Digital Collection Contexts

iConference 2014 Workshop Report

Organizers

Carole L. Palmer,[1] Antoine Isaac,[2] Karen M. Wickett,[3]
Katrina Fenlon,[4] Megan Senseney [4]

Invited Panelists

Hur-Li Lee, Karen M. Wickett, Martin Doerr, Carlo Meghini,
Amy Rudersdorf, Sheila Anderson, Shenghui Wang, Paul Clough

Contributors

Sheila Corrall and Angharad Roberts

---

[1] Information School
University of Washington

[2] Europeana Foundation

[3] School of Information
University of Texas at Austin

[4] Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

CIRSS

CIRSS Technical Report 201503-01

March 2015

Table of Contents

1. Introduction

The "Digital Collection Contexts: Intellectual and Organizational Functions at Scale" workshop was held March 4, 2014, at the iConference in Berlin, Germany. The aim was to unite a community of faculty, students, system designers, and developers interested in digital collections, particularly in the context of cultural heritage aggregations. Organized by a team from the University of Illinois, the Europeana Foundation, and the University of Texas at Austin, the one-day workshop brought together an international group of experts representing diverse threads of current research and development to engage on the role of collections in the digital environment and to identify new directions for inquiry.

As large-scale aggregations of digitized, cultural heritage collections from libraries, museums, and archives gain critical mass, they demonstrate immense potential value for the public and for scholarly users. When collections are treated as a functional element in aggregations, they become useful for the organization, description, and retrieval of items, as well as for the comprehension, visualization and evaluation of the aggregation as a whole. As such, the collection -- as both concept and construct -- lies at the heart of numerous existing but disparate research threads. Trends in interoperable content and open data raise important questions on how to represent complex objects, curated and dynamic collections, and context in ways that benefit users and collecting institutions. The workshop was designed to address four goals:

- Broaden the conversation across an international community
- Further the research and development agenda for digital aggregations
- Relate conceptual advances to implementation goals
- Identify realistic approaches for collection representation, contextualization, and interoperability at scale

Each of the invited panelists submitted short position papers prior to the event. During the workshop, the panelists and participants considered collections in relation to the information needs of scholars, roles of cultural institutions, and international interoperability through a set of presentations, break out sessions, and full group discussion. This report on the Digital Collection Contexts workshop compiles the position papers and includes synopses of the presentations by the authors and ensuing discussions. We first present background on the convergence of two projects that inspired the workshop, and an overview of the workshop agenda and participants. An introductory position paper fully introduces the workshop and remaining position papers.

2. Background

Europeana brings together the digitized content of Europe's galleries, libraries, museums, archives and audiovisual collections. The Europeana prototype, funded by the European Commission's eContentplus program under i2010, was first launched in 2008. Today it is one of the most successful implementations of a large-scale cultural heritage aggregation in the world, providing access to 39.2 million items from

more than 2,500 institutions representing 36 European countries.[1] The Europeana Data Model (EDM) was developed to replace the initial Dublin Core-derived data model, known as Europeana's Semantic Elements (ESE).[2] EDM addresses outstanding issues related to domain-specific metadata standards in a cross-domain environment, participation in a Linked Open Data environment, and adding value to digital cultural heritage objects through data enrichment. EDM has been highly influential in the development of the recently launched Digital Public Library of America.

From 2003 to 2013, the IMLS Digital Collections and Content (IMLS DCC) initiative developed one of the largest and most diverse cultural heritage digital aggregations in the country, based on several phases of research and development on metadata harvesting, enhancement, and interoperability; collection and item-level metadata dynamics; aggregation workflows; subject access; content evaluation; and metasearch. While no longer growing, the current IMLS DCC aggregation now provides integrated access to more than 1,700 collections with over 1 million items, representing nearly 1,500 cultural heritage institutions, large and small, from 46 states.[3] The team has made significant advances in national-scale interoperable metadata (Shreeves et al., 2005), and development has adhered to principles derived from research on users of cultural heritage collections and the scholarly use of digital resources (Palmer, Teffeau, & Pirmann, 2009; Palmer, Zavalina, & Fenlon, 2010). Adapting research library collection assessment strategies for national digital collection development, the project also demonstrated the importance of policy-driven growth and the viability of extending access to additional primary and secondary sources, including integrating IMLS DCC content into Flickr photostreams. Perhaps most importantly, the aggregation approach retained the collection contexts and subject coherence that has proven vital to how scholars explore and interact with cultural heritage materials.

The IMLS DCC and Europeana initiatives share many common principles and processes. They bring together similar kinds of content from a range of digital cultural heritage institutions, and the basic mode of aggregation is the same: metadata are centralized and indexed providing integrated access to descriptions and thumbnails that link back to the digital object at the host data providers. Upon exploring opportunities for collaboration with Europeana, the IMLS DCC team identified the Europeana Data Model (EDM) as the best area for engagement, since it is more advanced than the DCC approach in terms of applicability to semantic web technologies. However, EDM does not accommodate explicit representation of collections. Therefore interaction around EDM also benefited Europeana's interests in increasing coherence and functionality through collection representation within their expansive aggregation.

Beginning in May 2011, representatives from the two initiatives explored potential synergies, first at a one-day workshop held in Crete in conjunction with the European Semantic Web Conference. A year later, partners convened a three-day working meeting at the University of Illinois. The group included the CIRSS researchers in the Collections and Curation core area and three key Europeana representatives: Antoine Isaac (Europeana), Carlo Meghini (Italian National Research Council), and Martin Doerr (Institute of Computer Science, Foundation for Research and Technology – Hellas). The three-day

[1] See http://statistics.europeana.eu/welcome for more statistics on Europeana.
[2] See http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation for documentation on the Europeana Data Model.
[3] http://imlsdcc.grainger.uiuc.edu/

meeting resulted in a collaborative white paper entitled "Modeling Cultural Collections for Digital Aggregation and Exchange Environments," which was publicly released in Fall 2013 (Wickett et al., 2013). The white paper provides initial recommendations for developing a collection description and representation model that is compatible with the Europeana Data Model (EDM) and has important implications for both aggregations and interoperability between them, and for DPLA and other national library initiatives.

3. Workshop Overview

Following release of the white paper, CIRSS and Europeana organized an internationally scoped workshop on Digital Collection Contexts at iConference 2014, the first meeting of the iSchools organization to be held outside North America. Workshop organizers included Carole L. Palmer (CIRSS), Antoine Isaac (Europeana), Karen Wickett (School of Information, University of Texas), and Megan Senseney (CIRSS). The workshop was designed to provide a forum for international engagement on this important topic and provide iSchools the opportunity to build a community around established strengths in research on collections in digital environments.

The workshop was divided into a morning session on Conceptual Foundations and an afternoon session on Practical Implications. Each session included a panel of experts from European and North American iSchools and projects developing large-scale digital cultural heritage collections. Prior to the event, panelists submitted brief position papers to the workshop organizers which were then distributed to registered participants along with the white paper. These position papers helped seed topics for breakout discussions that were organized after each panel.

Palmer moderated the morning session on Conceptual Foundations. Panelists included:

- Hur-Li Lee, School of Information Studies, University of Wisconsin-Milwaukee
- Karen Wickett, School of Information, University of Texas at Austin
- Martin Doerr, Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas
- Carlo Meghini, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche

Isaac moderated the afternoon session on Practical Implications. Panelists included:

- Amy Rudersdorf, Digital Public Library of America
- Sheila Anderson, King's College London
- Shenghui Wang, OCLC Research
- Paul Clough, Information School, University of Sheffield

In total, 38 participants registered for the workshop, hailing from 15 countries on four continents. Panelist bios are included in Appendix A, and a complete list of registered participants is included in Appendix B.

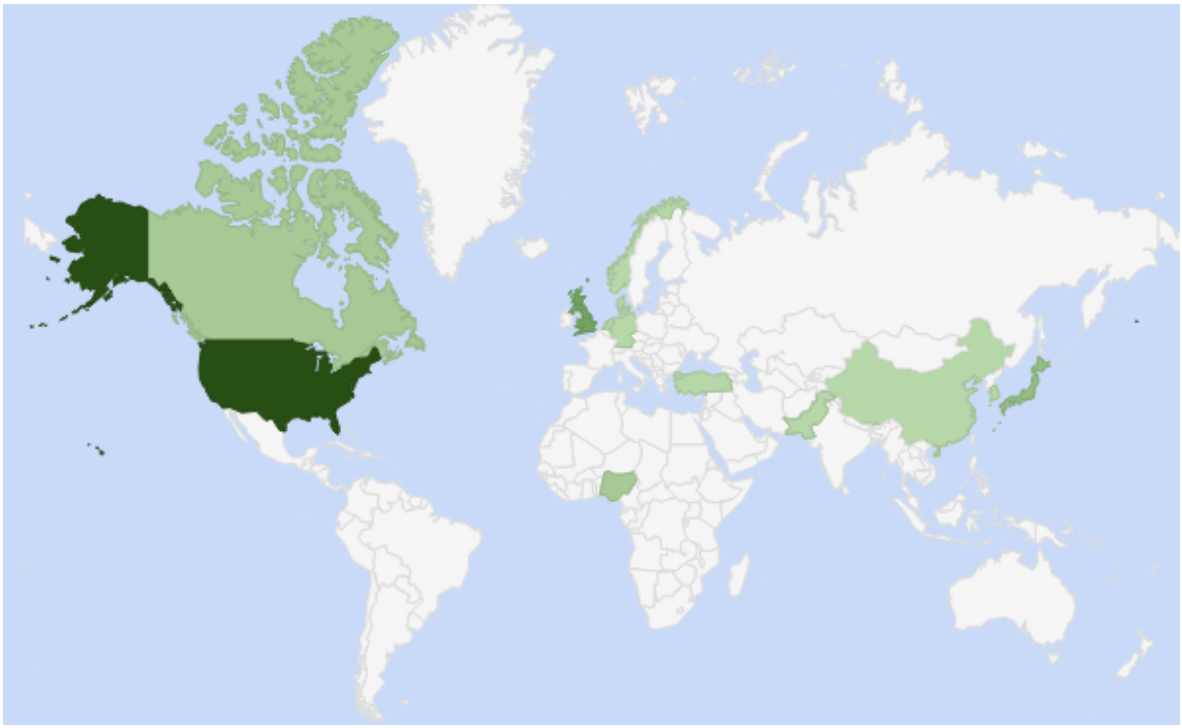

Figure 1. Map of participant demographics by country of residence

On digital collection contexts

Carole L. Palmer[*], Antoine Isaac[**] and Megan Senseney[*]

[*]Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
[**]Europeana Foundation

Building collections has always been central to the mission and functions of libraries, museums, and archives of all kinds and sizes. In fact, these types of institutions are now frequently grouped together and referred to as "collecting institutions." Many of these institutions are distinguished by their unique "special collections." Their collections are often aligned with service communities or promote certain genres or types of artifacts or works. Collections frame a range of institutional operations, as priorities are set and resources are allocated around them. Perhaps most importantly, collections are structures that provide organizational and intellectual contexts that direct how content is encountered and interpreted. We know that scholars, in particular, place high value on the context provided by intentionally, often expertly, grouped materials, and by the associated documentation recording the aims, criteria, history, and other information about a collection.

