Improving Online Access to Archival Data

Vittore Casarosa[1], Carlo Meghini[1], and Stanislava Gardasevic[2]

[1] ISTI-CNR, Pisa, Italy
[2] DILL International Master, University of Parma, Italy

Abstract. Archives are memory institutions whose original mission was to preserve and provide access to a set of carefully selected, arranged and described documents to a small number of scholars interested in their contents. For those specialists, the usual way to find information in an archive is by way of "finding aids", i.e. descriptions of the arc-hive contents that reflect the hierarchical structure by which data are physically arranged in an archive. With the increased availability of archival holdings accessible on the Web, archives are now widening the range of users, and the use of online finding aids has proved to be too complicated for the non-specialists. This is mostly due to the hierarchical nature of the description, usually represented on line with a standard called EAD (Encoded Archival Description). This paper is the synopsis of a Master Thesis, where a methodology has been developed to represent the information contained in finding aids with a different standard, namely EDM (Europeana Data Model), which is used by the Europeana digital library and is becoming the de-facto standard for metadata interoperability. EDM allows a much more intuitive representation of the archive content and the possibility to access data from many different access points.

Keywords: Archive, EAD, finding aid, EDM, Europeana Data Moodel.

1 The Structure of Archives

1.1 The Archival Fond

Archives differ from other memory institutions in the nature of materials they have. Contrary to libraries, where usually the material collected are just "copies" of books and journals, the material in archives and manuscript libraries are the unique records of corporate bodies and the papers of individuals and families. Therefore archival descriptions have to reflect this peculiarities, retaining all the informative power of a record, and keeping trace of the provenance and original order in which resources have been collected and filed by archival institutions.

This approach emphasize the central concept of archival science, which is "fond", i.e. "all of the documents naturally generated and/or accumulated and/or used by a particular person, family or corporate body in the conduct of personal or corporate activity". This definition leads to the fundamental archival principle (respect des fonds), which is dictating that resources of different origins are to be kept separate, in order to preserve the context in which they were found and the context in which they

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 153–162, 2013.

were created. Furthermore, documents or records kept in archive are usually related to other documents, and are grouped into identifiable subgroups. This kind of record keeping and describing fosters the use of a hierarchical model. The hie-rarchical structure of the archive expresses the relationships and dependency links between the records of the archive. Therefore, a fond is usually organized in sub-fonds, which in turn can be organized in series and sub-series, formed by archival units. Following this structure, archival descriptions also proceed from general to specific, and for every unit of description they show its relationships and links with other units and with the general fonds. Archival descriptions can be presented as a tree, as shown in Figure. 1.

