Conversion of EAD into EDM Linked Data

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

[1] Humboldt-Universität zu Berlin, Institut fr Bibliotheks- und
Informationswissenschaft
Dorotheenstrae 26, 10117 Berlin, Germany
{steffen.hennicke,marlies.olensky}@ibi.hu-berlin.de
[2] Vrije Universiteit Amsterdam, Department of Computer Science
De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands
{v.de.boer,aisaac,j.wielemaker}@cs.vu.nl
[3] Europeana, Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5, 2509 LK Den Haag

Abstract. We report on ongoing work in Europeana on the conversion of EAD-XML based archival data to an RDF-based representation using the newly developed "Europeana Data Model" (EDM) ontology. This short paper is based on [4].

Keywords: archive, EAD, semantic web, finding aid, RDF, EDM, Europeana

1 Introduction

The project Europeana[4] 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 cultural heritage. Today, Europeana offers access to millions of objects from all kinds of cultural heritage communities. The aim of the current agenda of Europeana is to provide semantically contextualized object representations and new functionality based on an API approach and on integration into the Linked Data context [1]. To enable this vision the "Europeana Data Model" (EDM) has been developed.

A large portion of the cultural heritage metadata that is prospected to be accessible through Europeana is currently described in archival finding aids used by archives across Europe. The "Encoded Archival Description" (EAD) is an XML standard for encoding such finding aids. As such, a method for converting EAD-compliant metadata to EDM will greatly benefit Europeana's goals.

In this paper, we will describe the basic functionality of the EDM and explain some pivotal principles of archival description embodied in EAD-encoded archival finding aids. After this, we elaborate on the EDM-RDF representation of a concrete EAD encoded finding aid. We conclude with the perceived advantages of such a new data representation.

[4] http://www.europeana.eu

2 Europeana Data Model (EDM)

The EDM has been specifically designed to enable Linked Data integration and 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 [2]. It acts as a top-level ontology consisting of elements from standards like OAI-ORE[5], Dublin Core Terms[6] and SKOS[7] 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.

RDF(S)[8] 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`): This concept allows to distinguish between digital representations which are accessible on the Web and thus are modeled as `edm:WebResource`[9] and the provided object, represented as a `edm:ProvidedCHO`.

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 Dublin Core Terms describe the objects. SKOS is used to model controlled vocabularies which annotate the digital objects [5].

3 Archival Description and Finding Aids

Archival finding aids are guides into the archival material that an archive holds in the form of archival collections. Typically, printed versions of finding aids serve archival users in the reading room and archivists in the reference service as the means to identify relevant archival materials.

The archival material in an archival collection is organized into records. A record denotes a group of documents from the archival collection. It does not describe a single information object like a single book in the library domain.

According to the principle *respect des fonds* the description of the internal structure (original hierarchy and ordering) and the external structure (provenance) of an archival collection provides information necessary to understand context and content of the records and to guarantee their authenticity.

A finding aid contains such information in the form of an archival description. This archival description typically consists of several parts arranged in a

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

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

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

[8] "Resource Description Framework (Schema)": `http://www.w3.org/TR/rdf-primer/` [5.09.2011] and `http://www.w3.org/TR/rdf-schema/` [5.09.2011]

[9] The namespace prefix "edm" stands for the Europeana namespace "http://www.europeana.eu/schemas/edm/".

multilevel hierarchy. The top-most part describes the archival collection as a whole. The following descendant parts describe sub-parts of the previous parts with increasing detail.

The leaves of the descriptive tree are about different kinds of unit of records which constitute the smallest parts within the archival description. The smallest parts of the description do not necessarily correspond to the smallest parts of the archival collection. The unit of a record can be, for example, an item which corresponds to one record, or a file with one or more folders of records.

A call number for the unit of records is used to order one or more physical boxes with archival documents (photographs, legal documents, letters, et cetera) from the archive's depot. Typically, searching for archival material in an archival finding aid means identifying call numbers for units of records whose potential relevancy for one's purpose is judged by the contextual descriptions.

The archival descriptions we find in archival finding aids contain huge and rich amounts of contextual and implicit information (especially through inheritance) in order to enable archival users and archivists to efficiently and effectively locate and discover archival material [3].