"Collections" are prevalent and important, as a conceptual construct but also as very real constructions. And while the characteristics described above relate to both analog and digital collections, in the digital realm collections can easily lose their presence. This has been evident to the organizers of this workshop in our respective initiatives aggregating digital collections across the U.S. and Europe, where we have been essentially collecting collections, to provide a single point of access to content from a multitude of institutions. At the same time, digital collections and aggregations hold great potential for exposing and exploiting relationships among materials distributed around the world, for an international audience of institutions and individuals to create new collections not yet imagined. This set of position papers is the beginning of a broader dialogue for exploring both the problems and potentials of retaining digital collection contexts at the international scale.

The first set of papers offers very different but complementary views on the conceptual and representational aspects of collections. What are collections; what is their essence, and why do they matter? How do we represent collections, formally, to retain them, utilize them, but also to innovate with them? It is important to note that our assertion in convening this workshop – that

collections are, in fact, context, was a central theme in Hur-Li Lee's 2000 paper, "What is a Collection?" [2]. Here, she reflects on how the collection concept has been treated in our field and the variable ways collections are conceived. She outlines three basic dimensions of collections – characteristics, functions, and representation and organization – as a frame by which to exact the fundamental nature of collections. Another fundamental feature emphasized by Martin Doerr, unity criteria, is an idea that we may readily intuit but rarely make explicit about collections. This concept stands to make a significant analytical contribution to how we understand what differentiates a collection from any grouping of materials, and the meaningful relationships among different collections.

A solution to the loss of collection context in large-scale digital aggregations lies in Lee's "representation" dimension. Karen Wickett takes a serious look at one framework for contextual information [1] to assess the match with our proposed approach to collection modeling [3]. In doing so she opens a path to further clarifying the elusive character of collections and context. Carlo Meghini offers further formalization in recognition that collections are more than containers of objects; they have intension. He demonstrates collection modeling, expressing a collection's purpose as a predicate symbol, and reminds us that interoperability hinges on appropriate ontological choices.

Indeed, investigating the foundational frameworks for modeling collections and their contexts is crucial. Yet there are additional practical problems faced by providers and users of collection-based services. The first of these is gathering collection data: many digital libraries focus on harvesting information on individual objects, sometimes neglecting to treat their collection context as first-order information resources. Compounding the problem is the frequent lack of structured, machine-readable data about collections. Often the institutions that host the collections have incomplete data, either because it exists as textual information, not suited for machine services, or because it has simply never been produced. In this case collections fail to exploit their contextual potential and merely exist as simple containers of individual objects.

These two aspects naturally hinder the provision of collection-aware services, and call for action from aggregation platforms like Europeana and the Digital Public Library of America (DPLA). Amy Rudersdorf presents the specific situations encountered by DPLA with respect to harvesting collection data. A key motivator for properly tackling these issues is to provide practical, user-focused requirements for the establishment of collection-aware services. Such services are necessary to meet the needs of digital humanities researchers, as well as other researchers who rely on access to cultural heritage content. The CENDARI project, presented by Sheila Anderson, aims at gathering descriptions of collections that are relevant for medieval history and World War One studies.

Finally, Paul Clough and Shenghui Wang identify methods to employ automatic techniques for processing and enriching collection metadata. This can be pursued as part of a general effort to improve users' access to collection content through (meta)data enrichment and object linking, as presented by Clough for the PATHS project. Shenghui Wang presents an initiative to develop object clustering for use in the case of heterogeneous datasets, where even the

boundaries of collections have been blurred.

Together, this set of position papers is a step forward in substantiating the role of digital collections as coherent, intentional, contextual objects of value to institutions and scholars, and in laying out key practical challenges of implementing systematic and functional collection description and representation. Through the workshop we are seeding broader international engagement and exchange to build the research and development agenda necessary for a future with interoperable international digital collections and aggregations that retain the richness of collection contexts.

References

[1] Lee, C. (2011). A framework for contextual information in digital collections. *Journal of Documentation, 67*(1). doi:10.1108/00220411111105470.

[2] Lee, H.-L. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

[3] Wickett, K.M., Isaac, A., Fenlon, K., Doerr, M., Meghini, C.L., Palmer, C.L, & Jett, J. (2013). Modeling cultural collections for digital aggregation and exchange environments. CIRSS Technical Report 201310-1, University of Illinois at Urbana-Champaign. Retrieved from http://hdl.handle.net/2142/45860.

4. Conceptual Foundations

4.1. Position Papers

The notion of collection: A retrospective overview

Hur-Li Lee
School of Information Studies
University of Wisconsin-Milwaukee, USA

"Collection" is among the foundational concepts of library and information science (LIS). Like many others in this foundation, the concept, as well as the term representing it, was taken for granted for a long time. Information scientists began pondering the definition, essential characteristics, and functions of collections only after the emergence of digital libraries, mostly in view of the capabilities and challenges introduced by advanced technology in the new information environments. In the past 25 years or so, the notion of collection has indeed been increasingly clarified and expanded, enhancing collection functionality and improving information retrieval. At this workshop on Digital Collection Contexts, it seems necessary and important for us to repeat the same question asked by Lee, "What is a collection?" [1], and review the historical development of the notion of collection.

The term "collection" has many connotations depending on the context. Loosely defined, it may refer to any objects grouped together. In China, archaeologists have excavated oracle bones used, collected, and stored together thousands of years ago. Anthologies of poems as well as the joint contents of any library are commonly referred to as collections. Then, do we also refer to a group of websites as a collection when the gathering is automatically conducted by a search engine on a moment-by-moment basis? If the answer to this question is "yes" and we reject the need for a formal and rigorous definition of collection, what are the implications of this approach? On the other hand, if the definition embodies a number of fixed requirements, how will we deal with changes brought about by evolving technology or human information needs?

LIS is an applied social science, and its nature necessitates theory building and system development with consideration to applicability and functionality in social contexts. Under such an overall frame and premise, I propose to examine the notion of collection in three dimensions: (1) characteristics and elements of a collection, (2) objectives and functions of a collection, and (3) the representation and organization of a collection. The first concerns a collection's substance and the second deals with its functionality. Naturally, the last is indispensible, for it is difficult, if not impossible, to develop a collection to realize its intended characteristics and functions without an effective organizational scheme in place.

References

[1] Lee, H.-L. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

Is collection modeling contextual modeling?

Karen M. Wickett
School of Information
University of Texas at Austin

Efforts to model and describe collections in digital aggregation systems are often driven by the desire to take advantage of the contextual information supplied by collection membership. Recently, a collaborative study group has examined the potential roles for collection-level information to enhance access, stewardship, and interpretation of resources in digital aggregations, and has derived requirements for the representation of collections to support these roles [2]. These requirements are the basis for a recommended extension of the Europeana Data Model (EDM)[1] so that it can fully accommodate collections and collection description.

The study group has proposed the following representational requirements:

R1. Models must treat collections as individual resources within the aggregation and allow for the representation of properties of the collection.

R2. Models must be prepared to represent collection membership as a property that stands between resources. Item-level entities must be explicitly linked to collection-level entities.

R3. It is necessary to have a set of properties designed to describe collections in ways that support users and managers. Collection-level description include:

   a. Properties that record the institutions that have participated in the stewardship of resources; including institutions collecting and/or holding physical resources, institutions that host digital versions of resources, and institutions that have created descriptions of resources.

   b. Properties that can be used to reflect the contextual information implied by collection membership, including topical or subject properties, properties related to the principles used to determine membership in the collection, and properties about the intended audience for a collection.

R4. To the extent possible, property values in metadata should be identifiers of resources that the system can make actionable.

[1] http://pro.europeana.eu/edm-documentation

But are the things represented in collection models that meet these requirements really contextual?

Lee analyzes context, and proposes nine classes of contextual entities in order to aid the representation of contextual information in digital aggregation systems [1].[2] The requirements for collection representation proposed by the study group closely match Lee's classes of contextual entities. The classes most relevant for modeling collections to support the roles identified by the EDM/DCC study group are *Object*, *Relationship*, *Purpose*, *Agent*, *Time*, and *Place*. Specifically, the *Object* and *Relationship* contextual entities are addressed by extending EDM according to R1 and R2 respectively, while *Purpose*, *Agent*, *Time*, and *Place* are met by R3.

R1 states that collections should be treated as individual resources within an aggregation, and therefore gives each collection that an item might be a member of the status of an *object*. Lee notes that taking of a contextual view of information in digital aggregations forces the identification of a target entity that we are providing context for. In our case, an item (e.g. a photograph) is the target entity, and the representation of a collection that the resource is a member of (e.g. a collection of photographs taken by a historical figure) can be used to supply context for interpreting and using the item. The study group has recommended extending the EDM class hierarchy to include the class *edm:Collection* as a subclass of *dcmitype:Collection*, with the definition "a group of objects gathered together for some intellectual, artistic, or curatorial purpose."

Lee describes relationships as associations between entities that "cannot be reduced to or adequately expressed as a property of the entities." R2 states that representing collections in digital aggregation systems requires explicitly representing the membership relationship between items and collections. As we argued in Wickett, et al. [2], recording this relationship is an essential element of using the collection object to supply contextual information for items. In order to meet this requirement, the study group has proposed adding the property *edm:isGatheredInto* to the model to reflect the specialized semantics of collection membership.

The remaining relevant contextual entities from Lee (*Purpose*, *Agent*, *Time*, and *Place*) are reflected in the recommendations for collection-level description developed by the study group. The recommended collection-level property *dcterms:accrualPolicy* is intended to capture the collection development policy, and therefore the purpose behind the collection (this information may additionally be found in *dc:description* statements at the collection level). The properties *dcterms:temporal* and *dcterms:spatial* are intended to be used at the collection level to reflect aspects of time and place, respectively.