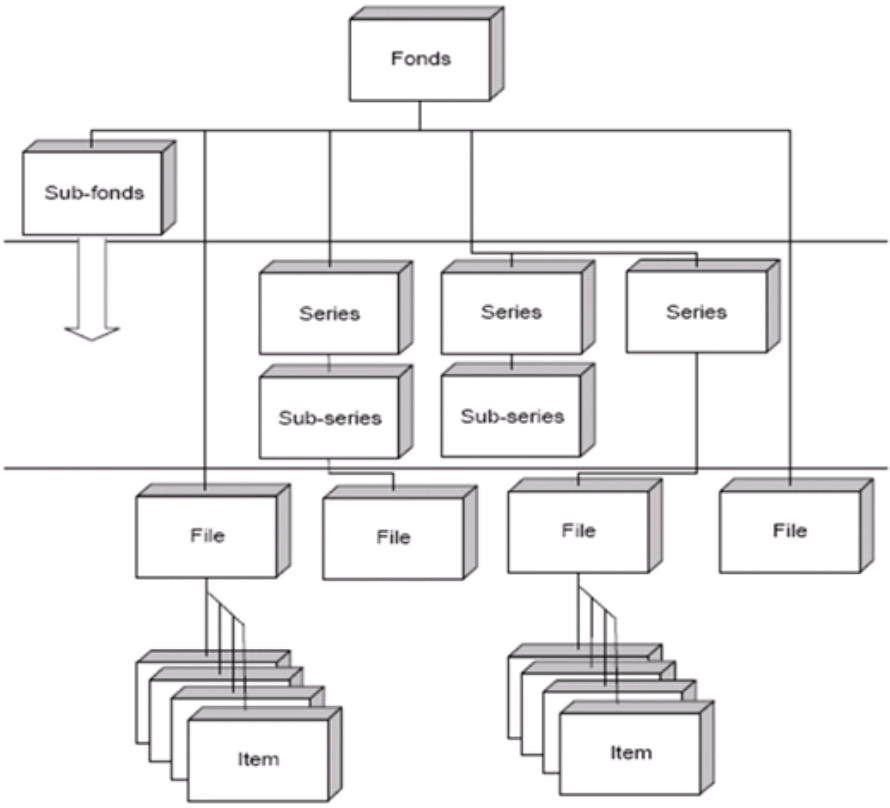

Fig. 1.

1.2 The Finding Aids

The gate to the archival holdings are finding aids, based on archival description practice. Finding aid is a term ordinarily used only in archives, and in general it can

include also card indexes for manuscript collections, administrative histories, and inventories for archives. Finding aids are used to access archival materials, and they contain far more information about a collection than can be found in a summary catalog record. Finding aids are generally created in the course of processing a collection and usually reflect the hierarchical arrangement of the materials. Often, many finding aids start by describing a large group of materials, usually the entire collection or record group, and then move to the description of the series of the firs level components, followed by the description of smaller and smaller components, such as subseries, files and possibly even items. The description of lower levels inherits the description of the preceding levels. At the same time, finding aid acts as a collection management tool for archivist and access point for the researchers

1.3 The EAD Standard

EAD, Encoded Archival Description, a standard for representing finding aids, was started in the early nineties. The design of EAD was based on the following criteria: "1) ability to present extensive and interrelated descriptive information found in archival finding aids, 2) ability to preserve the hierarchical relationships existing between levels of description, 3) ability to represent descriptive information that is inherited by one hierarchical level from another, 4) ability to move within a hierarchical informational structure, and 5) support for element-specific indexing and retrieval". Based on these requirements, XML was chosen as the formal syntax to represent the finding aids, so that an EAD encoded finding aid becomes an XML document written according to the specifications of the EAD XML Schema. Today, after several revisions, EAD is a really global standard being used by a wide variety of institutions throughout the world.

At the same time, EAD has also become the target of several critiques from archival theorists, many of them addressing its usability. The main problems that have been reported when using online finding aids encoded in EAD can be summarized as follows.

- the lack of alternative access points for users, because of the arrangement of materials according to provenance or original order of records;
- the complicated terminology; archivist should map technical terminology used as subject access points and for labeling data elements to a less technical vocabulary in order to facilitate resource discovery by non-expert users;
- finding aids consist of extensive contextual description of the circumstances surrounding the creation of its materials, and de-contextualized access to archival materials is very difficult;
- the length of the files and navigational complexity makes the process of discovery very hard;
- the administrative information that is woven throughout the finding aids is confusing;

- the collective and hierarchical description of the material and the lack of item-level description prevents an easy access at item level and a quick finding of a known item;
- the traditional finding-aid is designed to be used in an environment where the archivist acts as a mediator between the user and the finding aids, which is almost impossible over the Internet.