4 EAD-encoded Finding Aids

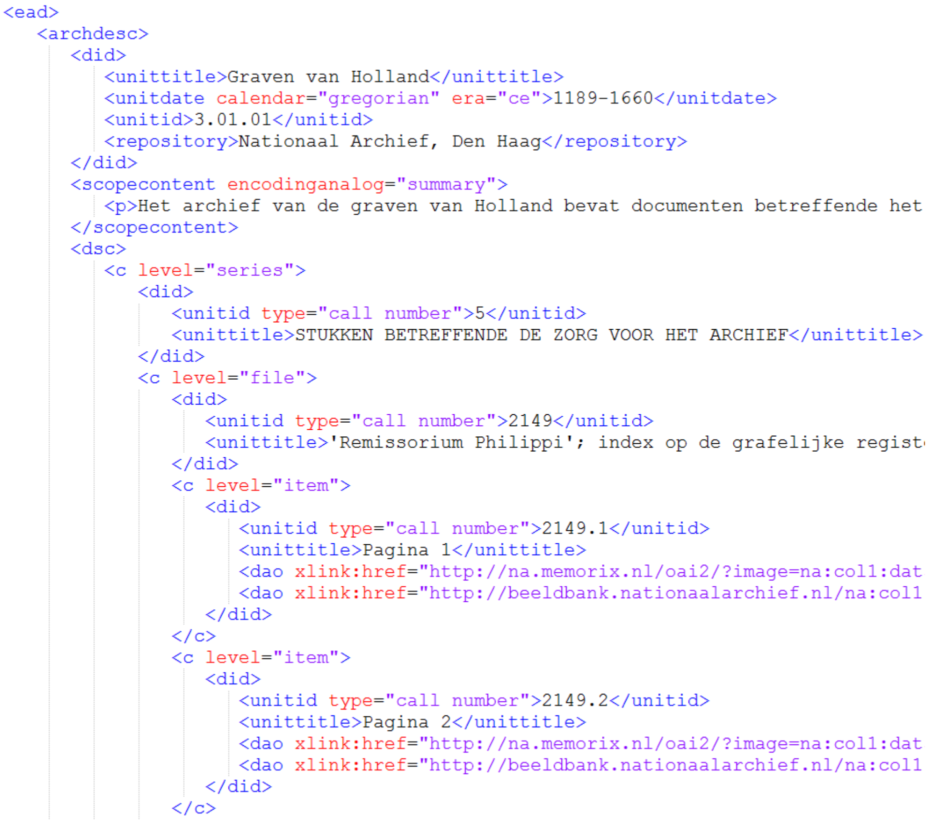
The "Encoded Archival Description"[10] (EAD) standard is the latest and most promising standardization effort for encoding archival finding aids for a digital environment. It provides the infrastructure to accommodate most designs of finding aids. Typically, an institution uses a subset of the full EAD model. Our conversion method is specifically designed for and tested with APEnet-EAD, which is currently developed by the APEnet project[11] within the context of Europeana. We expect, however, that our method is applicable to other EAD dialects with slight modifications to the script. An EAD-document typically contains the description of one archival collection in the form of a finding aid. The `<eadheader>` element[12] contains bibliographic and descriptive information to identify a finding aid document. Its sibling element `<archdesc>` holds information about the archival collection as a whole. In our example, the `<archdesc>` element 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>`).

Within the `<archdesc>` element, `<c>` elements of different types (classes, series, subseries, files, or items) represent the multilevel hierarchy of the archival description providing the intermediate structure and context for the archival material described in a finding aid. In our example we find a series which contains

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

[11] The "Archives Portal Europe" (`http://www.apenet.eu/`) is a data aggregator for the European archives.

[12] The element has been omitted in figure 1.



```
<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>
```

Fig. 1. An EAD-XML snippet taken from an EAD-encoded finding aid of the Dutch National Archives.

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. To suit the original purpose of the finding aids, it is crucial to retain all descriptive and contextual information about records when transforming this structure to RDF.