The study group on collection modeling paid particular attention to the representation of *agents* involved in the creation and stewardship of collections, analyzing the various stewardship roles reflected in the proposed properties for collection-level description. Institutional agents responsible for the steward-

[2] Although Lee uses the term "digital collections" in his title, the discussion indicates that he is using the term to refer to digital aggregations and repositories generally.

ship of collections are represented with the *edm:dataProvider* property, while those responsible for making content available online are represented with the *edm:provider* property. The agent responsible for applying the criteria for collection membership to individual items and thereby gathering them into the collection is represented with the *dc:creator* property (as applied to the collection). Additional agents may be indicated via other collection-level properties, and following R4 (and the general principles of EDM) these agents may also be treated as objects and have further descriptive information attached to them, thereby providing additional context.

Of course, in order for any of these collection-level properties to effectively supply context for items they must in some way exposed to users and administrators. Here R1 and R2 come into play by providing the technical means to display collection-level information along with items from the collections. Since the collection is itself an object in the aggregation, properties designed to capture collection-level context can be attached to it; and since the item is linked to the collection, those properties can be displayed and used to aid in retreival and access.

Given the intersection between the requirements developed by the study group, and Lee's contextual entities, it seems safe to say that collection modeling can act as contextual modeling for items in digital aggregation systems. Questions still remain about the most effective ways to generate collection-level metadata and the best strategies for integrating this contextual information into item-level displays and searches.

References

[1] Lee, C. (2011). A framework for contextual information in digital collections. *Journal of Documentation, 67*(1). doi:10.1108/00220411111105470.

[2] Wickett, K.M., Isaac, A., Fenlon, K., Doerr, M., Meghini, C.L., Palmer, C.L, & Jett, J. (2013). Modeling cultural collections for digital aggregation and exchange environments. CIRSS Technical Report 201310-1, University of Illinois at Urbana-Champaign. Retrieved from http://hdl.handle.net/2142/45860.

Unity criteria of collection contexts: Why are items together?

Martin Doerr
Institute of Computer Science
Foundation for Research and Technology – Hellas

Recent advances in making very large amounts of digital objects available on-line increasingly reveal the need to represent metadata of wider contexts than that of the form, creation history and aboutness of each individual object. The traditional concepts of collection have been very successful for contextualizing objects, presenting their relevance and guiding users to physical content kept in memory institutions. It is more and more obvious that there is more to the traditional concept of collection than just a work-around to the fact that users could not search directly individual objects so far.

Therefore it appears to be a promising challenge to find adequate digital equivalents to traditional concepts of collection. In order to do so, the functional role of these concepts and the associated human activities should be better understood. In this position paper we present the idea that "unity criteria" of a collection, i.e. the reasons why items are together in a collection, are a key to better understand both the processes of collecting and the information value the membership in a collection adds to the item. These criteria are typically a formulation of the decision-making that has been guiding the development of a collection and captures the curator's intentions, but also of passive cultural-historical incidents leading to the its current composition.

Unity criteria for collections are often expressed in characteristic collection titles, such as "Medieval Europe", "Waddesdon Bequest", "Roman Britain", "Ancient Europe 4000-800 BC", "Sir Hans Sloane Collection". From these examples, we can already roughly distinguish criteria relating to a common context of provenance of items from relating to the context of the collector and the incidental acquisition history. Further we have to regard criteria of completeness of subject coverage, such as "all etchings by Rembrandt", "correspondence between Newton and Hooke" or "all bird species of Europe" and, most importantly, criteria of cultural-historical relevance. A detailed account of a method for assessing and describing the relevance of cultural heritage objects and collections can be found in Russell and Winkworth [1]. More formally, we can distinguish four general categories that may be used to determine whether an individual item is suitable for membership in a collection:

1. Nature of the object: The individual construction or form of an item

provides evidence or information about the context of its creation, or represents an extraordinary artistic, scientific or technological achievement.

2. Example function: An item exemplifies a particular category or type of thing.

3. Witness function: An item was present at an event or in a period of interest, carrying direct evidence from that presence or simply serving as an illustration of the relevant context.

4. Aboutness: An item refers by form, depiction or content to some person, object, place, event, concept or other phenomena of interest.

Following the CIDOC CRM, the core concepts for modelling cultural-historical contexts are that of periods and activities in which people, things and ideas meet in space-time. We maintain that all these aspects and the above mentioned reasons for being in a collection are relevant for users to find and understand items, and that it is possible to develop a complete formal model for the representation of collection unity criteria that will enable formal reasoning methods to guide and provide users with relevant integrated item and collection information in Digital Libraries.

References

[1] Russell, R. and Winkworth, K. (2009) Significance 2.0: a guide to assessing the significance of collections. Collections Council of Australia Ltd, Australia. Available at: http://arts.gov.au/sites/default/files/resources-publications/significance-2.0/pdfs/significance-2.0.pdf

On the Logical Foundations of Digital Collections

Carlo Meghini
Istituto di Scienza e Tecnologie dellInformazione,
Consiglio Nazionale delle Ricerche, Pisa

Nowadays, digital libraries are one of the most common types of information system that can be found in everyday life: they range from those serving large societies, such as the web or Europeana, to those serving single individuals, such as those managing the music or photo collections in our phones or tablets.

The term *collection* occurs very often in the digital library discourse. Indeed, the notion of collection is considered to be a fundamental one for characterizing the content of a digital library, up to the point that in same cases the terms *collection* and digital library *content* are used as synonyms.

When analyzing the conceptual foundations of digital libraries, the question then naturally arises what a collection is in terms of the notions used in the logical analysis of discourse.

It is quite uncontroversial that collections are containers of items, and that the items in a collection may be inividual objects as well as (other) collections. The "container" metaphor is in fact very suggestive, as it directly relates to the physical realization of collections found, *e.g.*, in libraries or archives. The metaphor naturally extends to the digital world, which offers the concepts of folder or directory as candidates to adequately render all desired features of collections. From these metaphors, we are then brought to conclude that a collection is an individual having a composite nature, the parts of a collection being the inividual objects and the collections that it "contains".

However, defining collections solely in terms of their content, or *extension,* leads to the conclusion that two collections with the same content are indeed one and the same collection. While it would be rather uncontroversial that two works having exactly the same textual content are indeed the same work, the sameness conclusion is very counterintuitive if applied to collections. For example, the collection of best wines in Italy and the content of my cellar may consist, at some point, of the same bottles; but it would be very arguable that they are, *by definition*, one and the same collection.

In order to account for the difference between purely extensional objects, such as containers, and collections, we have argued [1] that collections, in addition to contain a set of objects, also have an *intension.* And therefore, the condition of identity for collections is more complex than sameness of extension in a specific situation: it amounts to sameness of content in *every possible situation.* This feature makes collections akin to predicates in logic, and as such

endowed with an extension and an intension, the latter accounting for the purpose that brings the collection into existence. Under this view, collections are modelled as predicate symbols, and

- membership of inividual objects in a collection is captured by predication; for example, membership of *tom* in the collection of *domestic cats* is expressed in natural language as the sentece "*tom is a domestic cat*" and represented in logic by the atomic sentence $\mathsf{DomesticCat(tom)}$;
- membership of a collection in another collection is captured by an implication statement having the member collection in the antecedent and the receiving collection in the consequent; for example, membership of the collection of *my cats* in the collection of *domestic cats* is expressed in natural language as the sentence "*my cats are domestic cats*" and represented in logic by the sentence $(\forall x)\mathsf{MyCat}(x) \rightarrow \mathsf{DomesticCat}(x)$;
- the purpose of the collection, or collection intension, is expressed as a predicate; for example, the intension of the collection of *my cats* is expressed in natural language as the sentence "*cats that live in my garden*" and represented in logic by the predicate $\mathsf{Live}(x, myGarden)$;
- the relation between a collection and its intension is expressed by an equivalence statement between the collection and the predicate representing the intension; for example, $(\forall x)\mathsf{MyCat}(x) \equiv \mathsf{Live}(x, myGarden)$. The statement may be weakened to be an only necessary or only a sufficient condition, but this does not affect the intensional nature of collections, but rather reflects our degree of knowledge of the collection intension.

On a practical level, the choice of an appropriate ontology is a necessary condition for tackling interoperability issues. And making collection interoperable is today a primary concern of many istitutions.

References

[1] Carlo Meghini and Nicolas Spyratos. Unifying the concept of collection in digital libraries. In Zbigniew W. Ras and Li-Shiang Tsay, editors, *Advances in Intelligent Information Systems*, volume 265 of *Springer Studies in Computational Intelligence*, pages 197–224. Springer Verlag, 2010. ISBN: 978-3-642-05182-1.

4.2. Breakout Sessions: Overview and Outcomes

The morning breakout sessions allowed participants to hold informal, small-group discussions on four topics raised in the Conceptual Foundations position papers: scholarly use of collections; non-scholarly use of collections; formalizing collection structures; and unity criteria. The following main points or themes emerged from the morning breakout sessions:

- Collections have a number of uses in digital aggregations. As digitization continues and aggregations grow, collections emerge as a viable and multifaceted tool for helping aggregators and users deal with the burgeoning scale of systems.
- Collections created by libraries, museums, and archives have value for non-scholarly users, but in what contexts or under what conditions will this value exceed the cost of describing and representing collections in large-scale aggregations? Participants suggested that user-generated collections may be a more valuable alternative for digital libraries oriented toward use by the general public.
- Users want to be able to create and keep their own collections, and to attach metadata to those collections. Most digital libraries and aggregations do not adequately afford this functionality, though it holds potential to add value to systems. For example, user-generated collections can contribute to improving metadata about items; crowdsourcing the large-scale curation of related sets of items; and creating shareable, reusable learning and teaching resources. The trend toward user-generated collections is provoking reconsideration of the metaphors we use to understand collections, such as: products, processes, streams, and narratives.
- As a type of user-generated collection, scholarly research collections are particularly interesting from the perspective of potential reuse. How do we facilitate reuse, ensuring the authenticity, trustworthiness, and shareability of user-generated collections?
- Participants disagreed about how to formally (mathematically) model collections underlying digital systems. How do our basic modeling strategies strike a compromise between our intuitions about collections, and having practical benefits for use in technical systems? Participants disagreed, in particular, about how to account for a collection-creator's intentions in a model. Competing alternatives discussed in these breakout sessions included modeling collections as series of events or as logical predicates.

The remainder of this section gives limited synopses of discussions in each of the four breakout sessions, based on detailed notes taken in each session.

4.2.1. Scholarly use of collections

In this breakout session led by moderator Carole Palmer and panelist Hur-Li Lee, participants discussed their experiences with, understanding of, and visions for scholarly uses of collections.

