A data model for cross-domain data representation

The "Europeana Data Model" in the case of archival and museum data

Steffen Hennicke[1], *Marlies Olensky*[1], *Viktor de Boer*[2], *Antoine Isaac*[2, 3], *Jan Wielemaker*[2]

[1]Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft
Dorotheenstrasse 26 - 10117 Berlin
steffen.hennicke@ibi.hu-berlin.de, marlies.olensky@ibi.hu-berlin.de

[2]Vrije Universiteit Amsterdam, Department of Computer Science
De Boelelaan 1081a - 1081 HV Amsterdam
v.de.boer@cs.vu.nl, aisaac@few.vu.nl, j.wielemaker@cs.vu.nl

[3]Europeana, Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5 - 2509 LK Den Haag

Abstract

This paper reports on ongoing work about heterogeneous and cross-domain data conversion to a common data model in EuropeanaConnect. The "Europeana Data Model" (EDM) provides the means to accommodate data from different domains while mostly retaining the original metadata notion. We give an introduction to the EDM and demonstrate how important metadata principles of two different metadata standards can be represented by EDM: one from the library domain ("Bibliopolis"), and one from the archive domain based on the "Encoded Archival Description" (EAD) standard. We conclude that the EDM offers a feasible approach to the issue of heterogeneous data interoperability in a digital library environment.

1 Introduction

The project Europeana was set up as part of the EU policy framework for the information society and media (i2010 strategy) aiming at the establishment of a single access point to the distributed European (digital) cultural heritage covering all four different domains: libraries, museums, archives and audio-visual archives. In November 2008 a first prototype of Europeana was released providing basic search functionalities over about two million digital object representations. Among other projects Europeana v1.0 and EuropeanaConnect work on completing Europeana's technical components and architecture (cf. Concordia et al., 2010).

2 Cross-domain interoperability

Europeana will be a digital library, a digital museum, a digital archive and a digital audio-visual archive. Its object representations come from heterogeneous sources. Data heterogeneity is a general problem, whenever digital libraries need to interoperate. Thus, issues of cross-domain data representation and different structural and semantic problems need to be addressed. Previous efforts on metadata harmonization include standardization and mappings/crosswalks (cf. Chan et al., 2006 and Zeng et al., 2006). Haslhofer et al. (2010) distinguish between three categories of interoperability approaches: agreement on a certain model, agreement on a certain metamodel, and model reconciliation.

The current metadata schema in use, the Europeana Semantic Elements (ESE) has solved the interoperability problem by agreeing on a common model and standardizing as well as converting the object metadata into flat, Dublin Core based representations. Thus, the original, richer metadata from the provider is lost during the conversion process. However, in the light of data enrichment, contextualization and semantic search functionalities it is important to use a data model that is able to reflect the richness of metadata from the original provider.

The Europeana Data Model (EDM) was developed as a co-effort of Europeana v1.0 and EuropeanaConnect (Isaac et al., 2010). It is an approach which combines two categories of interoperability techniques: the agreement

on a common meta-data-model and model reconciliation, i.e. mappings (Haslhofer et al., 2010).

In the following sections we will explain Europeana's approach to overcome cross-domain data heterogeneity in order to provide useful access to Europe's digital cultural heritage. To illustrate that this data model truly works across domains we will expand on two use cases taken from EuropeanaConnect's ongoing work on data conversion. We have converted the Bibliopolis'[1] metadata schema and the Encoded Archival Description[2] (EAD) standard into the EDM as part of proofing exercises. Bibliopolis is a database about the national history of the printed book in the Netherlands. The Encoded Archival Description (EAD) standard is maintained by the Library of Congress and is an established XML standard in the archival area.

3 The Europeana Data Model

To solve the problem of cross-domain data interoperability the EDM builds on the reuse of existing standards from the Semantic Web environment but does not specialize in any community standard (Doerr et al., 2010). The EDM acts as a top-level ontology consisting of elements from standards like OAI-ORE[3], RDF(S)[4], DC[5] and SKOS[6] and allows for specializations of these elements. Thus, richer metadata can be expressed through specializations of classes and properties. Some elements were defined in the Europeana namespace, yet contain referrals to other metadata standards. This allows for correct mappings and cross-domain interoperability.