5 Conversion of EAD-XML to EDM-RDF

Archdesc and each c-level are represented as a "`edm:ProvidedCHO - ore:Aggregation`" cluster (cf. figure 2). Both resources are connected via the property `ore:aggregates`. Their URIs are constructed by concatenating the apenet namespace prefix, the resource type (aggregation-, cho-), the type of the EAD-level (archdesc, series, file, item), and a guaranteed unique identifier, in this case the unitid of the respective c-level. By having a uniform URI creation scheme, objects referring to other objects can be easily represented in RDF by using URIs as objects. This way, if objects are added or metadata is updated, we ensure that existing objects receive their old unique URI while added objects receive a new unique URI. The metadata describing the cultural heritage resource

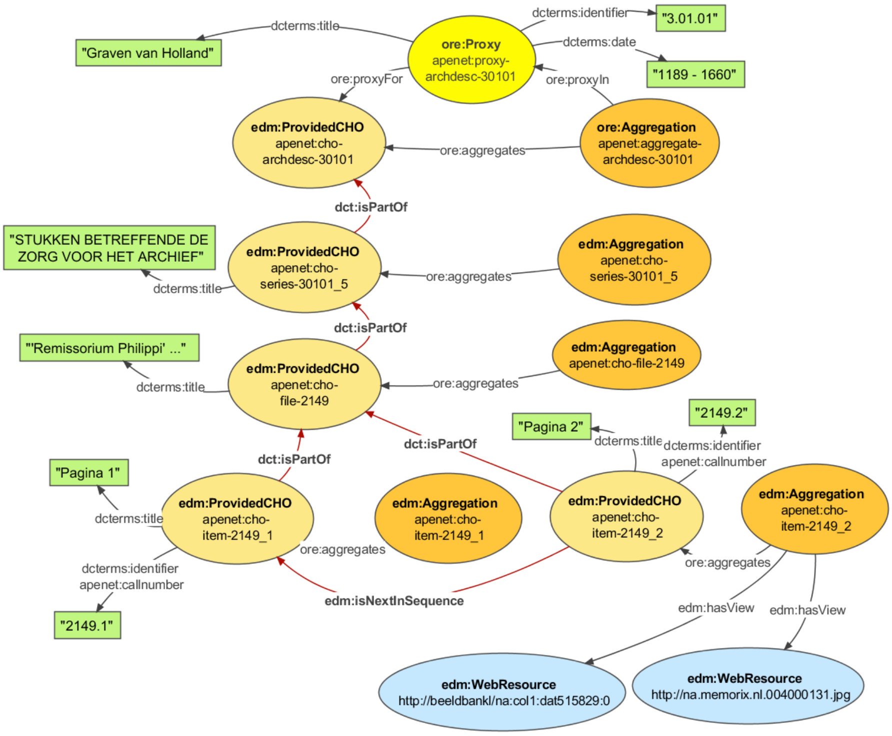


Fig. 2. Parts of the EDM-RDF representation of the EAD-encoded finding aid shown in figure 1.

itself (in our case, those are the different parts of the archival description describing and contextualizing records) can be either attached to the `edm:ProvidedCHO` directly or to a third resource called `ore:Proxy`. The proxy mechanism allows having different views, i.e. descriptions, of one and the same object. In such case a data provider can create proxies and attach the different views to it and thereby keep them distinct. Europeana itself, when doing semantic enrichment, will create proxies in order to retain the original structure and provenance information for the metadata. In our example, the proxy is not necessary as there are no conflicting descriptions. We only created a proxy for the first level (archdesc) to demonstrate the functionality.

The fourth EDM-class shown in figure 2 is `edm:WebResource` which represents associated web pages, thumbnail images or any other web resources and is attached to the aggregation. The URI for such a WebResource is typically the URL provided for the digital representation.

The descriptive 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 `<unittitle>` and `dcterms:title`), the DC Terms property is used directly. In the case where the match is not exact, an APEnet-EAD property is created in RDF which is specified as being a sub-property of the appropriate DC Terms property: For instance, `apenet:callnumber` is a `rdfs:subPropertyOf` of `dcterms:identifier`, as shown at the two leaves in figure 2. Interoperability at the EDM level is ensured through RDFS semantics by using the sub-property method. The language of the content of descriptive metadata fields can be specified by adding a language tagged-RDF literal as value.