Participants discussed how the increasing scale of digital collections offers both challenges and opportunities for scholarly research. While large-scale collections might reveal patterns that small-scale collections do not, information retrieval and comprehending massive collections (e.g., through visualization) are significant problems. Palmer suggested the value of collection representation for

organizing large systems of information. Citing the example of a debate over whether to distinguish a women's studies section in a large academic library, Lee noted that by prioritizing, separating, or making certain things visible within large information systems, collection-building can be a political act. Another participant noted that increasing scale places new demands on the skill sets of humanities scholars, a challenge that must be met by new opportunities for integrating humanistic hermeneutic approaches to analysis with retrieval and reading on a large scale.

Participants considered the key scholarly functions of collections, or what scholars actually want to do with collections. They considered how to contextualize collections for reuse in new disciplines, and how to make user-generated metadata and annotations interoperable between collections. Participants discussed how to productively shift our perspectives on collections. One participant noted that his work aimed to construe collections not as products or processes, but as streams, arguing that the metaphor allows for greater personalization of collection building. He suggested that finding ways of enabling scholars to create dynamic collections should be a priority. Another participant confirmed that scholars increasingly hoped to create personal, digital collections within the library, and to be able to add metadata to those collections. This trend had several implications: participants suggested that the concept of collection was not yet conceived with sufficient breadth to encompass all scholarly contexts. Another suggested that scholars use different resources for different reasons. One participant suggested considering not just transient (or dynamic) collections but transient systems, or rather adaptive and dynamic systems for collection making.

Participants discussed the opportunities and challenges for reuse of collections. Acknowledging that user-generated research collections change over time, one participant cast reusable collections as "iterative collections", and wondered how materials could be iteratively or gradually added, while maintaining the coherence and tracking the provenance of the collection. Another participant noted that gradual or iterative growth requires dedicated maintenance. Participants confirmed that in this light, authority, authenticity, and provenance become key issues for collection-development and maintenance, and these entail further questions:

- Authenticity is not an original condition but a process, which changes over time. What is the acceptable level of variation in a collection?
- How do you determine whether a collection is authentic and trustworthy rather than a variant? It will depend on institutional prerogatives, but is there a basic set of criteria?
- Collections often serve the identity and recognizability of institutions: how do we maintain that, while affording users the ability to create, manipulate, or change collections?
- How do we structure and represent research collections so that they are usefully shareable? For example, one participant noted that humanist scholars using the HathiTrust wish to conduct research at various levels of granularity, and that may require altering extant metadata. As another example, Palmer noted that scientists' data sets can be useless or useful to others depending on the granularity at which they are shared, how they contextualized, and what subsets are shared.

4.2.2. Non-scholarly use of collections

Antoine Isaac led a session that focused on participants' observations of, experiences with, and ideas for non-scholarly uses of collections, including uses in public libraries, in cultural heritage, aggregations, and generally. Participants came to the session with varying impressions of the usefulness of collections for

non-scholarly users of digital libraries. By the end of the discussion, participants seemed to acknowledge the utility of collections for non-scholarly users, but continued to disagree about the benefits, relative to costs, of representing collections on a large scale in an aggregation.

One participant discussed a survey of librarians and searchers (both scholarly and non-), which revealed competing conceptions of collections prominent among users: collections as entities versus collections as processes, and encouraged a user-centric view of collections. Participants noted some doubt that institutionally generated collections (such as special collections in libraries) will be adequate to serve users in massive, online aggregation environments. One participant suggested that prioritizing rich descriptions of items over collections would be more effective for helping users navigate massive information spaces. Another disagreed, noting that critical information about items often lies in descriptions of collections they belong to, especially information that contextualizes items and contributes to their meaning.

The discussion focused on the practicality of representing collections in large aggregations, such as DPLA and Europeana: where collection descriptions exist, are they readily integrated with an item-level library? And where they do not exist, how feasible is it to create them? Isaac, a representative of Europeana, discussed ongoing work to survey providers for collection descriptions in order to augment item information in the aggregation, focusing on subject, temporal, and geographic coverage information. Participants recognized that creating or recovering collection descriptions could be difficult, but noted many possible advantages to the representation of collections in aggregations, including improved browsing and navigation, especially of topical information.

Participants were particularly enthusiastic about the opportunities for user-generated collections, which are in demand among both scholarly and non-scholarly users. Users want to be able to create collections within information systems, and have those collections function in different ways. Participants noted that by allowing users to create and share collections, the system can leverage the pooled, curatorial work the users contribute to enrich metadata and access mechanisms. Participants agreed that it would be critical to create systems and platforms to facilitate user-generated collections, and to help people share not only their collections but their motivations for creating them. Participants also noted that collection visualization is a topic ripe for further research, and that helping users comprehend information spaces and collections in them will be important to making them useful.

4.2.3. Formalizing collection structures

In this session, led by panelists Carlo Meghini and Karen Wickett, workshop participants discussed open questions about how to formally model collections in such a way that those models can underpin accurate ontologies, effective description standards, and interoperable representations in digital systems. The discussion focused primarily on Meghini's suggestion for modeling collections in mathematical logic as predicate symbols, and the argument that since collections possess both intensions and extensions using predicates uniquely allows representation of both aspects of collections.

Participants explored the implications of modeling collections as predicates, from a logical perspective. Traditionally, a predicate symbol is associated with an intension that gives the meaning of the predicate and an extension that gives the individuals that satisfy the predicate. When this structure is applied to

collections, the purpose and unique character of the collection is represented by the collection intension, and the collection membership at any given time is represented by the collection extension. This approach means that an individual collection functions as a class, particularly in any digital library model that treats digital objects as logical individuals. Then the class that contains all collections is a meta-class, since its members are classes. Participants suggested that this is not an intuitive view of collections. While this view of collections is revisionary, Meghini argued, it bestows greater clarity on the concept of collection.

A primary objection to this approach also comes from the logical perspective on digital libraries and the fact that typically in a strictly first-order logical system it is not possible to assign properties or relationships to predicates. This means that it is not possible with this approach to record collection-level information in the same way that item-level information is recorded in a digital library. The collection intension information can be expressed with natural language documentation, but it will not be treated in the same way as item-level facts. According to Wickett, this contradicts the strategy to treat collections as individual objects in digital libraries, which was recommended in the whitepaper mentioned above.

Participants also expressed doubts about whether intension can be modeled; after all, the purposes for building a collection for integrating particular items may be ineffable, inconsistent, and otherwise out of keeping with the concept of a mathematical intension. Meghini argued that this perspective admits leeway for the imperfect, often intuitive, and interpretive ways in which people build collections. According to Meghini, a collection's intension does not, in practice, readily determine what items are gathered into a collection, because manifesting the intension is an interpretive process. Therefore, by this argument, someone building a collection may believe a given item fits the intension, though it does not. In Meghini's view, this does not compromise the integrity of the model.

4.2.4. Unity criteria

In this breakout session, led by panelist Martin Doerr, participants explored a concept at the heart of understanding collections and formalizing them: that of unity criteria, or the reasons (both intentional and historical) that items are brought together into collections. Doerr encouraged participants to view collection development as a process -- not a direct manifestation of a curator's intentions, but as determined also by social and historical circumstances and events.

In this view, Doerr clarified, he opposes Meghini's model of collections as predicates, because intension -- related as it is to a curator's intentions -- cannot be modeled -- only expressed as historical fact, alongside other historical facts that contribute to shape collections.

Participants discussed a range of factors that affect the concept of unity criteria:

- Collections can be created to help a researcher assess the viability of a new hypothesis.
- Collections are sometimes built to push political agendas.
- Collections can be built to assert causal relationships or influence between items.
- The arrangement of items in a collection (whether chronological, driven by aesthetic concerns, etc.) may be related to the unity criteria; this is evident in the differences between archival order (which is original order), the organizational schemes of libraries (which rely on the physical

features of documents, their subjects, etc.), and museum arrangements (in both exhibition spaces and behind the scenes).

- There is a distinction between scholarly collections (built to serve as evidence for research), institutional collections (from libraries, archives, and museums), and collections created for personal use, which are largely idiosyncratic: how do these kinds of distinctions affect the notion of unity criteria?

Doerr suggested that if we can arrive at a strong definition of unity criteria, we can employ the "gathered-togetherness" of different items, or the very fact that a group of items has been gathered together, to more accurately define and more richly describe collections.

5. Practical Implications

5.1. Position Papers

Implementing collection contexts, and metadata issues related to normalization and shareability

Amy Rudersdorf
Digital Public Library of America

The Digital Public Library of America (DPLA) brings together the riches of America's libraries, archives, and museums, and makes them freely available to the world. The DPLA aims to expand this crucial realm of openly available materials, and make those resources more easily discovered and more widely usable and used, through its portal (http://dp.la), platform (http://api.dp.la), and advocacy for open access to metadata and the resources it describes.

As of February 2014, the DPLA has 20 active "Hubs" and three others that will begin providing access to their data soon. Currently, Hubs are either identified as "Content Hubs" (large contributors of their own content, such as Smithsonian Institution and the National Archives and Records Administration), or "Service Hubs" (collaborations of many partners with a single aggregation point managed by one or more institutions, such as South Carolina Digital Library and the Mountain West Digital Library). The DPLA contains over 5.6 million records representing online resources from 1,100 libraries, archives, historical societies, and museums of varying sizes and data-creation mastery from across the United States.

With aggregations at scale come challenges of variations in metadata formats, standards implementations, and data quality [5, 2]. The DPLA's Hubs model[1] creates a more sustainable partner model; the work of aggregation, metadata remediation, and professional development is shared by the Hubs, their partners, and the DPLA. And while this model helps create more predictable metadata, it does not completely alleviate the issues that are inevitable in aggregations of heterogeneous institutions of this magnitude. This is where the work of metadata remediation comes into play. The DPLA's work in this area has focused to date on education about linked data and the use of URIs to identify entities and the source of text values, etc., machine identification of geographic names and application of coordinate values, normalization of date values, and beginning the work to develop better methods for the implementation of standardized rights statements.

A further challenge to aggregating at scale is maintaining the persistence of an identifier or other information about a collection, which can provide valuable contextual information about each digital resource [1]. The DPLA has taken

[1] http://dp.la/info/get-involved/partnerships/

first steps to capture this collection-level information in the *dpla:Collection* class in the metadata application profile (DPLA MAP).[2] As with item-level description, the quality and correctness of collection metadata varies. In some cases, collection description is expressed within the metadata record (i.e., a Dublin Core "source" or "relation" value), while in others the description is provided at the OAI-PMH set level. In a few cases, no collection is defined.