---

[1] "Bibliopolis": http://www.bibliopolis.nl/ [7.10.2010].

[2] "Encoded Archival Description": http://www.loc.gov/ead/ [7.10.2010].

[3] "Open Archives Initiative Protocol - Object Exchange and Reuse": http://www.openarchives.org/ore/ [7.10.2010].

[4] "Resource Description Framework (Schema)": http://www.w3.org/RDF/ [7.10.2010].

[5] "Dublin Core": http://dublincore.org/ [7.10.2010].

[6] "Simplified Knowledge Organization System": http://www.w3.org/2004/02/skos/ [7.10.2010].

RDF(S) is used as an overall meta-model to represent the data. The ORE approach is used to structure the different information snippets belonging to an object and its representation. It follows the concept of aggregations (`ore:Aggregation`) and allows to distinguish between digital representations which are accessible on the Web and thus modeled as `ens:WebResource` and the provided object, e.g., represented as a `ens:PhysicalThing`. Furthermore, different, possibly conflicting views from more than one provider on the same object can be handled in EDM by using the proxy mechanism (`ore:Proxy`). The DCMI Metadata Terms describe the objects. SKOS is used to model controlled vocabularies which annotate the digital objects (Isaac et al., 2010).

The EDM will replace the current metadata schema Europeana Semantic Elements (ESE) (Europeana v1.0, 2010) in the next release of Europeana ("Danube" release, scheduled for 2011). The ESE will then become an application profile of the EDM, which will thus be backwards compatible.

4 Bibliopolis

Bibliopolis is the electronic national history of the printed book in the Netherlands curated by the Dutch National Library. The collection consists of 1,645 images related to book-printing. These images are described by metadata records and are accompanied by a thesaurus containing 1,033 terms used as keywords for describing and indexing the images. Both thesaurus and metadata are bilingual (English and Dutch).

Figure 1 shows an example of a Bibliopolis object image and its metadata record. The Bibliopolis metadata is presented in an XML format and has a relatively simple 'flat' structure. Each object is represented by one metadata record `inm:Record`.[7] Individual metadata elements are denoted by single XML tags. The values of the metadata fields are free text terms, which can be present in the Bibliopolis thesaurus.

[7] "inm" is the original namespace of Bibliopolis. "bib" is the new namespace for the Bibliopolis data created during the conversion process.

The Bibliopolis example represents both the simple and the common case as many cultural heritage institutions have similarly structured metadata and thesauri. This example shows how such 'flat' metadata is represented in EDM and demonstrates the use of some of the central features of the model.

```
<inm:Record>
    <inm:NUMMER>6</inm:NUMMER>
    <inm:TITEL>Delftse Bijbel...</inm:TITEL>
    <inm:TITEL_EN>Delft Bible...</inm:TITEL_EN>
    <inm:MAKER>Yemantszoon, Mauricius : d</inm:MAKEF
    <inm:OBJECT>tekstbladzijde</inm:OBJECT>
    <inm:TECHNIEK>boekdruk</inm:TECHNIEK>
    <inm:DATERING>10 jan. 1477</inm:DATERING>
    <inm:CLASSIFICATIE>D</inm:CLASSIFICATIE>
    <inm:ORIGINEEL>Bijbel. Oude
        Testament...</inm:ORIGINEEL>
    </inm:REPRODUCTIE>
    <inm:TWNAAM/>
    <inm:TWOND>typografische vormgeving</inm:TWOND>
    <inm:TWOND>bijbels</inm:TWOND>
    <inm:TWGEO>Delft</inm:TWGEO>
    <inm:OMSCHRIJVING>Eerste bijbel die in het
        Nederlands verscheen...</inm:OMSCHRIJVING>
    <inm:OMSCHRIJVING_EN>The first Bible to
    appear in the Dutch language...</inm:OMSCHRIJVIN
    <inm:AFMETINGEN>27 x 20 cm</inm:AFMETINGEN>
    ...
</inm:Record>
```