2 EDM, the Europeana Data Model

The Europeana Data Model (EDM) is a model for structuring the data that the Europeana Digital Library will be ingesting, managing and publishing. Europeana is a major effort of the European Union to create a digital library containing the cultural heritage of Europe. Today it already contains about 18 millions items, provided by a number of memory institutions all over the world. EDM was defined not only to support the richness of the content providers' metadata but also to enable data enrichment from a range of third party sources and to facilitate the publishing of (some of) Europeana content in the Linked Open Data cloud. The main requirements considered for the design of EDM were:

- distinction between "provided object" (painting, book, movie, archaeology site, archival file, etc.) and the digital representation(s) of the object
- distinction between the object and the metadata record describing the object
- multiple records for the same object should be allowed, containing potentially contradictory statements about an object
- support for objects that are composed of other objects
- compatibility with different abstraction levels of description
- provide a standard metadata format that can be specialized
- provide a standard vocabulary format that can be specialized
- allow data integration in an open environment, where it is impossible to anticipate all the data that will be contributed
- allow for rich functionality, possibly via extensions
- re-use existing (standard) models as much as possible

These design criteria have been the basis for the choice of the Semantic Web principles for EDM, providing a model which can be seen as an anchor to which various finer-grained models can be attached, making them (at least partly) interoperable at the semantic level, while retaining original expressivity and richness of original data.

The low level syntax for representing resources and their properties is RDF (Resource Description Framework), usually represented as graphs for "human consumption" or as XML documents for "computer consumption"; the high level syntax is OAI-ORE (Object Re-use and Exchange), which easily supports the ideas of the

Linked Data approach, emphasizing the re-use and linkage of richly described resources over the web. Fundamental for EDM is the OAI-ORE notion of "aggregation", which allows to link together an object and its digital representation(s), and the notion of "proxy" which allows to represent different views on the same resource. In Figure 2 we illustrate these ideas using as an example the painting of Mona Lisa.

Fig. 2.

The top element is an OAI-ORE aggregation, identified by the URI ex1:aggregation000PE025604, which links together (ore:aggregates) the resource Monna Lisa, identified by the URI ex1:object000PE025604, provided (dc:creator) by the Direction des Musèes de France and and two digital representations (ens:WebResorce) of this resource. Additional information about the resource (dc:creator, dcterm:title, i.e. metadata records) are provided through the proxy ex1:proxy000PE025604. This allows to attach to the same resource another proxy (possibly coming from another provider) containing additional information for that same resource, and maintaining a clear distinction about the provenance of the two different sets of information.

As can be seen from the example above, in addition to defining terms in its own name space (abbreviated in ens:), EDM (re) uses as much as possible existing name spaces (i.e. their semantic), such as those defined for RDF, RDFS, SKOS, OAI-ORE, Dublin Core.

3 Mapping EAD to EDM

The EAD data has a hierarchical structure with descriptions associated with the nodes of the hierarchy. For this reason it is convenient to divide the general problem of mapping EAD into EDM into two parts: the structural mapping, i.e. the

transformation of an EAD hierarchy into an equivalent EDM aggregation; and the metadata mapping, that is the transformation of the descriptions found in the EAD nodes into an equivalent EDM metadata record. It is important to remember that in EAD the description associated with a node inherits all the descriptions of its ancestors.

3.1 Transforming an EAD Hierarchy into an EDM Aggregation

The steps for the first part of the transformation are as follows:

1. transform each EAD tree node C into an EDM Aggregation A
2. associate an OAI-ORE Proxy P to the Aggregation A, by means of the OAI-ORE property ore:proxyIn;
3. use the Proxy P as a representative of the real-world entity that node C is about, i.e. the content described by node C;
4. use the Dublin Core property dc:hasPart to relate the proxy P with the proxies defined for the children of node C in the EAD tree. In this way, the EAD tree is represented by the tree induced by the dc:hasPart property;
5. retain the order of the sibling nodes of C by means of the property ens:isNextInSequence.

The first steps of the transformation is shown in Figure 3, where the proxies have been omitted.