The `edm:ProvidedCHO` resource carries not only descriptive metadata but also properties which are used to relate other objects. During conversion the EAD hierarchy has been translated into a hierarchy relation between the `edm:ProvidedCHO` resources which are connected by `dct:isPartOf` properties. This hierarchy mirrors the XML-structure of the multilevel archival description of the EAD file. At the same time these relations represent, on a more abstract level, the different levels of generality of digital object ”packages” submitted via the EAD file to Europeana. The `dct:isPartOf` properties conceptually reflect the documented structure of the archival material, i.e. the archival collection (archdesc) incorporates a series which has a file which holds two items as parts.

The two c-levels of type item at the bottom in the XML structure are in an intentional and meaningful sequential order. This sequence is expressed by asserting an `edm:isNextInSequence` statement between the resource with title ”Pagina 1” and the resource with title ”Pagina 2”.

6 Discussion: EAD in a Linked Open Data Environment

We demonstrated how an EAD-XML encoded archival finding aid can be modeled in a RDF-based representation. The representation in an RDF graph makes implicit information explicit, for example, the hierarchical and sequential relations between the different parts of the archival description. We aimed at a conversion which produces a RDF representation which stays as close as possible to the original structure of the APEnet-EAD model. This way, we have a conversion template which is feasible for many different variations of EAD encoded finding aids. The method also entails, however, that not all implicit information pieces in the descriptive metadata have been made explicit or have been connected: for example, the unit titles on the different levels remain only indirectly connected to each other. A user being on the level of one of the leafs of the descriptive tree, most certainly needs to know that he is on page 1 (Pagina 1) of the book ”Remissorium Philippi (...)” of the ”counts of Holland”. In order to use this data one needs a special reasoner or a Linked Data browser which brings those information pieces from different levels together.

Another option is to merge intermediate levels of the archival description without digital representations and the descriptive metadata we find there into the leafs of the descriptive tree already during the conversion. In the context of Europeana, each `edm:ProvidedCHO` is an object which can be found via searches. If those objects have no digital representations or only partial descriptions then their information value in the context of Europeana can be questioned. Data providers creating mappings to EDM, on the one hand, have to consider how they want to represent their data in the context of the Europeana information space, but, on the other hand, enjoy great flexibility regarding data modeling with EDM.

At the same time we showed the capability of the EDM to accommodate such a particular archival domain model. The EDM is able to accommodate EAD and other different standards as we demonstrated elsewhere [4]. One of the main reasons to use EDM as the ontology (instead of some specific EAD-RDF model) for an EAD conversion is, that the archival data can now be connected to museum, library and other archival data within the Europeana information space. Contextualization through external resources is now possible. For example, person names can be linked to concepts in a controlled vocabulary like VIAF[13]. Such contextualization allows, for example, to disambiguate meaning or to relate the original object to other cultural heritage objects annotated with the same person name. Europeana is planning to do enrichments for a number of fields like, for example, person or place names.

References

1. Concordia, C., Gradmann, S., Siebinga, S.: Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. IFLA Journal 36(1), 61–69 (2010)
2. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The Europeana Data Model (EDM) (2010), http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf
3. Haworth, K.M.: Archival Description: Content and Context in Search of Structure. In: Pitti, D.V., Duff, W.M. (eds.) Encoded Archival Description on the Internet, pp. 7–26. Haworth Information Press, Binghamton NY (2001)
4. Hennicke, S., Olensky, M., Boer, V.d., Isaac, A., Wielemaker, J.: A data model for cross-domain representation: The "Europeana Data Model" in the case of archival and museum data. In: Griesbaum, J., Internationales Symposium für Informationswissenschaft 12, .H., Hochschulverband für Informationswissenschaft (eds.) Information und Wissen: global, sozial und frei?, vol. 58, pp. 136–147. vwh Hülsbusch, Boizenburg (2011)
5. Isaac, A.: Europeana Data Model Primer (2010), http://version1.europeana.eu/web/europeana-project/technicaldocuments/

[13] "Virtual International Authority File": http://viaf.org/ [16.07.2011]