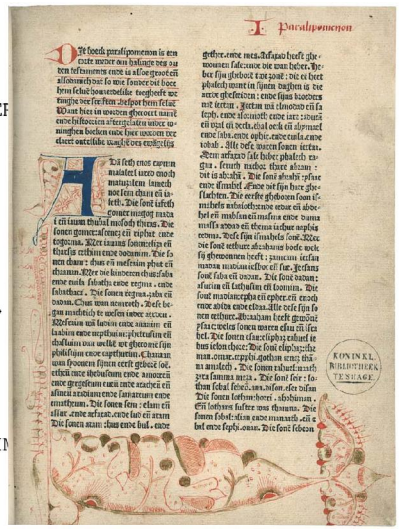

Figure 1: Example Bibliopolis metadata record and the described image

Each `inm:Record` in the original metadata becomes a PhysicalThing-Proxy-Aggregation cluster in the EDM representation (cf. Figure 2). Each of these three EDM resources receives a URI, constructed by concatenating the Bibliopolis namespace prefix, the resource type (`proxy-`, etc.) and a guaranteed unique identifier, in this case the number (`inm:NUMMER`). By having a uniform URI creation scheme, objects referring to other objects can be easily represented in RDF by using URIs as objects. EDM specifies the relations that hold between these resources (`ore:proxyIn`, `ore:aggregates`, etc.) and these are added to the data.

In EDM, the metadata describing the cultural heritage resource itself (e.g., painting, book…) is attached to the `ore:Proxy` using DC Terms properties. The Bibliopolis metadata fields can be represented in EDM in two ways: In the case where an original field exactly matches a DC Terms property (for example `inm:TITEL` and `dcterms:title`), the DC Terms property is used directly. In the case where the match is not exact, a Bibliopolis property is created in RDF which is specified as being a sub-

property of the appropriate DC Terms property (for example `inm:TECHNIEK` is a `rdfs:subPropertyOf` of `dcterms:medium`). Interoperability at the EDM level is ensured through RDFS semantics by using this sub-property method.

Some Bibliopolis metadata fields are actually the identical properties with different language values (for example `inm:TITEL` and `inm:TITEL_EN`). In EDM/RDF these are represented using the same property and a language tagged-RDF literal as value. Figure 2 shows an example.

In EDM associated web pages, thumbnail images and other web resources are attached to the aggregation. As Figure 2 shows, in the case of Bibliopolis, the landing page (the main access page for an object) is represented by `bib:landingPage` which is a sub-property of `ens:landingPage` and has the aggregation as subject. In EDM, the `ens:PhysicalThing` resource of the triangle does not have any properties itself and is only used to relate objects as described by multiple aggregators and represented then through multiple proxies. Also, the relation to a thesaurus (`skos:Concept`) is depicted.

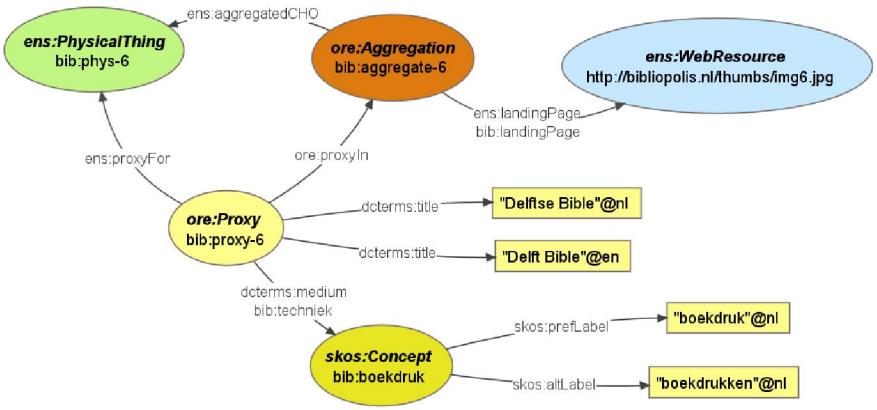

Figure 2: This RDF graph shows part of the converted metadata of a Bibliopolis object.