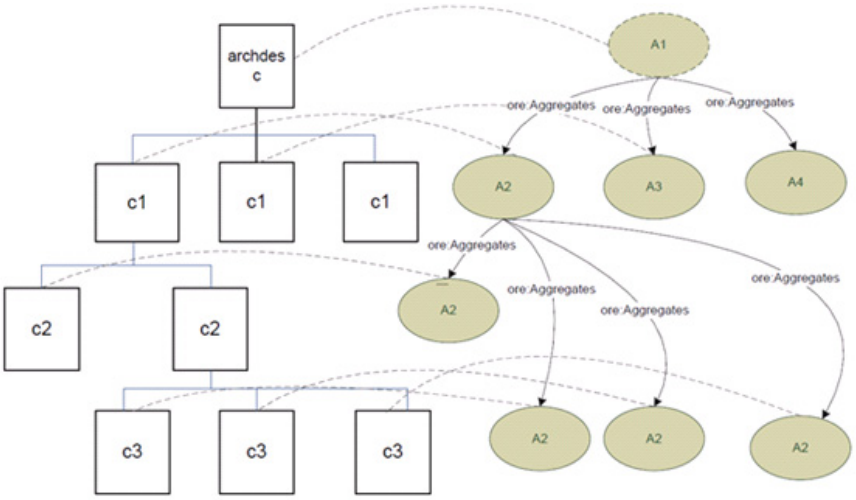

Fig. 3.

3.2 Mapping EAD Values into EDM Values

In the second part of the mapping, EAD elements and their possible attributes are mapped to corresponding EDM properties. To find in EDM a property equivalent (or as close as possible) to a source element, the EDM element specification should be consulted in order to see the definitions, constraints and examples of usage for all EDM classes and properties. When mapping to EDM properties, one should choose those properties carrying as much as possible semantic similarity to the elements or attributes of the original EAD schema, in order to retain as much original information as possible. EDM offers a range of properties, which are mostly defined in Dublin Core and Europeana namespaces, and to which more specialized ones can be attached and declared as subproperties.

The core idea behind converting EAD data into EDM is that every complex element, i.e. an element carrying all the information related to its ancestors (the EAD hierarchy) maps to a resource, i.e. a node of the EDM aggregation (more precisely, it maps to the proxy of the EDM node representing the corresponding node in the EAD hierarchy), and every atomic attribute maps to an attribute of this node. It should be remembered that, based on the EDM model, the metadata values are attached to the proxy of a resource, and not to the resource itself.

3.3 Validation of the Process

The process described above was validated by applying it to archival data coming from the Multimedia Archive of Accademia Nazionale di Santa Cecilia (ANSC). ANSC is a musical academy located in Rome, Italy and one of the oldest musical institutions in the world. The entire patrimony of this institution is about 120,000 volumes and publications, mainly scores, monographs and periodicals about music. Two fonds of this archive (Ethnomusicology Fond and Audio Video Fond) were mapped to EDM for the purpose of validating the method and analyzing the process.

The descriptions of the two fonds chosen was made available as two separate EAD XML files, and each fond was processed separately. The separation of the different levels found in the description of the fonds was performed by using ad hoc software developed at ISTI-CNR. For each extracted level a separate XML file was created, and each level was analyzed to make sure that the mapping of the nodes of a given level would cover all the possible elements at that level.

The result of this work was summarized in two metadata mapping tables, one for each fond. An excerpt of one table is shown in Figure 4. In column (a) there is the path in which the EAD elements were encountered in the original file; in column (b) there is the meaning (the semantics) of these elements; in column (c) there is the default values of these element and in column (d) there is the most appropriate EDM counterpart. Finally, column (e) contains the the RDF objects created for the composite elements.

(a)	(b)	(c)	(d)	(e)
<c>	fond , highest node		create instance of ens:ArchivalFond, domain of: ens:IsPartOf (to recordgrp Proxies)	create subclass of ens:NonInformationResource called ens:ArchivalFond
#level	Fond	fond	dc:type	
#id	Identifier		dc:identifier	
#audience	internal			/
<did>				/
<unit-title>		Archivio di Etnomusi-cologia	dc:title	dc:title
<uitid>	call number/reference code, value not mapped		/ ens:currentLocation	create instance 1 of ens:Place
#country-code	IT			ens:country (to 1:Place)
#repository-code	ANSC		dc:source (to 1:Place) instance 1 of class:Agent, skos-altlable:ANSC, this URI will hold all the data on ANSC, address..+ skos altlable ANSC (to 1:Agent)	

Fig. 4.

