Cultural Heritage Collections : From Content Curation to Semantic Services and the Semantic Web

Luxembourg
Muriel
Foulonneau

Collecting and aggregating resources

The action of collecting intentionally resources for a specific purpose, to organize them for personal use or for a particular audience, creates meaning. It has in itself a value which can be shared and used, just like descriptions or annotations to enable retrieval and manipulation of resources.
All the actions of content creators, managers (e.g., librarians or museum curators) and users may be thought of through the concept of *collection*. Content creators often create a set of resources. Managers collect resources for a particular audience (e.g., the manuscripts of James Joyce or a collection of resources to support researchers in high energy physics). Users collect resources and organize them in their environment.
Nevertheless, resources are most often described at item level and more rarely at collection level. The standards for the description of collections are not as stable and consistently used as standards for item level descriptions. As a result, while the work around resources is conditioned and driven by implicitly or explicitly created collections, those are often not represented in resource management systems.
Recent advances in online services have emphasized interactions with users who can create their own collections and share them in Web 2.0 applications. Semantic representations of resources have also led to widening our conception of valuable resources because anything can be a resource of equal importance, a picture, a book, a city, an idea, and therefore also a collection.
This article provides an overview of collection description practices, the integration of collections in different services, the metadata models for collection level descriptions, and the representations of collections on the Semantic Web.

Different traditions: the computer science domain, museums, libraries, archives

There are multiple definitions of a collection. The criteria used to aggregate resources in a collection are extremely different according to sectors and curatorial traditions. The traditional interpretation of library collections is associated with tangibility, ownership, a user community and a service (Lee, 2000). The Conspectus methodology was even created by the Research Libraries Group in the U.S. to assess the strengths and weaknesses of research libraries collections and therefore engage libraries in developing complementary collections on a regional or national basis, for instance. Andy Powell (1998) has described the diversity of traditions for collection definition: *almost always, the collections of 'archives' delineate themselves: they relate normally to a specific person or institution. The collections of 'libraries', on the other hand, should be de-lineated by the purpose for which the library exists: by the information needs of their user popu-lations. In contrast, the collections of 'museums', are - again - delineated somewhere between those two extremes. They can perhaps best be conceived as a bridge between the collecting desires and interests of specific people or institutions; and the information needs - in the widest sense - of those who might use the resulting collection.*

In addition, in the digital environment, a collection has long represented a search target, i.e., a set of items available through a single access service (e.g., Lagoze et al., 1998). This definition of collections is derived from the services implemented on top of the resources rather than on the content itself. It is then possible to analyze for instance the terms most often used in a particular search target. However, OAI-PMH repositories have showed that in many cases, a particular institution sets up a service on top of very heterogeneous contents. The University of Michigan OAI repository provides access to more than 300 collections, including journals, poetry, pictures. In this case, the OAI repository was organized according to collections defined on content criteria, rather than on service criteria.

Several studies have been conducted on the criteria that collectors used to define a collection. The criteria used to divide the content of an OAI-PMH access interface into OAI sets are very diverse. They include issues for journals, departments and research centres in institutional repositories, subject and publication status in e-print archives, and finally content related criteria in cultural heritage repositories (Foulonneau et al., 2006). Renear et al. (2008) defined collections according a *curatorial intent*. Indeed, resources can be aggregated automatically because they represent the same work (e.g., a PDF and a Word version of an article). Although it is possible to create an aggregation with these items, it only becomes a collection if there is an intent to apply similar curatorial procedures to the aggregated items. Studies have focused on personal collections (Rousseaux et al.,2008 ; Beagrie, 2005) and the way in which people either accumulate things for building collections, as in a child's instinct for accumulating and organizing (Rousseaux et al., 2007), or only organize their personal information environment, i.e., their computer (Boardman et al., 2004). Then personalized retrieval systems can be elaborated on top of this organization, for instance. Finally, Stvilia et al. (2009) studied collections created by users on a community Website, i.e., Flickr. They found that individual users define collections mainly according to activities (including an event for instance), places and persons, and less often by artistic or photographic techniques, things, time, quality, or even randomly (e.g., when people compete to have the highest number of items included in their collection). More and more community Websites allow the sharing of personal information and personal collections (e.g., Merlot, the Multimedia Educational Resource for Learning and On-line Teaching[2] and Google Books).

