Uniting Libraries And Archives: How An Integrated Metadata Strategy Can Produce A Connected Research Environment

Richard Gartner [a]
Centre for e-Research, King's College London, 26-29 Drury Lane, London, UK

Received 31 January 2014, accepted 14 February 2014

Abstract. This article details the work of the European CENDARI (Collaborative EuropeaN Digital Archive Infrastructure) project which aims to create a unified query environment for historical archives and form the basis of a digital eco-system on which research infrastructures can be built. The long-established division between metadata practices in the archive and library domains and its obsolescence in the context of the digital information environment are discussed. The CENDARI project has devised an XML-based architecture which aims to bridge this divide. To enable this, a new schema, the CENDARI Collection Schema (CCS) has been constructed which links archival records to library catalogues and also to the Semantic Web. In this way, these historical boundaries are eroded and the full potential of collections can be realised.

Citations: Richard Gartner. Uniting Libraries And Archives: How An Integrated Metadata Strategy Can Produce A Connected Research Environment – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(16)** 2014 – Pp. 9-13.

Keywords: metadata; archives; libraries; archival practice; cataloguing practice; digital eco-systems; semantic Web.
Short title: Uniting Libraries and Archives.

Introduction

Although libraries and archives, both key resources in academic research, are inevitably joined symbiotically in many ways (including often in their administrative and physical co-location), they are usually perceived as far apart in their approaches to metadata. For historical reasons, each domain has evolved its own standards for this, often for practical reasons dictated by their divergent functions but in many cases following traditional imperatives which have their origins in the history of their development. In the analogue era in which many of these approaches were initially conceived such disparities could operate without any significant impact on the effectiveness of their respective operations: in the digital era, however, where the boundaries between libraries and archives become much more fluid, they can present major impediments to facilitating research.

In the contemporary research environment, the distinction between archival and library resources is essentially irrelevant for most users of collections. In the digital world, it is necessary to move beyond any suggestion of polarised approaches and seek out methods for integrating resources into dynamic research environments. Such environments not only include pre-existing collections and the metadata necessary to find and utilise them, but also dynamically-created content produced as the research process proceeds. They can, therefore, no longer be regarded as static objects produced by domain experts or practitioners (such as the archivist or cataloguer) but as 'digital eco-systems' [1], constantly evolving systems of which research sources are only one component.

This article examines one method by which the divergent worlds of archival and library metadata practice can be integrated in order to allow them to act as the core of such a digital eco-system. The approach described here was constructed as part of the European CENDARI (Collaborative EuropeaN Digital Archive Infrastructure) project [2], which is attempting to provide a unified enquiry environment for existing archives and resources in the areas of medieval and modern European history. To enable this, the project has produced an XML metadata schema, known as the CENDARI Collection Schema (CCS), which is designed to act as an intermediary between established schemas in multiple domains and as a kernel on which the dynamic content of an eco-system can be built.

1. Divergent approaches to metadata

In the archival sector, the primary method of documenting the contents of a collection is the *finding aid*. This is traditionally a single record which aims to describe the *fonds*, a set

[a]Corresponding author, email: *richard.gartner@kcl.ac.uk*

of documents which are considered to share the same source. In a finding aid, the *fonds* is usually divided hierarchically into subsidiary components, ranging from collections (the next level down), though series, sub-series and folders down to individual items. This hierarchical description of the contents of a *fonds* usually forms the bulk of a finding aid, but it is often supplemented by textual commentaries on such facets as the history of the collection, biographical information on those involved in its creation, information on the repository which holds it, and core administrative information such as restrictions on accessing its contents.

The principles underlying this approach have their origins over 150 years ago: they are generally considered to have been codified by the historian Natalis de Wailly who in 1841 suggested that the archivist should aim "to gather together by fonds, that is to unite all the deeds (i.e., all the documents) which come from a body, an establishment, a family, or an individual, and to arrange the different fonds according to a certain order" [quoted in 3]. The principle enunciated here, generally known as *respect des fonds*, establishes two principles, shared provenance and the assignment of an ordering of materials, which continue to this day; these underlie the contemporary finding aid, in both its scope and its (usually hierarchical) arrangement.