Institutions define the concept of "collection" differently, as well [3, 4]. For example, a collection may simply contain all of the records from a single contributor ("Beltrami County Historical Society"), all items from a contributor in a particular format ("Beulah Glover Photograph Collection"), a single item ("Brief History of Moscovia by John Milton"), or a designation that simply identifies the data as a DPLA-only set ("SSDPLAWashington"). Identifying what comprises each of these collections will be a difficult undertaking to automate. While collection data is not currently presented through the portal, the DPLA intends to present this information when some of these variations are addressed.

As stated, the challenges outlined here are no different than those faced by other large-scale aggregations. The DPLA is an advocate for empowering Hubs to perform their own data remediation so that their improvements can be implemented locally, before their data ever reaches the DPLA servers. Services such as the "DPLA Regional Hub Extraction and Transformation Services,"[3] in development at the University of Minnesota, are examples of the work being done to create metadata remediation modules and to take local control over metadata transformations to the DPLA data structure. Minnesota's open-source tool (and others like it) will allow users to "turn on" only the remediation scripts required by a data set. (For example, one script may strip periods from the end of subject terms extracted from MARC. If this is not an issue in a data set, however, the script could be ignored.)

Wide adopted by other institutions, or even hosting of tools like these by the DPLA for institutions without the infrastructure to implement it themselves, would ensure that data cleanup, errors in implementation standards, or normalization and enhancements could be applied prior to harvesting by DPLA. Sustainability, improved data at the source, and speed of harvest are just a few of the benefits to this model.

References

[1] Curral, J., Moss, M., & Stuart, S. (2000). What is a collection? *Archivaria 58*. Accessed February 2, 2014. http://journals.sfu.ca/archivar/index.php/archivaria/issue/view/417/showToc.

[2] Dushay, N. & Hillmann, D.I. (2003). Analyzing metadata for effective use and re-use. In *Proceedings of the 2003 international conference on*

[2] http://dp.la/info/developers/map/

[3] https://github.com/chadfennell/dpla.services

Dublin Core and metadata applications: supporting communities of discourse and practice—metadata research & applications (DCMI '03). Dublin Core Metadata Initiative, Article 17, 10 pages. Accessed February 4, 2014. http://hdl.handle.net/1813/7896.

[3] Isaac, A., Schlobach, S., Matthezing, H., & Zinn, C. (2008). Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies, *Library Review, 57*(3), pp.187-199.

[4] Renear, A.H., Wickett, K.M., Urban, R.J., Dubin, D., & Shreeves, S.L. (2008). Collection/item metadata relationships. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, Berlin, Germany, September 22-26, 2008. Accessed February 2, 2014. http://hdl.handle.net/2142/9144.

[5] Shreeves, S.L., Riley, J., & Milewicz, L. (2006). Moving towards shareable metadata. *First Monday, 11*(8). Accessed February 12, 2014. http://firstmonday.org/issues/issue11_8/shreeves/index.html.

The CENDARI Knowledge Framework: a dynamic research environment and a grown knowledge framework

Sheila Anderson
Centre for e-Research
King's College London

1 Introduction

Over the last decade the European Commission FP7 programme has funded the development and building of Research Infrastructures designed to meet the needs of humanities researchers. This includes the Collaborative Digital Archive Research Infrastructure (CENDARI). CENDARI aims to provide integrated access to the archives of two domains: First World War studies and medieval history. The project is highly collaborative spanning multiple areas of expertise and knowledge including computing scientists, information scientists, historians, archivists, and librarians working within a programme of technical research informed by cutting edge reflection on the impact of the digital age on scholarly practice. The primary goal is to facilitate and enhance research by increasing access to and use of records of historic importance across Europe, creating a powerful new platform for accessing and investigating historical sources in a transnational and comparative fashion and overcoming the national and institutional information silos that now exist.

The starting point for CENDARI was to understand (using surveys and interviews) the constraints scholars face when working in Archives and their archival research practices, and to envisage, through a series of participatory design workshops, what tools and services might be required of a digital research infrastructure. A number of key challenges emerged including:

- Fragmented and dispersed archives and collections, for example, the archives of Hans-Gunther Adler who documented life in the Terezin Ghetto are dispersed across four institutions in four countries.
- The differences between the epistemologies of historians and archivists / libraries and thus between the way in which each will extract, describe, and create knowledge from archival records. Archives and libraries undertake knowledge making in the form of structuring, arranging, describing and classifying the archives and collections in their care, primarily for

the purpose of information and document discovery and retrieval and the long term care of the materials. Researchers in the humanities undertake knowledge making as a hermeneutic process, as part of a practice of argumentation, and its exploitation and further development takes place through engagement with previous knowledge within the discipline.

- The way in which archival finding aids are structured as text streams rather than structured data. For example, the Encoded Archival Description (EAD) is modelled closely on the traditional archival finding aid and so tends more towards the prose description as its primary content rather than fine-detailed interoperable metadata. It therefore produces a record of a collection which is more descriptive than potentially analytical. It is also primitive in terms of data interoperability by the standards of today's Semantic Web, providing limited facilities for fine-grained semantic linking of the type facilitated by RDF triples and their use of URIs (Universal Resource Identifiers).

- The requirement for universal descriptions to meet the needs of multiple research topics. For example, in the writing of the histories of the First World War literary studies investigating changes and differences in the portrayal of disability in fiction and in official records sit next to historical studies investigating the 'green cadres', groups of armed deserters, who hid themselves in forested areas, staging raids on livestock and crops, and attacking the local gendarmerie and military. Universal descriptions cannot hope to capture the depth of information required to fulfill all these niche areas.

2 The CENDARI Knowledge Framework

To address these challenges CENDARI is seeking to connect dispersed archives, manuscripts, and manuscript fragments; connect and reflect the knowledge of both archives and scholars; extract knowledge and meaning to aid information retrieval and to answer research questions; and to make sure this work is rooted in the hermeneutic practices of historians. The knowledge framework for CENDARI combines the development of an integrated metadata strategy with the development of dynamic domain ontologies. Because CENDARI is conceived as a dynamic eco-system rather than as a portal to resources, its metadata needs to avoid the common paradigm of a centrally-provided collection of information produced by expert practitioners, and instead function as a kernel on which newly-created, constantly evolving, layers of researcher-generated content can accrete in a logical way. This requirement for extensibility makes XML (eXtensible Markup Language) the most logical choice for CENDARI's metadata syntax. However, despite the substantial advantages of using XML, it is also important for the CENDARI metadata to integrate with the Semantic Web. To allow this, the project is also making its metadata available as RDF triples.

At an early stage of metadata design for CENDARI, it became clear that the two core communities would have different emphases for their respective metadata requirements. For the medievalists the core requirement enunciated was for a rich and detailed description of complex digital objects. The WW1 historians foreground the contextual, collection-related, environment within which objects (often much simpler than their medieval counterparts) are found. Thus three distinct levels of granularity emerged: at the top is a description of the holding institution. Below this is the collection itself, which requires descriptions of a richness and flexibility to meet the demands of the twentieth-century historian in particular. At the base of this conceptual hierarchy is the item description, which in itself may be of an equal complexity to any higher-level object. The metadata strategy makes use of both new and existing metadata schemas at these three levels: an amended version of EAG is used at the institutional level; MODS with TEI extensions is used at the item level; due to the semantic limitations of EAD a new CENDARI specific schema was created at the collection level, but that can generate EAD using an XSLT transformation.

The domain ontologies within CENDARI will not form a static database of knowledge separate from the rest of CENDARI, but will form a dynamic knowledgebase integrated into a wide range of other CENDARI services. The CENDARI domain ontologies will be:

- Integrated with institutional, collection, and item-level metadata records within the CENDARI environment, enabling researchers to navigate between related records.
- Integrated with researchers' notes in the virtual research environment and any associated annotation tools (e.g., Pundit), facilitating the organization of personal research notes as well as the discovery of other researcher's work (where permission is given).
- The foundation for additional CENDARI services, such as the development of Pineapple, a CENDARI service that is designed to answer questions rather than queries through the use of semantics.
- Enhanced as a Named Entity Recognition service identifies entities and relationships from additional metadata records.
- Enhanced as researchers using the CENDARI virtual research environment build new entities and relationships into the domain ontologies.
- A knowledgebase to be queried in its own right.

The Europeana Data Model will be extended for the CENDARI ontologies, with separate extensions for each of the two research communities:

Within CENDARI the focus is primarily on the transformation and curation of ontologies that already exist, although this is supplemented by instances created through named entity recognition (NER) and user contributions captured through the CENDARI virtual research environment note-taking tool.

CENDARI contextual classes	EDM contextual classes
Places/spaces	Places
Persons/role	Agents
Institutions	Agents
Dates	Timespans
Events	Events
Topics	Concepts

Table 1: Suitability of EDM to meet the needs of CENDARI researchers

3 Conclusion

The knowledge framework research described above provides the foundation steps towards the establishment of a viable research infrastructure eco-system. The metadata strategy and its complementary ontologies allow much more interoperable access to historic records transnationally than has previously been achievable and make possible the creation of a dynamic enquiry environment underpinned by a shared knowledge framework.

Hunting for semantic clusters in aggregations

Shenghui Wang
OCLC Research, Leiden, Netherlands

1 What is the problem?

More and more large-scale digital libraries and aggregators have become available, such as Euopeana,[1] the Digital Public Library of America[2] and, also, Worldcat.org.[3] These aggregators provide access to large numbers of heterogeneous digital objects, however, also bring challenges to users to explore such large aggregations.

Aggregating metadata from heterogeneous collections raises quality issues such as uneven granularity of the descriptions, ambiguity between original and derivative versions of the same object, even duplication if different providers give access to a same object. Also, simple, common-denominator vocabularies such as Europeana Semantic Elements (ESE) are inappropriate for capturing internal semantic links between objects (e.g., parts of an object, adaptations of a work, objects representing others) or external links to contextual entities (*e.g.*, places or persons related to an object). Many data providers do not have resources to provide richer and interoperable metadata as instructed in the CIDOC-CRM[4] and the new Europeana Data Model.[5]

Traditional usage of a search box and a query-response mode of interaction are no more sufficient when users do not have clearly defined information needs, or when they want to gain an overview over collections. More and more browsing and exploration functionalities based on thesaurus, facets, or clustering are being proposed to improve user search experiences. However, overviewing and exploratory browsing have still not been investigated much [1].