The variety of collection definitions and of collecting practices illustrates of building a system which is based on collections rather than or in addition to items. Nevertheless, in order to enable the creation, storage, management, sharing and distribution of collections, applications must be not only item centered but also collection centered, and allow the manipulation of collection level descriptions in addition to item level descriptions.

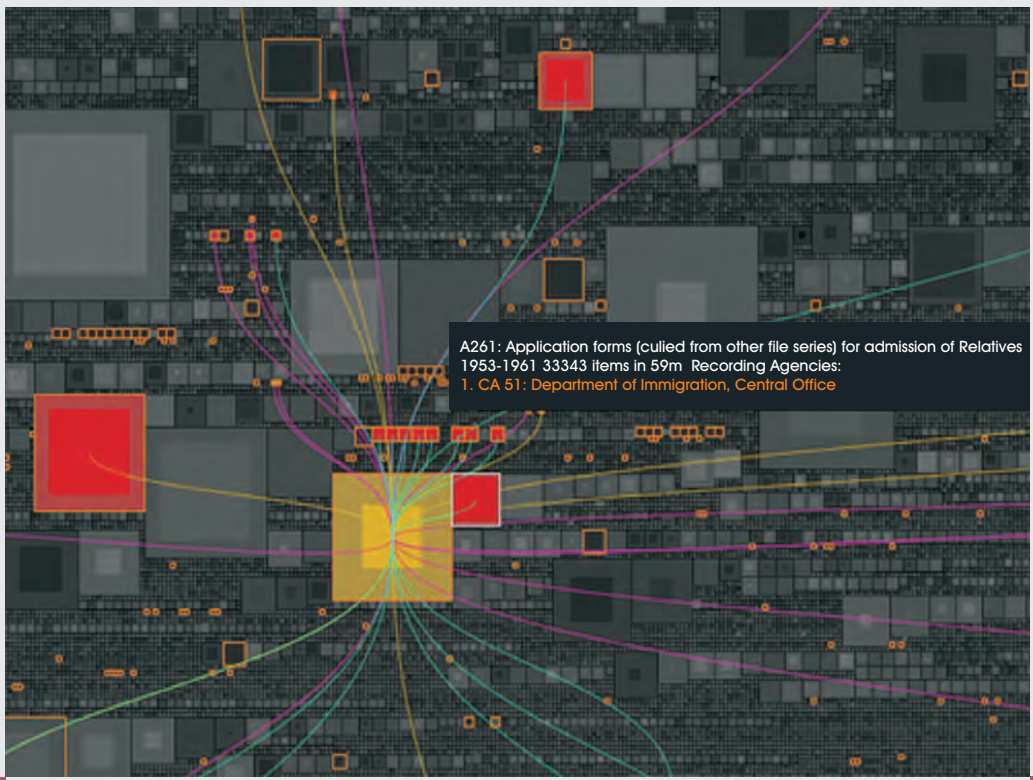
Collection registries and the creation of an information landscape

The creation of collection level records, either as a replacement for or as a complement to item level descriptions, was encouraged in such projects as the New Opportunities Fund in the UK[3], the Patrimoine numérique Website in France[4] and many initiatives around the world (Foulonneau et al., 2003). Collection registries (Entlich, 2000) have then been developed on a large scale. The

MICHAEL portal at European level[5] and the IMLS Digital Collections and Content (IMLS DCC) in the US[6] gather together descriptions of digitized collections of heritage resources. Both have been used as a basis for the harmonization of item level resource descriptions in the digital environments and the creation of access services at both item level and collection level.

These registries use a definition of collections provided by the metadata creators. As a result, collections are very heterogeneous. Metadata creators use the traditions of museums, libraries, archives and others. In addition, collections are defined at various levels of granularity. A 100-item collection is represented next to a 50.000-item collection. Michael Heaney (2000) described *information landscapes* as the organization and access to information at various degrees of granularity and specialization, through the definition of collections.