By contrast the library sector has tended to avoid notions of a discrete, closed *fonds* or the imposition of any ordering of collections above the level of the individual item. Libraries have usually concentrated on the unitary object, usually the book on the shelf. This item-centric approach to metadata applies even in the case of multi-item library objects such as the entire run of a journal, which generally receives a single entry in a catalogue as if it were a monograph. These conventions also owe their origins to a major figure of the 19th century, in this case Anthony Panizzi, whose *Ninety-One Cataloguing Rules* from 1841 [4] still underlie the principles of much contemporary cataloguing practice.

These divergent approaches have been carried forward into the electronic age and into the metadata standards which attempt to move their respective cataloguing traditions into formats more suitable for the imperatives of digital metadata. In the archival world, the Encoded Archival Description (EAD) [5], an XML schema for encoding and exchanging information of the contents of archives, effectively translates the structures and conventions of traditional finding aids into a machine-readable syntax. This is particularly evident in its document-centric architecture which retains much of the structure of the printed finding aid, and its hierarchical arrangement with the *fonds* at its top level.

The library sector, on the other hand, remained firmly focussed on its item-level viewpoint when it devised the MARC (MAchine-Readable Cataloguing) standard [6] in the 1960s. This essentially translates the conventions of the card catalogue to the machine-readable age, maintaining many of its conventions which are essentially irrelevant for digital data (such as its differentiation between main and supplementary entries). Despite the limitations imposed by its origins, the MARC standard has revolutionised library science, allowing an interoperability which has allowed the creation of extensive union catalogues, such as WorldCat [7], which are such essential features of the contemporary researcher's resources.

For the researcher, however, archives and libraries are often equally important resources and this divide is an impediment to resource discovery rather than an aid to it. To produce a seamless enquiry environment for researchers which allows them to access archival and library holdings together requires a metadata strategy which integrates these approaches and allows their divergent approaches to become invisible to the user.

2. The CENDARI project

One current initiative which is attempting to do this is the European CENDARI (Collaborative EuropeaN Digital Archive Infrastructure) project [2], a collaboration between 14 universities and libraries in Ireland, UK, France, the Czech Republic, Germany, the Netherlands, Serbia and Italy. The project aims to build a research infrastructure which will integrate digital archives in the subject areas of medieval and modern European history. One of its deliverables is a unified enquiry environment for existing archives and resources in these two subject domains.

These have polarised emphases in their metadata requirements which correspond neatly to the archive/library divide: the medievalists are particularly concerned with complex objects at the item level (for instance, manuscripts), the modern historians more with finding presently undiscovered materials in existing archives. The former are therefore more interested in detailed item-level descriptions, often with complex codicological information for medieval manuscripts, the latter require sophisticated collection-level descriptions to facilitate resource discovery. Uniting the two into a coherent, unified metadata environment is necessary to allow the two domains to integrate into a single research tool.

Some components of this environment can already be encoded in pre-existing schemas; wherever possible the project uses these, adapting them if necessary to the particular requirements of the intended research environment. Descriptions of the collection-holding institutions themselves, for instance, can readily be accommodated in the pre-existing Encoded Archive Guide (EAG) schema slightly modified to allow more precise descriptions of some elements [8].

For item-level descriptions which mesh with library metadata practices, two pre-existing standards can also be used in conjunction. The more generic elements for these can be encoded in MODS (Metadata Object Description Schema) [9], an XML schema for bibliographic descriptions which is particularly designed for digital objects. MODS is useful for integrating with library collections as it is designed specifi-

cally to interoperate with the MARC standard to which the majority of its elements can be mapped. Using MODS thus allows one link in the intended chain between archives and libraries to be established.

Unfortunately the MODS element set (approximately 80 components) is not in itself specific enough for some of the requirements of medievalists. It is particularly lacking in codicological information necessary for describing manuscripts from this period in sufficient detail. MODS does however, allow its element set to be extended when necessary: for the purpose of the CENDARI project this is done to incorporate a detailed set of manuscript description elements from the TEI (Text Encoding Initiative) [10].

The TEI is a long-established standard for encoding textual objects: because of its modular architecture and extensive set of elements it is often used for digital editions of manuscripts. One of its optional components is a detailed set of elements for describing the features of manuscripts: *the msDesc* (Manuscript Description) [11] set includes elements for such important facets as the physical descriptions of manuscripts, information on scripts used, decorations, bindings, layouts and their provenance in addition to detailed descriptions of their contents.

