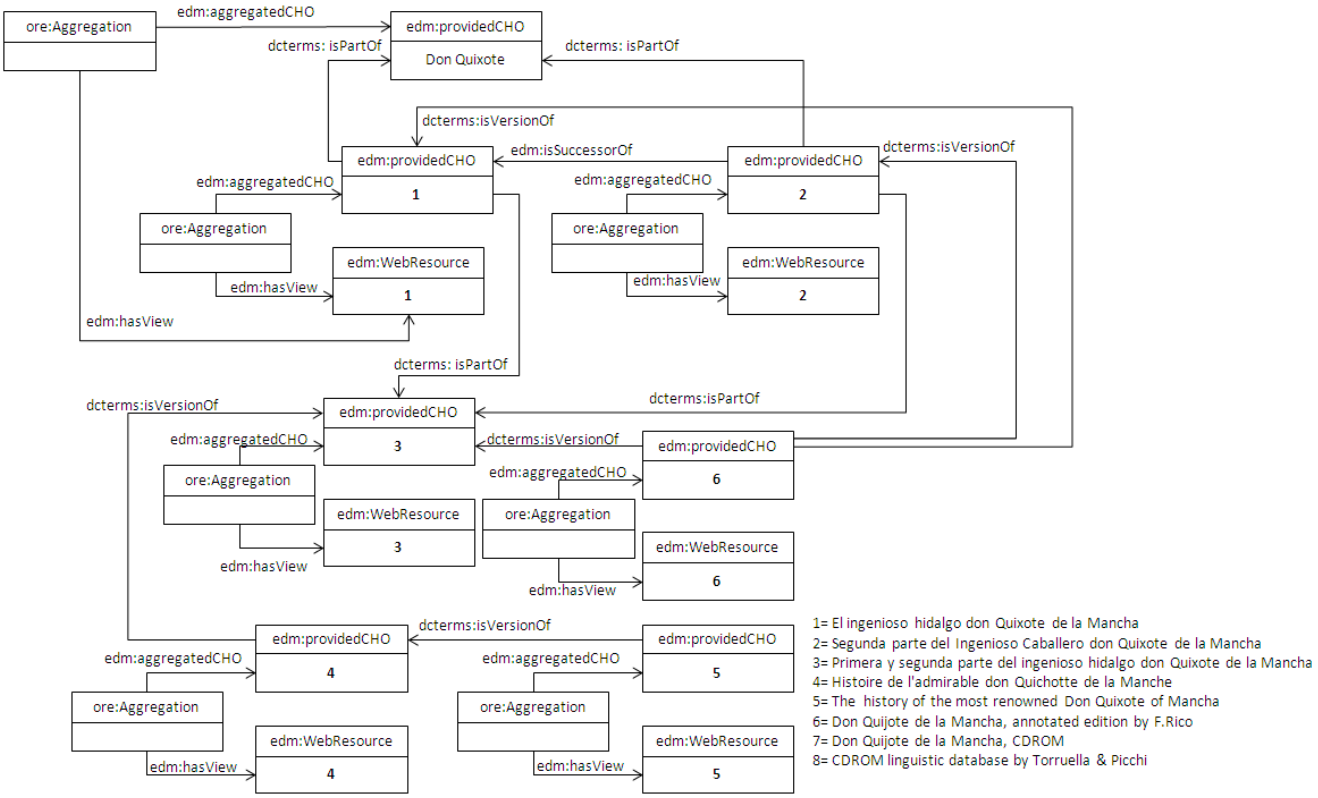

Figure 5: The EDM representation derived by BIBFRAME – EDM mapping

a) the aggregation of this Instance (and the Work whose instance it is) by Europeana, b) the mapping of the BIBFRAME path "Work – hasInstance – Instance" to one *ProvidedCHO* class in EDM and c) an instantiation of the edm:webResource class for this one and only *ProvidedCHO*.