2 Detecting semantic clusters of content in aggregations

We propose a bottom-up approach: finding related digital objects at different levels of similarity, which potentially reflect different semantic relations between them. As shown in Fig. 1, after calculating the clusters at level 80 (with the

[1] `http://europeana.eu/`
[2] `http://dp.la/`
[3] `http://digital.experimental.worldcat.org/`
[4] `http://www.cidoc-crm.org`
[5] `http://pro.europeana.eu/edm-documentation`

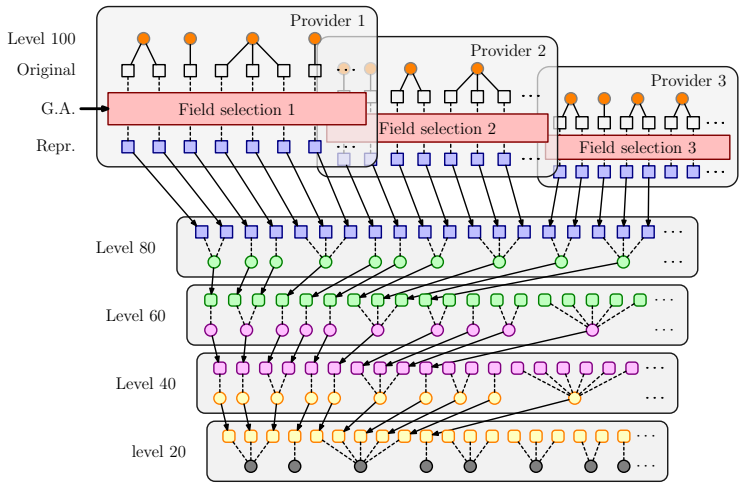

Figure 1: Hierarchical structuring of CHOs at different similarity levels.

similarity of 80%), we generate an artificial record from each cluster, gathering in each metadata field its values for all clustered records. These artificial records, together with all the records which could not be clustered at level 80, will join the clustering process at level 60. We again cluster at level 40 and 20 in the same way.[6] In the end, hierarchies of records are generated, so that one can have some structural information about these records, instead of quickly getting drowned in the sheer amount of data.

3 Open issues

3.1 Clustering objects around user-defined topics

Manual inspection of the results of this naive unsupervised clustering algorithm reveals that the digital objects in clusters are similar in various dimensions, e.g., pages of the same book, photos of the same building, work by the same author, objects about the same theme, etc. However these dimensions sometimes do not align with the expectations and needs of users. For example, an archive user is looking for archives which are about "the women's movement" or "Albert Einstein and his religious views." These topics are often not explicitly assigned to existing archival data, while users have better ideas what kind of archival materials they are looking for. We are currently developing a new algorithm by combining current clustering algorithm with user-defined constraints in order to generate topical clusters that are more meaningful for users. This would potentially bring more flexibility in providing topic-based exploration of digital objects.

[6] See more technical details in [2]

Figure 2: Multidimentional similarities between digital objects

3.2 Multidimensional clustering

Digital objects can be linked along many different dimensions. As illustrated in Figure 2, one picture of Sir James Eyre could be clustered with another picture of his, while for some users, it might be interesting to explore a cluster that contains pictures of the people who shared the same office. Current clustering algorithm based on lexical similarity would not be able to distinguish such different dimensions. More semantic enrichment which identifies semantic elements (e.g., person, organisations, locations, etc.) within metadata would allow us to experiment multi-dimensional clustering.

References

[1] M. Hall and P. Clough. Exploring large digital library collections using a map-based visualisation. In *Proceedings of 17th International Conference on Theory and Practice of Digital Libraries*, pages 216–227, Valletta, 2013.

[2] S. Wang, A. Isaac, V. Charles, R. Koopman, A. Agoropoulou, and T. van der Werf. Hierarchical structuring of cultural heritage objects within large aggregations. In *Proceedings of 17th International Conference on Theory and Practice of Digital Libraries*, pages 247–259, Valletta, 2013.

Supporting users' navigation and exploration of large digital collections: Experiences from the PATH project

Paul Clough* & Mark Stevenson**
*Information School
**Department of Computer Science
University of Sheffield (UK)

Abstract

In this short paper we summarise our experiences from the EU-funded PATHS project. The aim of this project was to support various types of users with their navigation and exploration of large digital cultural heritage collections. A dataset derived from Europeana was gathered and enriched using techniques from natural language processing and information retrieval. This enabled the design of a system incorporating various navigational and exploratory search aids, such as subject hierarchies, recommendations, links to related Wikipedia articles, a workspace and map-based visualisations. The system also allowed users to create narrative-like structures through the collection through trails/paths that can be used as collection guides or used for educational purposes.

1 Introduction

> *Cultural heritage involves rich and highly heterogeneous collections that are challenging to archive and convey to the general public* [5].

The PATHS (Personalised Access To cultural Heritage Spaces) project[1] was funded under the European Commission's FP7 programme and consisted of partners from multiple (cultural heritage, library and information science, and computer science) and from academic and non-academic institutions. The project aimed to support expert (e.g., scholars, curators) and non-expert users (e.g., students, the general public) with navigating and using materials from large and heterogeneous cultural heritage collections [2, 4]. A selection of 1,701,672 artefacts (i.e., metadata records) from Europeana, the European aggregator for museums, archives, libraries, and galleries, was used as the dataset,

[1]PATHS project website: http://www.paths-project.eu/

but additional semantic enrichment was carried out to support the provision of enhanced search and browse functionality.

During semantic enrichment, two issues identified with Europeana metadata were: (i) limited information associated with many items and (ii) the lack of a unified indexing scheme across aggregate collections. These issues were addressed through semantic enrichment[2] that included: (i) identifying key entities, such as people, locations and dates; (ii) identifying the similarity between pairs of artefacts, including categorising how items were similar (e.g., similar description, similar location or similar event); (iii) identifying'background links' to relevant Wikipedia articles; and (iv) the automatic creation of data-driven subject hierarchies to organise items [1]. The enrichments were used in various ways. For example, to provide links to similar items and related background information from Wikipedia when viewing an item; subject hierarchies were used to provide navigational support, implemented in various ways — a thesaurus, term cloud view and a map-based visualisation. The enriched data was encoded using a custom format (ESEPaths) derived from Europeana Semantic Elements (ESE).

In developing the PATHS system we adopted a user-centred approach — identify requirements, build prototype and evaluate. This involved a range of expert (and non-expert) users of cultural heritage in establishing a functional specification for the system [3]. A prototype system was developed that includes novel functionality for exploring the collection based upon the data-driven subject hierarchies, map-based visualisations of the semantic space, supporting the manual creation of guided tours or paths and the use of personalized (and non-personalized) recommendations to promote information discovery and help convey the rich content of Europena to various types of user.

One of the key features of the project has been investigating the design of functionality to support the manual creation of paths or trails through the collection. This has included a workspace feature to store items during exploration of the collection, a path editing feature for arranging the gathered items and forming narrative-like structures, and functions for sharing the paths. The resulting paths can be used as a means of navigating items in the collection based on a theme or topic, along with forming tangible learning objects for education purposes. The PATHS system has been extensively tested using various forms of evaluation, including task-based lab evaluations and field trials.

2 Concluding remarks

There is a need to develop information access systems that allow different types of users to unlock the potential of rich content available in large digital collections of cultural heritage materials. Providing users with the necessary tools and functionalities to help them to navigate and make sense of digital material requires understanding and recognising the needs of end users. The design

[2] A freely available web service can be accessed at: http://ixa2.si.ehu.es/paths_wp2/paths_wp2.pl

and development of information access tools must involve experts from multiple academic disciplines, including the arts and humanities, information science and computer science, as well as practioners and the providers of digital collections. Aggregated collections of cultural heritage can aid discoverability, but the resulting mixture of metadata standards and vocabularies used and heterogeneous information available makes providing consistent and usable information access features a challenge. As Hardman et al. state, it is the rich and heterogeneous collections that are challenging to archive and convey, especially to non-expert users including the general public [5]. The use of natural language processing and text mining can enable the semantic enrichment of digital collections, creating exciting opportunities for the creation of novel features to better support the identification, interpretation and use of relevant artefacts in large-scale digital collections.

References

[1] Agirre, E., Aletras, N., Clough, P., Fernando, S., Goodale, P., Hall, M., Soroa, A., & Stevenson, M. (2013). PATHS: A system for accessing cultural heritage collections, In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, Sofia, Bulgaria, August 4-9 2013, pp. 151-156. .

[2] Fernie, K., Griffiths, J., Stevenson, M., Clough, P., Goodale, P., Hall, M., Archer, P., Chandrinos, K., Agirre, E., de Lacalle, O., de Polo, A., & Bergheim, R. (2012). PATHS: Personalising access to cultural heritage spaces, In *Proceedings of 18th International Conference on Virtual Systems and Multimedia (VSMM 2012)*, pp.469-474.

[3] Goodale, P. et al. (2012) User-centred design to support exploration and path creation in cultural heritage collections, In *Proceedings of the 2nd European Workshop on Human Computer Interaction and Information Retrieval (EuroHCIR 2012)*, pp. 75-78.

[4] Goodale, P., Clough, P., Hall, M., Stevenson, M, Fernie, K., Griffiths, J., and Agirre, E. (2013). Pathways to Discovery: Supporting Exploration and Information Use in Cultural Heritage Collections, In *Proceedings of Museums and the Web Asia 2013*, Hong Kong, 9-12 December 2013. Available online: http://mwa2013.museumsandtheweb.com/paper/pathways-to-discovery-supporting-exploration-and- information-use-in-cultural-heritage-collections/

[5] Hardman, L., Aroyo, L., van Ossenbruggen, J. and Hyvönen, E. (2009). Using AI to access and experience cultural heritage, *IEEE Intelligent Systems, 24*(2), pp. 23-25.

5.2. Breakout Sessions: Overview and Outcomes

The afternoon breakout sessions allowed participants to hold informal, small-group discussions on four topics raised in the Practical Implications position papers: use scenarios for collections (which expanded on the morning discussions of scholarly and non-scholarly uses of collections); digital library aggregation and interoperability; data enrichment for collections; and visualization of collections. The following main points or themes emerged from the afternoon breakout sessions:

- The ways people use (or want to use) collections generates new value for collections themselves, for the aggregations they are part of, and for the items they gather together. Digital libraries and aggregations must find ways to take advantage of that added value, for example by offering functions for sharing user-generated collections and for contributing new metadata or feedback on existing metadata about collections and items.
- Technical challenges confronting the effort to make collections from all over the world interoperable are numerous. However, the solutions may be more social than technical. Participants discussed how systems can facilitate improved communication between data providers and aggregators, and between users and aggregators, in order to solve problems inherent in trying to reconcile resources from different kinds of cultural institutions.
- One of the principal problems impeding the usefulness of aggregations and collections is data quality. How do we improve and enrich the information available in collections and aggregations, in such a way that aligns with what users variously need? The participants had more questions than answers, suggesting the importance of, and interest in, this area of inquiry.
- Another area of interest and ongoing investigation revolves around the visualization of collections: how can we allow users to make "intelligent dives" into huge masses of content, to help them comprehend the whole visually, and to understand the layers of context on top of items and collections?