Including the TEI msDesc as an extension to MODS provides metadata records of sufficient detail to enable medievalists to incorporate these objects into their research while retaining the interoperability with library cataloguing practices allowed by the MODS schema. Some problems can arise with this strategy, however, owing to potential duplications and redundancies between the two schemas. In many cases, the same concept can reasonably be encoded in either schema: both, for instance, include elements for physical descriptions. These can readily be obviated by drawing up precise cataloguing guidelines that detail which schema should be used for each concept, so preventing ambiguities and redundancies.

3. The CENDARI Collection Schema

For the collection-level descriptions used particularly by the twentieth-century historian EAD was initially considered but found inadequate for the interoperability requirements demanded by the project. As stated earlier, EAD is modelled on the traditional paper finding aid and so is designed essentially as a way of encoding the information that would be found in such documents. Much of its architecture is, therefore, populated with textual fields designed to contain descriptive prose. These elements are relatively poor as mechanisms for interoperability as they are inevitably semantically broad and imprecise.

For the CENDARI project, a new schema was devised which offers the potential for a more precise method of referencing the components of a collection description and so making it more possible to link such a description to the wider information environment. This schema was constructed following a discussion with domain experts in archives who were asked to define the components that they considered central to their requirements. A total of fourteen such components (or facets) were defined:

1) collection description (identifiers/titles etc.);
2) holding institution;
3) subject coverage;
4) languages of materials;
5) bibliographies of related literature;
6) rights information;
7) contents of the collections;
8) source information;
9) dates;
10) relationships to external objects;
11) lacunae (gaps) in the collection;
12) impediments to using it effectively;
13) information on the collection's likely future availability;
14) information on the metadata record itself.

Many, but not all of these, have counterparts in EAD's element set: the exceptions to this are lacunae, impediments and information on the collection's future. Even where there is some degree of overlap between EAD and CCS elements, this internal structure of these is often very different owing to the divergent emphases of each schema.

This is most evident in the extensive use of XML attributes to provide semantically precise qualifiers to each facet. For instance, a lacuna in a collection can be described as presented in Table 1.

Table 1. Lacuna description.

```
<lacuna lang     = "en"
         type    = "missing component"
         typeURI = "http://cendari.edu/id/lacunatypes/missingcomponent"
        cause    = "mice"
        causeURI = "http://cendari/edu/id/lacunacauses/mice"
      coverageID = "cendari-sample-1-component1"
       startDate = "1923-02-02"
         endDate = "1924-12-12"
        calendar = "gregorian">
 <p>Years 1923-25 are missing as a result of being eaten by mice</p>
 </lacuna>
```

Table 2. Form of <relation> element.

```
<relation   type    = "part"
            typeURI = "http://purl.org/dc/terms/hasPart"
          target    = "item:3903456"
          targetURI = "http:/cendari.eu/id/item/3903456"
        coverageID = "cendari-sample-1-component1"/>
```

In addition to the textual description of the gap, which in EAD would be recorded in a generic <note> element, this element records the cause of the gap, its chronological boundaries (and the calendar in which these are recorded), the type of gap and the part of the collection in which it occurs (given by *coverageID*). This is a much richer set of information; more importantly, because it is encoded in discrete data components, it is amenable to machine-readable analysis and processing.

This schema provides a rich metadata set for describing collections, but is intended to form only part of a wider network of information. It is designed specifically to act as an 'intermediary' schema, that is a schema which is not necessarily intended as a final delivery mechanism for metadata, but as a mediator between other established schemas [12]. This is achieved partly by mapping the schema to its more established counterpart (in this case EAD) and by using its extensive linking facilities.

4. Establishing the linkages

Extending this capability beyond a single CCS record is made possible by the schema's extensive use of URIs (Universal Resource Identifiers). These are identifiers which precisely reference any concept or thing anywhere on the Internet, and form the basis on which the semantic interoperability of linked open data is built. Several sets of linkages are made possible in this way.

A primary linkage is to item-level records encoded in MODS. This may be achieved either from the CCS document to the MODS file or vice versa. In the former direction, linkages may be formed by using a <relation> element available in the CCS element set which allows any type of relationship to an external entity to be specified. For instance, to specify an item which forms part of the collection, the <relation> element may take this form as presented in Table 2 where *targetURI* records the URI of the MODS record for this item and *coverageID* contains the identifier for the part of the collection in which it is found. The linkage in the opposite direction is achieved by the use of a <relatedItem> element within MODS, which references the URI of the CCS file.