4 Conclusions

The main purpose of transforming the EAD representation into EDM is an attempt to make online access to finding aids of archives more “user-friendly” for the casual user. From the insight gained through the validation of the transformation process, despite the limited size of the archival data used, it seems that (at least to some extent) the goal has been achieved. The main improvements to on-line access for the general public of archives can be summarized as follows:

- The specialized archival terminology, through the mapping defined in the mapping tables, is translated to the more general terms used in EDM, making access more intuitive and easy; in addition, it eliminates the many inconsistencies of the terms

used in different archives (and on their web sites), which makes archival research even more confusing.

- In the hierarchical structure of finding aids discovery is usually available through a top-down approach, while using EDM as a query language any node can now be reached directly, and from there a user can go in any possible direction.
- In EAD the information is often buried so deep in the hierarchical structure of the file, that the Web crawlers have problems in indexing it; in the mapping to EDM, the information from the inner levels is extracted and is equally accessible to search engines as the one from the upper ones.
- If the mapping to EDM from different archives is done in a consistent way, it would allow to search for information over more than one archive, providing the same functionality as the union catalog for libraries.

Along the lines of the last point, we might add that the use of existing authority files for person names and for geographical names would provide a great added value to the archival data. As a general rule, genealogists and historians account for more than 50% of archive users, and they usually search for information starting with person or place name. Authority files would help overcome problems caused by different spelling for the name of a person or a location, or to account for the change of names over time.

In a broader perspective, we should consider also that once that the archival data is available in EDM representation it would be possible to overcome the "principle of provenance", i.e. the fundamental archival principle by which records of different origins (provenance) should be kept separate in order to preserve their context. The consequence of this principle is that archival researchers often need to access several fonds in order to collect material of interest that is kept (and described) in separate fonds, but that is logically connected in some way. By applying the ideas of Linked Open Data, and creating links between archival collections and other(re)sources on the Web it would be possible for a researcher to easily discover contextually related material of different provenance, possibly getting new (and may be unexpected) perspectives on the subject of interest.

References

1. Carpenter, B., Park, J.: Encoded Archival Description (EAD) Metadata Scheme: An Analysis of Use of the EAD-Headers. Journal of Library Metadata 9(1), 134 (2009)
2. Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization–A Study of Methodology. Part I. D-Lib Magazine 12(6) (2006)
3. Coats, L.R.: Users of EAD - Finding Aids: Who Are They and Are They Satisfied? Journal of Archival Organization 2(3), 25 (2004)
4. Definition of the Europeana Data Model elements Version 5.2.1, Europeana v1.0 (2011)
5. Europeana Data Model Primer. Europeana v1.0 (2010)
6. International Council on Archives, Statement of Principles Regarding Archival Description. Archivaria 34 (1992)

7. Meghini, C., Isaac, A., Gradmann, S., Schreiber, G., et al.: The Europeana Data Model. In: ECDL Workshop on Very Large Digital Libraries, Glasgow, September 10 (2010)
8. Pitti, D.V.: Encoded Archival Description: An Introduction and Overview. ESARBICA Journal 20, 71–80 (2001)
9. Pitti, D.V., Duff, W.M.: Encoded Archival Description on the Internet. Haworth Information Press, Binghamton (2001); Also published as Journal of Internet Cataloging 4(3/4)
10. Theodoridou, M., Doerr, M.: Mapping the Encoded Archival Description DTD Element Set to The CIDOC-CRM. Technical Report 289, ICS-FORTH (2001)