Attempts to mix different levels of granularity in the same portal raise major issues, such as the constraints created by various user expectations on what they will find on a given portal. Nevertheless, the lack of consideration for collections also creates challenges, for instance because the item level descriptions in certain collections are too similar and create biases in the digital library system (Foulonneau, 2007). The possibility to collect both item level descriptions and collection level descriptions can however create opportunities, to compensate for incomplete records or add context to decontextualized records fetched in a third party aggregation such as an OAI-based portal (Foulonneau et al., 2005).



Representation of collections in the Visible Archive Project[7]

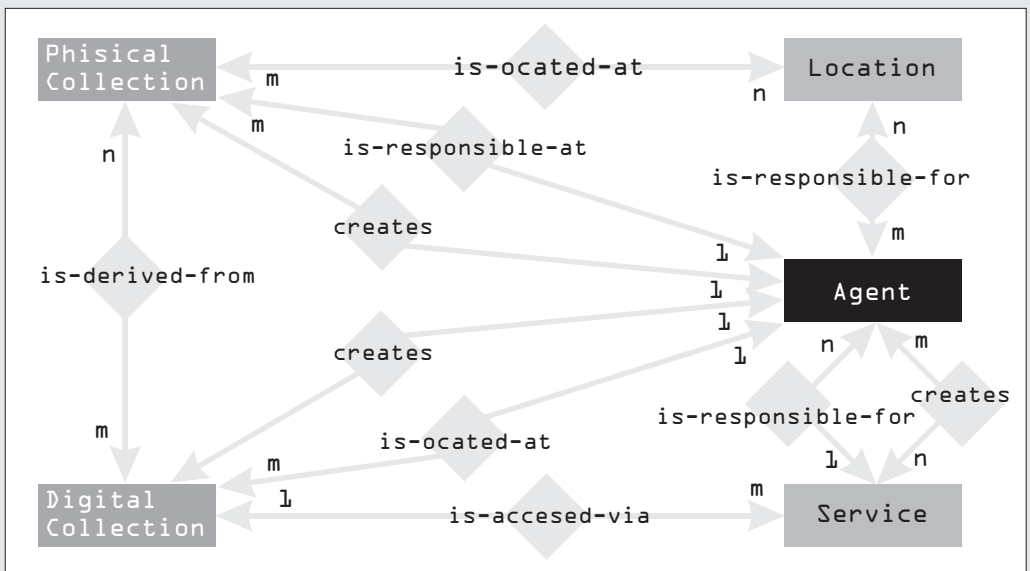
1

Moreover, visualization technologies illustrate new possibilities to represent an information landscape and to provide a sense of a collection's strength and the parts of a collection that match a particular query (e.g., the Visible Archive Project and the ArchivesZ[9] project) (Figure 1). Urban et al. (2010) have worked on the creation of dashboards to visualize resources through their aggregation into collections. These technologies can help extract new information out of the representation of collections.

Describing digital collections

In order to support the description of heterogeneous collections, registries need to create a collection description model for all types of digital cultural content, whether from libraries, archives, museums or archaeological sites, for instance. Certain metadata models, such as MARC, were mainly conceived for item level descriptions but also used for collection level descriptions. MODS guidelines[10] specify a particular use of certain elements in the context of a collection level description, e.g., *extent*[11]. VRA-Core[12] for visual resources explicitly defines a model for collection level descriptions in addition to item level and work level descriptions. EAD[13], which is used in the archival domain, includes both collection and item level descriptions, as well as information on the organization of different collections.

The Research Support Libraries Programme[14] in the UK designed a simplified collection description format for libraries. It then inspired the work conducted for the Dublin Core Collections application profile[15], and thereafter the IMLS DCC collection description format and the metadata format used for the MICHAEL collection registry (Figure 2). Collections were even integrated into the CIDOC-CRM model, i.e., the Conceptual Reference Model for cultural heritage resources (Lourdi, 2009).