Beyond establishing these linkages, the CCS file can also be used to generate EAD files directly, so allowing the integration of CENDARI records with legacy data already encoded in that schema. As such, CCS operates as an 'intermediary' schema as outlined above. Using this technique allows the project to continue employing schemas which have embedded themselves in their respective communities (such as EAD) but to link them into a coherent whole, so reconciling to some extent their divergent metadata strategies.

A further level of integration may be achieved by employing the CCS schema to generate metadata for the Semantic Web. To achieve this, a simple transformation is written to produce RDF (Resource Description Framework) [13] 'triples', subject-predicate-object units of semantic information which form the atomistic components on which the Semantic Web is built. RDF triples function best when URIs are used for their constituent components, as these allow their precise semantic delineation in a form which should be unique throughout the internet. The consistent use of these URIs in the CCS schema makes the generation of these triples straightforward and allows the ready generation of RDF metadata. The overall set of linkages achieved in this way may be summarised schematically as presented in Fig. 1.

There are many reasons why using XML in this way, rather than encoding these linkages directly into RDF-based ontologies, may be more practical for a working, unified environment. The atomistic approach of RDF, in which each semantic component is encoded in a single subject-predicate-object 'triple', rapidly produces information networks of great complexity involving potentially thousands of triples when objects or collections of any size are involved. Maintaining such networks, and particularly transferring their constituent metadata between systems, is highly complex: for these reasons, using the readily-packaged XML syntax is the better option in working environments.

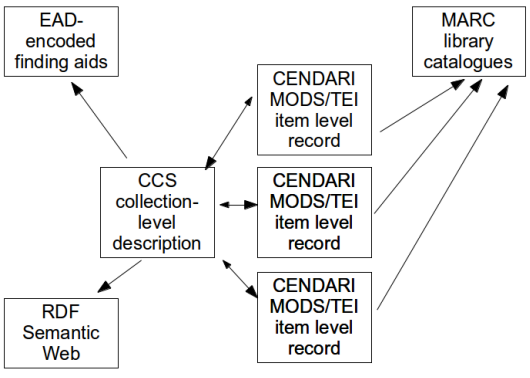

Fig. 1. Schematic of CENDARI linkages.

Conclusions

The imperatives of the digital eco-system have rendered the long-established divide between the archival and library worlds at best irrelevant and at worst a major impediment to research practices. The erosion of boundaries between research resources which has been made possible by the advent of digital technologies and is further realised by the Semantic Web requires a means of making joins across these borders while retaining the key advantages gained by established practices in both domains. The CENDARI project, in particular the CCS schema, should form a solid basis on which these joins can be made and eco-systems built.

It is because, most established schemas were not designed with linkages of this type as part of their functionality that it becomes necessary to employ mediating schemas of the type proposed by CENDARI. By employing these, and incorporating semantic linking features as their core design feature, it becomes possible to allow these sophisticated networks of components to be integrated into a coherent whole. In this way, a unity between the divergent strategies and methodologies of archives and libraries becomes a real possibility and the now obsolete divisions between the two can, at last, be discarded.

Acknowledgement

The research leading to this article has received funding from the European Union's Seventh Framework Programme [FP7/2007-2013] under grant agreement N 284432.

References

1. Harold Boley & Elizabeth Chang: Digital Ecosystems: Principles and Semantics, <http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=8914187>, accessed 2007.
2. <http://www.cendari.eu/>, accessed 2013 12 12.
3. Michel Duchein. Theoretical Principles and Practical Problems of Respect des fonds in Archival Science. – *Archivaria* 16 (1983) 64-82.
4. Anthony Panizzi. Rules for the compilation of the catalogue. – London: British Museum. Department of Printed Books, 1841.
5. <http://www.loc.gov/ead/>, accessed 2013 12 12.
6. <http://www.loc.gov/marc/>, accessed 2013 12 12.
7. <http://www.worldcat.org/>, accessed 2013 12 12.
8. <http://www.cendari.eu/wp-content/uploads/APEX-customizingEAG_Medves.pdf], accessed 2014 01 30
9. <http://www.loc.gov/standards/mods/>, accessed 2013 12 12.
10. <http://www.tei-c.org/index.xml>, accessed 2013 12 12.
11. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-msDesc.html>, accessed 2013 12 12.
12. Richard Gartner. Intermediary schemas for complex XML publications: an example from research information management. – *Journal of Digital Information* 12 (2011).
13. <http://www.w3.org/RDF/>, accessed 2013 12 12.