The remainder of this section gives limited synopses of discussions in each of the four breakout sessions, based on detailed notes taken in each session.

5.2.1. Use scenarios for collections

This breakout session, led by Carole Palmer and panelist Sheila Anderson, covered three main themes. (1) Participants revisited the concept of journeys or paths through collections, from Paul Clough's panel presentation on the PATHS project. (2) Participants discussed challenges related to assessing and developing for a wide variety of potential user needs. (3) Finally, participants discussed options for helping users comprehend huge aggregations of collections and make intelligent dives into the content.

(1) Participants were excited by the concept of user-created paths or journeys through collections and aggregations, especially as that idea opens up new possibilities for imagining and structuring collections. Different ways of working through content can be understood to generate new templates for different kinds of collections, from linear, narrative structures to clusters. In particular, participants were hopeful that opening up new kinds of collection templates could yield observations on how people make collections for themselves, which could in turn inform institutional collections.

(2) Participants considered the wide variation in user needs, even among users of the same cultural heritage aggregation. How should collection-developers or aggregators set about trying to understand and accommodate that variety? Participants were particularly eager to resolve the conundrum that without a solid understanding of specific user needs, collection-developers or cultural institutions tend to build systems that cater to general needs; but general systems, in turn, run the risk of meeting no one's specific needs. Potential solutions to this conundrum have been considered, such as treating aggregations as platforms for third-party development, which can in turn cater to niche communities of interest. But that solution must be qualified: participants noted that successful "apps" tend to dominate or exercise undue influence over further developments of a system. Participants agreed that the key will be to retain modularity and openness for development. Participants also considered whether advances in standardization and control (to increase the interoperability of collections) may inhibit development toward systems that cater to diverse needs.

(3) Helping users to see and comprehend the whole of a large collection or aggregation, and facilitating intelligent, purposeful "dives" into the content, is a major, acknowledged challenge confronting large collections and aggregations. Participants expressed optimism that research on clustering and visualization have potential to help appease this problem. But the technical solution of clustering raises new complications.

Facilitator Carole Palmer noted that the research collections that humanities scholars create for themselves exhibit features that complicate how we understand topical clustering and relevance. Humanities scholars often add to their collections based on intuition about what may be important in the future; they capture things for assessment later, according to an intuition (rather than any obvious evidence) that something may prove useful. This kind of assessment of the relevance of an item to a collection can be seen to operate on a different level from the algorithms or conscious selection policies that drive topical clustering. In light of the subtlety of this kind of use, several further questions were raised:

- How do we generate clusters that may be of interest?
- How do we anticipate at what semantic level we should be representing collections, topics, or clusters of things to users?
- How should the clusters function?
- How do we capture and keep them?

The breakout session ended with an inconclusive discussion of a theme prevalent in several of the breakout sessions: how to leverage what people are actually doing with content from our collections, and how to feed that back into our systems to improve them. The particular case discussed in this session was that of education. It is known that teachers use cultural heritage collections in the classroom. Incorporating cultural collections into lesson plans or research assignments generates new knowledge about those collections or their contexts. But that knowledge is never fed back into the collection or aggregation, to enrich the metadata or forge new links between things. Participants wondered how to take advantage of these and other value-generating uses of cultural collections, to enrich those collections.

5.2.2. Digital library aggregation and interoperability

This session, led by Antoine Isaac and Amy Rudersdorf, focused on the barriers and solutions to aggregating digital collections in such a way that they become interoperable.

Participants discussed strategies for handling the differences between the preferred metadata standards of different kinds of cultural institutions. One approach involves top-down reconciliation of major, mutually relevant standards. For example, participants discussed an ongoing initiative to ensure that FRBR concepts are somehow mappable to concepts in CIDOC-CRM.[4] A second approach is transformation. The challenges of transformation garnered significant discussion among participants.

One participant noted that continuous processes of transformation are an inevitable aspect of maintaining interoperability between different kinds of data from different providers. Transformation is continuous, in that as standards evolve and the data themselves change, transformations will have to be altered and repeated. Part of the problem is subjectivity: even mapping between different implementations of the same standard can be challenging, as different institutions or collection developers interpret and apply standards differently. The process of transformation thus entails its own challenges, such as:

- Tracking the provenance of data through multiple transformations,
- Making the transformation process efficient, and
- Creating sustainable relationships between data providers and data aggregators.

Panelist Amy Rudersdorf noted that enforcing modularity in transformation processes (e.g., by breaking them down into subprocesses that are chained together) is one key to keeping the processes sustainable, flexible, and adaptable. Rudersdorf also noted a prototype tool for transforming DPLA data, which visually highlights anomalies for more efficient processing and communication between aggregator and provider.[5]

One participant suggested that detailed documentation of how data are transformed could be communicated to data providers as feedback. Data providers could use that information to improve data preparation for aggregation. With improved communication between aggregators and data providers, transformation could entail less labor (especially manual intervention) over the long term. The participant also suggested that the community might need a reference model that outlines different transformation processes and assigns responsibility for their implementation either to aggregator or provider.

Finally, participants discussed selection and collection policies for aggregations. Both DPLA and Europeana have grown rapidly through the institution of feeder hubs, which exist to assume some of the burden of aggregation, and to feed data from smaller institutions into the aggregation. DPLA has been opportunistic in data collection, relying on the selection policies of the contributing institutions. Participants acknowledged a tradeoff between rapid growth, which targets gaining critical mass, and selective growth, which prioritizes coverage and the curation of strong research areas.

[4] http://www.cidoc-crm.org/frbr_inro.html
[5] https://github.com/chadfennell/dpla.services

5.2.3. Data enrichment for collections

This breakout session, led by Paul Clough and Shenghui Wang, generated many questions about (and a few proposed solutions to) data enrichment for digital collections.

- What does data enrichment entail, and how does it differ from ensuring data quality? Participants suggested that to enrich data means to add value to it, but that this can take many forms. Enrichment could include adding new information to a data set, or removing noise. Adding information may mean linking to external resources (e.g., authority linked data sets), adding metadata (e.g., identifying named entities), linking to other collections, and normalization of data. Participants drew a very fine, and often ambiguous line between measures to improve data quality and measures for augmenting data.
- How do we ensure compatibility between actual use and measures taken to enrich data? For example, how do we consider who the users are and what they want in our enrichment processes? How do we orient enrichment toward effective downstream use?
- How do we ensure that data enrichment is done effectively and with transparency? For example, participants suggested representing the reliability, authority, or confidence of the person (or the algorithm) providing data enrichment. In addition, participants encouraged the preservation of the original alongside the enriched data. In addition, participants suggested that enrichment should be standards-driven and rigorously evaluated where possible.

5.2.4. Visualization of collections

This attendee-driven session, led by Karen Wickett and proposed by Marian Dörk, explored ways that visualizations can serve digital collections. Participants distinguished visualizations for different kinds of users, including scholarly and non-scholarly users of collections and aggregations, and even for the purposes of collection or aggregation administrators. In order to consider the possibilities for collection visualization more concretely, participants explored a handful of current visualization projects. Techniques that used subject descriptors to build dynamic views of library holdings seemed particularly applicable to collections in cultural heritage aggregations.

Participants also discussed the potential for visualizations of collections and aggregations to serve as research tools for scholarly interpretation. Visualizations designed to aid or facilitate public interaction with a collection can help with navigation, searching, browsing, and can help reveal what is unique or significant about a collection. Additionally, visualizations can be used by administrators, developers, and users to help characterize a collection. Participants agreed that there is no single, "correct" visualization of a data set that reveals all there is to know; rather, more visualizations yield more distinct views on a data set (or collection or aggregation).

6. Closing Discussions

A closing discussion among the full group of participants aimed to generate ideas for harnessing the workshop's momentum on research questions related to collections.

The participants expressed a strong desire to continue a dialogue on the questions raised in the workshop, and to form a community of interest on collections. A number of platforms for continuing discussion were proposed, including a wiki space, listserv, or discussion group. Participants suggested they would like to meet again at future iConference meetings, and workshop organizers are actively exploring areas for formalizing continued engagement.

Participants also discussed specific research agendas they plan to pursue, including:

- Reconstructing or generating anew the context lost when items are pulled from their original collections into large-scale aggregations
- Visualizing collections
- A reference model for data transformation, allowing aggregators and data providers to effectively share responsibilities for transforming and improving data as it is aggregated
- Evaluation to guide collection development efforts
- User-defined collections, and assessing and then exploiting users' criteria for grouping items together
- Collections in contexts other than cultural heritage, such as in science
- Dynamic representations of collections and items for different user needs
- Collections as tools for communication

As a first step toward solidifying what emerged, during the workshop, participants were invited to contribute post-workshop position papers for inclusion with this report. Sheila Corrall and Angharad Roberts submitted a response paper, titled “Collection as thing, process, and access: Two proposed models”, partially based on Roberts’ doctoral research at the University of Sheffield.

Collection as thing, process, and access: Two proposed models

Sheila Corrall[*] and Angharad Roberts[**]

[*]School of Information Sciences, University of Pittsburgh
[**]Knowledge and Library Services, Barts Health NHS Trust

Our position paper outlines two models of collection in the digital world presented in recent doctoral research. Both models are based on dimensions of collection as "thing", "process", and "access", identified using a mixed-methods research design including interviews, a survey, catalog searches, and a case study of the British Library's collection for the subject area of social enterprise. Our research revealed a considerable degree of shared understanding of the concept of "collection" by library and information professionals and ordinary people engaged in the field of social enterprise, whether users or non-users of library and information services.

The research [10] identified the following elements of collection:

- Collection as thing, including:
  - Collection as a group of materials
  - Collection as a group of subgroups (organised groupings)
  - Collection as quantity
  - Collection as container or store
  - Collection as a whole
- Collection as access, including:
  - Collection and connection
  - Collection for use
- Collection as process, including:
  - Collection as selection
  - Collection as search
  - Collection as service

Management Level	Collection definition	Example
Strategy	Collection as thing	Policies for: identifying and prioritizing subject areas; scoping collections (local and system-wide); collaborative collection development; preservation.
Tactics	Collection as access	Links to web-based materials and collections; interoperable systems; embedding libraries and librarians within non-library networks.
Operations	Collection as process	Support for community-created content; patron-driven collection; dynamic collection creation; linked data.