The MICHAEL-EU Dublin Core Application Profile Data Model (simplified version)[16]

Indeed, the initiative to create a simple generic description format for collections was conducted under the auspices of the Dublin Core Metadata Initiative. The Dublin Core Collection application profile released in 2007[17] added specific terms to the Dublin Core vocabulary, such as *dcterms:accrualPolicy* or the Custodial History of a collection (*dcterms:provenance*). In addition, different roles were distinguished, e.g. the Collector versus the Owner of the collection. Certain properties are specific to the aggregation *per se* such as the Date Collection Accumulated, others to the items that are part of the collection such as the Date Items Created to represent that a collection of 16th century objects for instance was collected in the 1920s.

The coexistence of properties specific to the aggregations with others specific to the items makes the propagation of values from collections to items very difficult. For instance, if a collection has both a *dcterms:type StillImage* property and a *dcterms:type Text* property , it is impossible to infer which of the items are of type *StillImage* and which ones are of type *Text*. It is however likely that none is of type Audio. Foulonneau et al. (2005) and Renear et al. (2008) in particular have studied these mechanisms and the way in which they can be used to enrich digital library services.

There are therefore specific challenges to the representation of both collections and items in an information system. Collection level description models have been designed to support cross-domain access services for libraries, archives, and museum resources. Nevertheless, on the Semantic Web, different types of resources co-exist, including items, collections, and even metadata. The accurate representation of each resource is even more important to ensure that humans can find their way in huge amounts of heterogeneous data published on the Semantic Web.

Representing collections on the Semantic Web

More and more, cultural institutions are becoming aware of the potential of new modes of publication of their data, through the web architecture rather than repository or service oriented architectures. On the Semantic Web, data are published as structured data, either inside web pages (e.g., RDFa) or as datasets (e.g., RDF-XML). They are accessible through browsing (dereferenceable URIs), through query interfaces (SPARQL endpoints), or even through data dumps updated on a regular basis. Data are not only readable (like XML) but also interpretable by applications, thanks to RDF formatting and the documentation of ontologies (or models) on which they are based. On the Semantic Web, resources are represented by URIs. A document is represented by a URI, but so are people www.dbpedia.org/resource/Albert_Einstein , places www.dbpedia.org/resource/Italy, and concepts http://id.loc.gov/authorities/subjects/sh2002000569.html#concept (the Semantic Web concept as defined by the Library of Congress).

Early on, the Semantic Web community felt the need to describe data collections. Instead of describing collections of external things, these collections are datasets, i.e., metadata or structured data collections. The VoID (Vocabulary of Interlinked Datasets) model supports the description of Linked datasets. It includes properties useful for the processing of a dataset in an application, such as a SPARQL endpoint to access the dataset, as well as the number of triples

included in the dataset. VoID can typically be used to describe metadata of cultural heritage resources published as Linked Data on the Semantic Web.

Several initiatives have proposed representations of collections in semantic environments. W3C POWDER[19] enables the definition of the criteria for the construction of a dynamic collection for instance all Flickr pictures that have been assigned a tag *architecture*. Although this mechanism can help organizing resources, it carries the risk that the collection properties will not be valid when the collection dynamically grows or changes focus over time.

In contrast, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[20] defines a RDF-based model to represent aggregations of resources as enumerations. Resources are included by reference, so that anybody can publish on the Semantic Web a map of a collection. This map contains URIs of items distributed over the entire Web. The OAI-ORE specification advises the use of core descriptive metadata properties, such as Dublin Core metadata.

OAI-ORE even defines the concept of Proxy resource to enable the metadata creators to define collection-specific item level metadata. For instance, a caricature by Honoré Daumier can be included in a collection about the representation of lawyers and in another collection about 19th century French caricaturists. In the first case, its topic and description should reflect particular traits of lawyers emphasized by Daumier, such as venality, whereas in the second case, the particular description should rather reflect the artistic characteristics of the Daumier's style. Whereas both types of metadata are relevant, they may have different importance in the context of different collections. Moreover, in certain cases, characteristics have to reflect the specific relation between items in a collection, for instance the fact that a resource has been used in conjunction with another one in the context of a particular collection (e.g., a collection of research papers in the context of a conference). The mechanism of proxies offered by OAI-ORE aims to create different views of a resource in particular contexts. This is particularly interesting for the Semantic Web, where different actors can make statements (i.e., metadata) about resources, specialists and non specialists, from multiple institutions or individuals. Certain metadata created for a particular work environment (e.g., personal annotations) may only have value in their original context of creation.