5 Encoded Archival Description (EAD)

While Bibliopolis exemplifies a simple and very common case of object centric data representation, EAD represents finding aids which describe one

or more archival collections which themselves consist of many files or items organized according to provenance in sequential order and in a contextualizing hierarchy. In other, non-archival terms we can say that an EAD file is one huge record containing many single objects which are contextualized by a hierarchical and sequential order.

Several different EAD dialects exist each of which are subsets of the full EAD model. We use APEnet-EAD which is currently developed by the APEnet project[8] within the context of Europeana. However, the core notion and structure of an EAD representation remains the same.

The `eadheader` element contains bibliographic and descriptive information to identify a finding aid document. Its sibling element `archdesc` holds information about the archival collection as a whole and – within subsequent descendant `c` elements – information about classes, series, subseries, files, and items represented in a hierarchical and sequential order. Files or items generally constitute the smallest unit within the archival description and potentially hold digital representations of the possibly many single items (e.g., paper pages) it contains. All other intermediate levels normally structure the context for a file. The described structure is intrinsic to archival documentation practice and theory. The single file loses most of its information value if it is not properly represented within its context of provenance.

The Bibliopolis example demonstrates central and standard features of the EDM and the conversion process like the mechanism of sub-properties for descriptive metadata or the creation and assignment of URIs to resources. Here we will focus on advanced features for the representation of hierarchical and sequential order in EDM.

Figure 3 shows a simplified snippet from an EAD-XML representation of a finding aid of the Nationaal Archief in Den Haag.[9] The `archdesc` element

[8] APEnet project homepage: http://www.apenet.eu/ [18.10.2010].

[9] The original presentation of this archival fond can be found at http://tinyurl.com/EAD-NatArch [6.11.2010],
the equivalent representation in ESE is at http://tinyurl.com/EAD-EurSemEle [6.11.2010], and a first technical demo of the EDM representation is available at http://tinyurl.com/EAD-EurDataMod [6.11.2010].

contains several descriptive metadata fields which hold information about the title of the whole archival fond (`unittitle`), the time span the material covers (`unitdate`), a call number (`unitid`), the name of the repository where the material is kept (`repository`), and a summary of the contents (`scopecontent`). Further down the hierarchy we see several `c` levels which are of different types: a `series` which contains a `file` which holds two `items`. All these levels have a call number and a title which are constitutive parts of the contextual description. The two items also link to digital representations (`dao`), e.g. digital images, of their contents.

```
<ead>
   <archdesc>
      <did>
         <unittitle>Graven van Holland</unittitle>
         <unitdate calendar="gregorian" era="ce">1189-1660</unitdate>
         <unitid>3.01.01</unitid>
         <repository>Nationaal Archief, Den Haag</repository>
      </did>
      <scopecontent encodinganalog="summary">
         <p>Het archief van de graven van Holland bevat documenten betreffende het
      </scopecontent>
      <dsc>
         <c level="series">
            <did>
               <unitid type="call number">5</unitid>
               <unittitle>STUKKEN BETREFFENDE DE ZORG VOOR HET ARCHIEF</unittitle>
            </did>
            <c level="file">
               <did>
                  <unitid type="call number">2149</unitid>
                  <unittitle>'Remissorium Philippi'; index op de grafelijke regist
               </did>
               <c level="item">
                  <did>
                     <unitid type="call number">2149.1</unitid>
                     <unittitle>Pagina 1</unittitle>
                     <dao xlink:href="http://na.memorix.nl/oai2/?image=na:col1:dat
                     <dao xlink:href="http://beeldbank.nationaalarchief.nl/na:col1
                  </did>
               </c>
               <c level="item">
                  <did>
                     <unitid type="call number">2149.2</unitid>
                     <unittitle>Pagina 2</unittitle>
                     <dao xlink:href="http://na.memorix.nl/oai2/?image=na:col1:dat
                     <dao xlink:href="http://beeldbank.nationaalarchief.nl/na:col1
                  </did>
               </c>
```

Figure 3: Simplified snippet from an EAD-XML representation of a finding aid of the Nationaal Archief in Den Haag.