At this point it must be noted that the adopted path – oriented approach [10, 13] enabled explicit semantic expressions and mappings between the source and the target data model, as defined in models' current specifications. This mapping will probably change as the BIBFRAME evolves. BIBFRAME is currently under development; its classes and semantics change and evolve. An indicative example that caused discussion was the case of Holding versus HeldMaterial classes. According to BIBFRAME vocabulary - List View [3], domain of the bf:electronicLocator property is bf:HeldMaterial (sub-class of Annotation), while in the BIBFRAME Annotation Model [4] there is a Holding subclass of Annotation and not a bf:HeldMaterial subclass.

Our mapping is a first attempt and future investigations regarding interoperability between BIBFRAME and EDM must be made. In these future investigations there are some issues that we have identified for further research. Even though libraries hold mostly monographs they keep other types of material too. BIBFRAME representations of other types of material and their mappings to EDM must be included in a future BIBFRAME-EDM application profile. Non digitized materials must also be taken into account. EDM is focused on digitized material, while BIBFRAME is a model for describing library materials. Therefore it must also be investigated if information regarding non digitized material should be integrated into the EDM and if yes how this may be achieved. While Europeana uses the concept of the ore:Proxy [9] in order to contextualize the ingested descriptions of the Cultural Heritage Objects, use of proxies was not discussed neither in the EDM library data alignment report [1] nor in our mapping. Europeana is an aggregator portal interested in hosting various descriptions about the same cultural heritage object without losing information about its data providers' contributions [9]. A study on how BIBFRAME data could be mapped to EDM using proxies would contribute to interoperability as well.

6. REFERENCES

[1] Angjeli, A., Bayerische, M., Chambers, S., Charles, V., Clayphan, R., Deliot, C., Eriksson, J., Freire, N., Huber, A., Jahnke, A., Pedrosa, G., Phillips, V., Pollecutt, N., Robson, G., Seidler, W., Rühle, S. 2012. *D5.1 Report on the alignment of library metadata with the European Data Model (EDM) Version 2.0*. Europeana Project.

[2] Bekiari, C., Doerr, M., Bœuf, P. Le and Riva, P. 2013. *FRBR object-oriented definition and mapping from FRBRER, FRAD and FRSAD (v.2.0)*.

[3] BIFRAME. c2014. *List View BIBFRAME Vocabulary*.

[4] Denenberg, R., Ashton, J., Boughida, K., Danskin, A., Deliot, C., Fallgren, N., Fons, T., Ford, K., Godby, J., Ogbuji, U., Guenther, R., Heuvelmann, R., McCallum,S., Miller, E., Shieh, J., Trail, N. and Wiggins, B. 2013. *BIBFRAME Annotation Model BIBFRAME Community Draft, 26 August 2013*.

[5] Digital Public Library of America. 2013. *Digital Public Library of America: Metadata Application Profile, Version 3*.

[6] Doerr, M., Gradmann, S., LeBoeuf, P., Aalberg, T., Bailly, R., and Olensky, M. 2013. *Final Report on EDM – FRBRoo Application Profile Task Force*. Europeana.

[7] Europeana. 2013. *Europeana Semantic Elements Specification and Guidelines*.

[8] IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. *Functional Requirements for Bibliographic Records. Final Report*. International Federation of Library Associations and Institutions.

[9] Isaac, A (Ed.). 2013. *Europeana Data Model Primer*. Europeana.

[10] Kondylakis, H., Doerr, M. and Plexousakis, D. 2006. *Mapping Language for Information Integration*. Technical report FORTH-ICS-TR385.

[11] Library of Congress. c2014. *MARC standards*. Retrieved March 2, 2014 from http://www.loc.gov/marc/

[12] Miller, E., Ogbuji, U., Mueller, V. and MacDougall, K. 2012. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Service*s. Library of Congress.

[13] Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M. and Gergatsoulis, M. 2007. Ontology-based metadata integration in the cultural heritage domain. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers: 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007*. LNCS 4822, Springer. 165–175. DOI= http://dx.doi.org/10.1007/978-3-540-77094-7_25