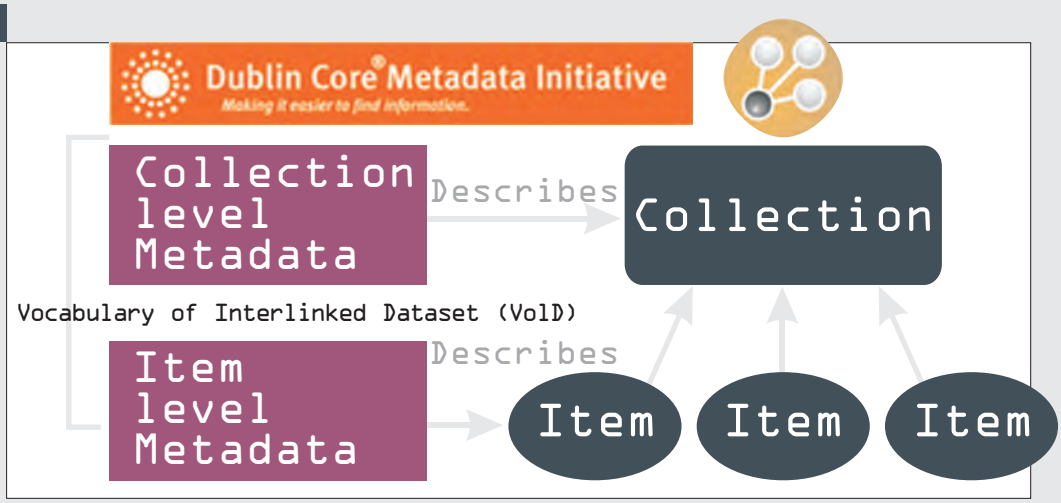
Whereas both W3C POWDER and OAI-ORE have been created for different types of aggregations, it is possible to use them in order to represent collections defined according to a curatorial intent. These models allow describing the structure of aggregations as compound objects (like METS or MPEG-DIDL) as well as supporting the description, re-composition, curation and manipulation of cultural heritage collections.

Conclusion

The perspective on collections has therefore been widened to include not only collections as a way for institutions to organize and manage their resources but also collections created potentially by third parties, who may cluster resources according to different criteria and at different stages of the resource lifecycle from its creation to its consumption and sharing.

Whereas collections for cultural heritage resources used to be defined according to cultural institutions' curatorial traditions, the frontiers between cultural institutions and other actors, including their users, is more and more blurred. Collections are built according to other logics and

3
Complementarity between standards for the representation of heritage collections on the Semantic Web



perspectives. Several cultural institutions have started using Flickr and similar sites to publish their resources. Users can create their own collections.
Moreover, tools for the publication of data on the Semantic Web are growing rapidly. This allows envisioning that different actors can create and publish collections, while only publishing a "map" of that collection (e.g., in OAI-ORE) and its description, without collecting the content itself. The Dublin Core Collection application profile or a derivative implementation can be used for describing the collection content. VoID can also be used for the description of the metadata sets published on the Semantic Web. Figure 3 illustrates the complementarity between core vocabularies to support services for cultural heritage items and collections in digital library systems as well as on the Semantic Web. Nevertheless, the co-existence of multiple levels of granularity and of extremely heterogeneous types of resources remains a challenge for which new visualization tools and smart services designed for large datasets provide new answers.

Notes

1. http://quod.lib.umich.edu/cgi/o/oai/oai?verb=ListSets
2. http://taste.merlot.org/personalcollections.html
3. www.enrichuk.org
4. www.numerique.culture.fr
5. www.michael-culture.org
6. http://imlsdcc.grainger.uiuc.edu/
7. www.visiblearchive.blogspot.com/
8. www.visiblearchive.blogspot.com/
9. www.archivesz.com
10. www.loc.gov/standards/mods/userguide/
11. www.loc.gov/standards/mods/userguide/physicaldescription.html#extent
12. www.loc.gov/standards/vracore/
13. www.loc.gov/ead/
14. www.rslp.ac.uk/
15. www.dublincore.org/groups/collections/collection-application-profile/
16. www.ukoln.ac.uk/metadata/michael/michael-eu/dcap/
17. www.dublincore.org/groups/collections/collection-application-profile/
18. http://imlsdcc.grainger.uiuc.edu/
19. www.w3.org/2007/powder/
20. www.openarchives.org/ore/

References

Beagrie Neil. *Plenty of Room at the Bottom? Personal Digital Libraries and Collections.* D-Lib Magazine, 2005 vol. 11 no 6.