Figure 4 pictures a simplified graph representation of the example in figure 3 which shows how hierarchies and sequences are modeled in EDM. `Archdesc` and each `c` level are represented by an aggregation with a proxy

for the descriptive metadata.[10] The URI of a resource indicates the type of each level.

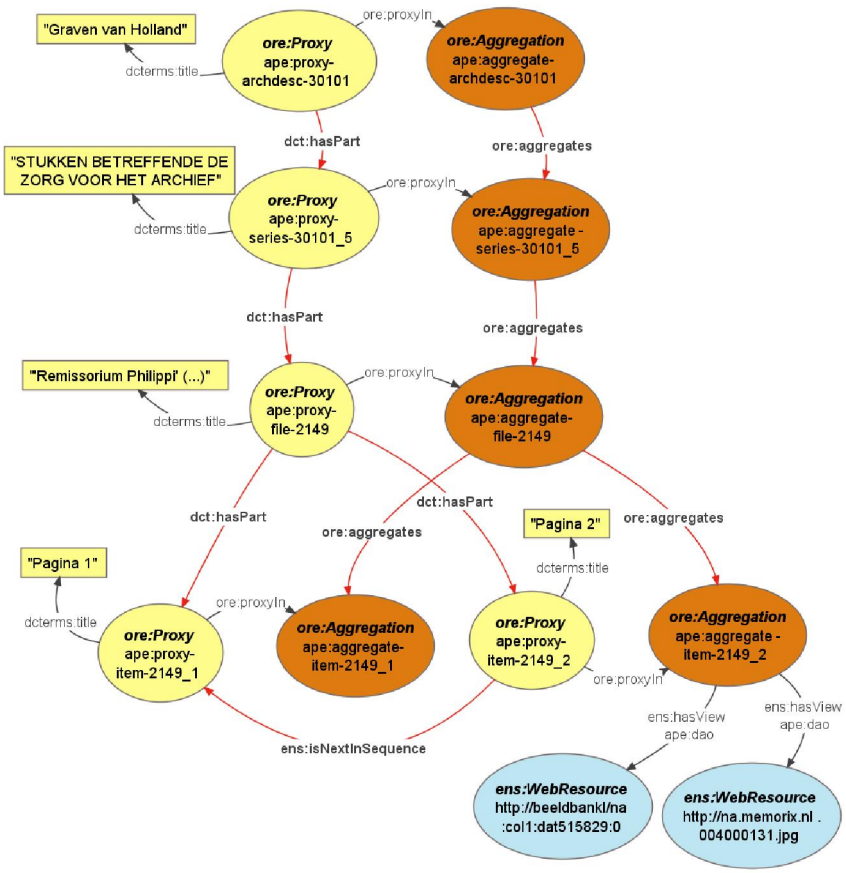

Figure 4: Simplified EDM representation of an EAD structure.

During conversion the EAD hierarchy has been translated into a double hierarchy: The `ore:aggregates` properties between the aggregations mirror the XML-hierarchy of the documentation in the EAD file. At the same time these relations represent, on a more abstract level, the different level of genericity of digital object "packages" submitted via the EAD file to Europeana. The `dct:hasPart` properties between the proxies conceptually

[10] All `ens:PhysicalThing` are omitted, `ens:WebResource` are displayed only for one of the items and the `ens:Proxy` only hold the title of each level.

reflect the documented physical hierarchy of the archival material as it exists in the actual archive. This line of hierarchy says that the archival fond (`archdesc`) incorporates a `series` which has a `file` which holds two `item` as parts. This way the original hierarchical context of description is retained and every part of the complex object EAD file is represented distinctively.

In the XML structure the two `item` elements are in an intentional and meaningful sequence. To express that the `item` with title "Pagina 2" is second in sequence with regard to the `item` with title "Pagina 1" we assert an `ens:isNextInSequence` statement.

This small example shows how EDM models complex hierarchical objects. The `archdesc` level and each `c` level from the EAD file have been converted to aggregations constituting objects in their own right which have been linked together with inter-object properties. In the same way other inter-object relations can be modeled, for instance derivative relations between different translations of a book with the property `ens:isDerivativeOf`.