Boardman Richard, Sasse Angela M. *Stuff Goes Into the Computer and Doesn't Come Out: a Cross-tool Study of Personal Information Management. In: CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York: ACM Press, 2004, p. 583-590.

Lagoze Carl, Fielding David. *Defining Collections in Distributed Digital Libraries.* D-Lib 1998 vol. 4 no 11. www.dlib.org/dlib/november98/lagoze/11lagoze.html [Viewed 10 March, 2011]

Entlich Richard. FAQ: *Is There a Good, Comprehensive Catalog of Web-accessible Digitized Collections Available on the Internet?* RLG DigiNews 2000 vol. 4, no.6.

Foulonneau Muriel, Arms Caroline, Shreeves Sarah L. *Sharing Resources by Collection: OAI Sets and Set Descriptions.* Digital Library Federation Spring Forum, Austin, TX 2006. www.new.diglib.org/spring2006schedule [Viewed 10 March, 2011]

Foulonneau Muriel. *Report Analysing Existing Content. Minerva Project.* www.minervaeurope.org/intranet/reports/D3_1.pdf [viewed 10 March, 2011]

Foulonneau Muriel et al. *Using Collection Descriptions to Enhance an Aggregation of Harvested Item-level Metadata.* JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. 2005 p. 3241.

Foulonneau Muriel. *Information Redundancy Across Metadata Collections.* Information Processing & Management. 2007 vol. 43 no 3 p. 740-751.

Heaney Michael. *An Analytical Model of Collections and their Catalogues.* Bath: UKOLN, 2000. www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf [viewed 10 March 2011]

Lee Hur-Li. *What is a Collection?* Journal of the American Society for Information Science, 2000 vol. 51 no 12, p. 1106-1113.

Lourdi Irene, Papatheodorou Christos, Doerr Martin. *Semantic Integration of Collection Description: Combining CIDOC/CRM and Dublin Core Collections Application Profile.* D-Lib 2009 vol. 15 no 7-8. www.dlib.org/dlib/july09/papatheodorou/07papatheodorou.html [viewed 10 March 2011]

Powell Andy. *Collection Level Description - a Review of Existing Practice.* eLib, 1998. www.ukoln.ac.uk/metadata/cld/study/ [viewed 10 March 2011]

Renear Allen H. et al. *Collection/Item Metadata Relationships.* Berlin: Humboldt University, 2008 p. 80-89 www.edoc.hu-berlin.de/conferences/dc-2008/renear-allen-80/PDF/renear.pdf [viewed 10 March 2011]

Rousseaux Francis, Bonardi Alain. *Benefiting from Piaget to improve our Collections Browsing Tools?* IADIS Applied Computing 2007 Salamanca, 18-20 February 2007. www.iadis.net/dl/final_uploads/200702C051.pdf [viewed 10 March 2011]

Rousseaux Francis, Bonardi Alain. *Parcourir et constituer nos collections numériques.* CIDE 2008 http://lodel.irevues.inist.fr/cide/index.php?id=269. [viewed 10 March 2011]

Stvilia Besiki, Jörgensen Corinne. *User-generated Collection-level Metadata in an On-line Photo-sharing System.* Library & Information Science Research. 2009, vol. 31 no 1, p. 54-65.

Urban Richard J, Twidale Michael B, Adamczyk Piotr. *Designing and Developing a Collections Dashboard.* In: Jennifer Trant and David Bearman (eds). *Museums and the Web 2010: Proceedings. Toronto: Archives & Museum Informatics.* 2010. www.archimuse.com/mw2010/papers/urban/urban.html#ixzz1K3J7z292 [viewed 25 March 2011]