From a data modeling point of view no structural issues arose. EDM easily represents complex, hierarchical, and sequential objects. The EDM leaves room for data providers to consider different modeling options: For example, with regard to search and retrieval, it is possible to include the `eadheader` as a separate aggregation which describes a printed finding aid as a separate object. It is also possible to consider only levels which hold digital representations worth finding and therefore dismiss all other levels from the EDM representation. In our example above each `c` level in an EAD file is considered as a retrievable object in its own right.

6 Conclusion

Four community workshops[11] confirmed the feasibility of the EDM for the different domains represented in Europeana. It is important to stress that EDM does not make assumptions about the domain models. The two

[11] Held for archives and museums in Berlin, libraries in Amsterdam, and audiovisual archives in Pisa during March and April of 2010.

examples discussed in this paper focus on the difference of flat and hierarchical structures of the metadata, but EDM also accommodates, for example, event-centric models. It is designed to be applied to different metadata structures and our examples provide the proof of concept for two of them.

Currently prototyping continues and additional data sets are converted to EDM. These data sets will be integrated into a demonstrator called ThoughtLab[12] which shows the use of the cross-domain data representation in search and retrieval functionalities envisioned for Europeana.

This work is part of the current development of the EDM and the restructuring of the Europeana information space, which enables new functionalities like semantic search (Gradmann, 2010). It is important to note that the issue of data modeling is a separate step from the issue of data visualization: not all complex data needs to be rendered in end-user interfaces.

The EDM is an approach to interoperability of heterogeneous data in a digital library environment. We showed how EDM accommodates metadata representations from two different domains while building on existing standards and leaving room for specializations. The EDM is aggregation-oriented and abstracts from the domains by remaining minimal in its modeling approach. It demonstrates how a domain-independent ontology defined by an RDF model is a feasible approach to integrate different metadata perspectives by providing a layer of generic properties and classes which at the same time can be specialized. Thus, it is possible to accommodate flat metadata representations like in the case of Bibliopolis but at the same time very complex structures like in the case of APEnet-EAD.

7 References

Chan, L. M., Marcia L. Z. (2006). Metadata Interoperability and Standardization - A Study of Methodology. Part I: Achieving Interoperability

[12] "Semantic Searching Prototype, ThoughtLab": http://www.europeana.eu/portal/thought-lab.html [18.10.2010].

at the Schema Level. D-Lib Magazine 12 (6) June 2006. Retrieved January 12, 2011 from http://www.dlib.org/dlib/june06/chan/06chan.html

Concordia, C., Gradmann, S., Siebinga, S. (2010). Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. In: International Federation of Library Associations and Institutions 36 (1), pp. 61–69. http://dx.doi.org/10.1177/0340035209360764

Doerr, M., Gradmann, S., Hennicke, S. et al. (2010). The Europeana Data Model (EDM). Paper presented at the World Library and Information Congress: 76th IFLA General Conference and Assembly 10-15 August 2010, Gothenburg, Sweden. Retrieved October 29, 2010 from http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf

Europeana v1.0 (2010). Europeana Semantic Elements Specification, Version 3.3. Retrieved October 18, 2010 from http://www.version1.europeana.eu/web/guest/technical-requirements

Gradmann, S. (2010). Knowledge = Information in Context. On the Importance of Semantic Contextualisation in Europeana. Europeana White Paper, 1. Retrieved October 18, 2010 from http://version1.europeana.eu/web/europeana-project/whitepapers

Haslhofer, B., Klas, W. (2010). A survey of techniques for achieving metadata interoperability. In: ACM Computing Surveys 42 (2), S. 1–37. Retrieved January 4, 2011 from http://portal.acm.org/citation.cfm?doid=1667062.1667064

Isaac, A. (ed.) (2010). Europeana Data Model Primer. Retrieved October 18, 2010 from http://version1.europeana.eu/web/europeana-project/technicaldocuments/

Zeng, M. L., Chan, L. M. (2006). Metadata Interoperability and Standardization - A Study of Methodology. Part II: Achieving Interoperability at the Schema Level. D-Lib Magazine 12(6) June 2006. Retrieved January 12, 2011 from http://www.dlib.org/dlib/june06/zeng/06zeng.html