Table 1: Proposed collection development hierarchy [3].

Some elements echo earlier discussions of the concept of collection. For example, Lee identifies key characteristics of collection including "access" (p. 80); "selectivity" (pp. 72, 76); "subcollections" (p. 73) and "subject" (p. 76) [9]. Feather and Sturges (pp. 80-81) suggest collection can refer to "all the information resources to which a library has access" [5].

Collection as process is described in Atkinson's discussion of the "process of importation into the control zone," [1] and by Lagoze and Fielding's presentation of collection as "a set of criteria for selecting resources" [7].

The first model based on these dimensions of collection is described by Corrall and Roberts [3], elaborating on a collection development hierarchy based on Edelman [4], Gorman and Howes [6], and Corrall [2] to connect ideas of collection as "thing", "access", and "process" to the management levels of strategy, tactics and operations (Table 1).

In this model, "collection as thing" describes how the boundaries of collection are defined, whether in a physical, virtual or hybrid space. This space may be defined in relation to a single individual or organization, between a group of individuals, or across a range of organizations. "Collection as access" represents the tactics of encouraging and facilitating collection use, such as linking out to web-based content, or developing interoperable systems, such as those which enable movement between separate repositories. "Collection as access" also utilizes physical world tactics, such as printed QR (Quick Response) codes to link people viewing printed material to online content, or embedded librarians who can assist users' access to content in their own real world situations. Finally, "collection as process" describes operational level activities which support the creation, growth or reduction of collections. This element of the hierarchy may take the form of patron-driven acquisitions, dynamic collection creation based

on newly emerging areas of interest, or the automated inclusion or exclusion of particular items or objects (physical or digital) based on particular criteria.

The second model of collection reinterprets the dimensions of "collection as thing", "collection as access", and "collection as process". Instead being presented as elements of a hierarchy, the three aspects of collection are presented here as types of context about content within a collection. Lee (p. 1111) describes collection as context – "sometimes physical, sometimes institutional and sometimes intellectual" [8]. More recently, Wickett et al. demonstrate the value of representing contextual information in collection descriptions and show how aspects of context can be expressed using properties included in the Europeana Data Model. Some of these properties reflect the dimensions of collection described in this position paper, such as "access properties" (pp. 31-32) [11].

Examples of aspects of context suggested by the dimensions of collection as thing, access, and process include:

- Collection as thing
  - Grouped together with
  - Organised by/for
- Collection as access
  - Connected to/from
  - Used by/for
- Collection as process
  - Selected by/for
  - Searchable from/found by searching for
  - Presented as/delivers service

Within this model of collection, interactions with "collection as thing", "collection as access", and "collection as process" may add new context or remove existing context. These may include interactions by collection professionals or by users. Capturing changes in context over time – as well as describing intrinsic context derived from the original collection entity or the items of which it comprised – may add further value to collection content.

Our paper has described three dimensions of collection – "collection as thing", "collection as access", and "collection as process" – suggested by recent doctoral research. We suggest two models which apply these three dimensions: first, to suggest a new interpretation of an existing collection development hierarchy; and, secondly, to explore types of context which collection adds to content. Although developed with specific reference to library and information collections, these three dimensions of the concept of collection may have broader relevance and could be applied to other cultural collections in the digital world.

Acknowledgement

The doctoral research described here was funded at the University of Sheffield by a British Library Concordat Scholarship.

References

[1] Atkinson, R. (1996). Library functions, scholarly communication and the foundation of the digital library: Laying claim to the control zone. *Library Quarterly, 66*(3), 239-265.

[2] Corrall, S. (2012). The concept of collection development in the digital world. In M. Fieldhouse & A Marshall (Eds.), *Collection development in the digital age* (pp. 3-25). London: Facet.

[3] Corrall, S., & Roberts, A. (2012). Information resource development and "collection" in the digital age: Conceptual frameworks and new definitions for the network world. Libraries in the Digital Age (LIDA) 2012. Retrieved March 31, 2014, from http://ozk.unizd.hr/proceedings/index.php/lida2012/article/view/62/33

[4] Edelman, H. (1979). Selection methodology in academic libraries. *Library Resources & Technical Services, 23*(1), 33-38.

[5] Feather, J., & Sturges, P. (Eds.) (2003). *International encyclopedia of information and library science* (2nd ed.). London: Routledge.

[6] Gorman, G. E. ,& Howes, B. (1989). *Collection development for libraries.* London: Bowker Saur.

[7] Lagoze, C., & Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib Magazine, 4*(11). Retrieved March 31, 2014, from http://www.dlib.org/dlib/november98/lagoze/11lagoze.html

[8] Lee, H.L. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

[9] Lee, H.L. (2005). The concept of collection from the user's perspective. *Library Quarterly, 75*(1), 67-85.

[10] Roberts, A. (2013). Conceptualising the library collection for the digital world: A case study of social enterprise (PhD thesis, University of Sheffield). Retrieved March 31, 2014, from http://etheses.whiterose.ac.uk/5186/

[11] Wickett, K.M., Isaac, A., Fenlon, K., Doerr, M., Meghini, C.L., Palmer, C.L, & Jett, J. (2013). Modeling cultural collections for digital aggregation and exchange environments. CIRSS Technical Report 201310-1, University of Illinois at Urbana-Champaign. Retrieved from http://hdl.handle.net/2142/45860.

7. Conclusion

The workshop succeeded in rallying a community of interest around a topic of increasing importance to digital libraries and aggregations, especially but not only in the cultural heritage arena. The discussions in panels and breakout sessions throughout the day highlighted several important thematic questions for ongoing research on the representation of collections in digital aggregations:

- How can digital aggregations facilitate user-generated collections and other mechanisms that allow users to add curatorial value back into the system?
- Collections created by libraries, museums, and archives have high potential value, but in what contexts or under what conditions will this value exceed the cost of describing and representing collections in large-scale aggregations?
- How can we facilitate reuse of collections, while ensuring their trustworthiness and authenticity, and maintaining their provenance?
- What technical measures could improve the processes of data aggregation and data enrichment?

In addition, the workshop was successful in meeting the goals for fostering engagement and pushing forward the research agenda around collections in digital aggregations. The workshop broadened the conversation to an international audience by bringing together 38 registered participants from 15 countries on four continents. The ongoing challenge, for the whole community of interest, is to continue the conversations begun at the workshop in a formal venue, and with ongoing meetings at international conferences. The workshop furthered the research and development agenda for digital aggregations by distilling a wide-ranging and detailed set of research questions surrounding collections, and by helping participants identify and share directions for fruitful investigation. The morning's discussions on the conceptual foundations of collections helped contextualize the afternoon's conversations about the challenges confronting practical implementation in digital libraries or aggregations. By examining both the conceptual and practical aspects of collection representation and modeling, the workshop participants were able to explore realistic approaches for collection representation, contextualization, and interoperability at scale.

The challenges for creating representations and tools around collections in digital aggregations are rooted in conceptual, technological and organizational issues. The workshop participants brought a diverse set of perspectives on these issues, resulting in a refined set of directions and targets for an active community of researchers interested in advancing the state of the art for the representation and use of collections in digital environments.

8. References

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is 'quality' metadata 'shareable' metadata? The implications of local metadata practice on federated collections. In H.A. Thompson (Ed.) Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN. Chicago, IL: Association of College and Research Libraries.

Palmer, C. L., Teffeau, L. C., & Pirmann, C. M., (2009, January). Scholarly information practices in the online environment: Themes from the literature and implications for library service development. (Report). Dublin, OH: OCLC Online Computer Library Center, Inc.

Palmer, C. L., Zavalina, O., Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. In Proceedings of the ASIS&T Annual Meeting. (Pittsburgh, PA, Oct. 22-27).

Wickett, K. M., Isaac, A., Meghini, C., Doerr, M., Fenlon, K., Jett, J., & Palmer, C.L. (2013). Modeling cultural collections for digital aggregation and exchange environments. CIRSS Technical Report 201310-1, University of Illinois at Urbana-Champaign.

Appendix A: Workshop Organizers and Panelists

Sheila Anderson, Professor of e-Research, Centre for e-Research, King's College London. Anderson conducts research on digital libraries, digital archives, scholarly publishing, and digital assets management. She has an academic background in Sociology, Social Policy and Social History, and she has held a variety of leadership positions in association with the History Data Service at the University of Essex, the UK Data Archive, the Arts and Humanities Data Service (AHDS), and the Centre for e-Research.

Paul Clough, Professor, Information School, University of Sheffield. Clough is currently Scientific Director of an EU-funded project called PATHS (Personalised Access To cultural Heritage Spaces), running an AHRC-funded project on recommender systems for WorldCat.org with OCLC Inc. His research interests pertain to information storage and retrieval, particularly multilingual searching of texts and images; evaluation of retrieval systems; natural language processing, text reuse and plagiarism detection.

Martin Doerr, Research Director, Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas. Doerr leads the development of systems for knowledge representation and terminology, metadata and content management. He has led or participated in a series of national and international projects for cultural information systems, including the development of CIDOC CRM. He also holds a PhD in experimental nuclear physics from the University of Karlsruhe, Germany.

Katrina Fenlon, Research Assistant, Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Fenlon is currently pursuing a doctoral degree in library and information science from the University of Illinois, Urbana-Champaign where she also received an MS in 2009. Her research interests include national and large-scale digital library initiatives; digital collections, aggregations, and their users; metadata semantics and interoperability; and context and cohesion in and among cultural heritage resources online.

Antoine Isaac, Scientific Coordinator, Europeana Foundation. Isaac's research interests include the application of Semantic Web/Linked Data technology in the cultural heritage domain with an eye toward representation of metadata and controlled vocabularies and the use of ontological mappings to access heterogeneous collections. He completed his PhD studies at the French National Institute for Audiovisual and Université Paris Sorbonne.

Hur-Li Lee, Associate Professor, School of Information Studies, University of Wisconsin-Milwaukee. Her research interests include classification theory, traditional Chinese bibliography and knowledge organization, role of classification in scholarship, and social and cultural aspects of information services. She received both a MLS and a Ph.D. from Rutgers, the State University of New Jersey as well as a BA from National Taiwan University.

Carlo Meghini, Prime Researcher, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche. Meghini conducts research related to digital libraries, digital preservation, and conceptual modeling. Recent projects include Europeana version 1.0, Advanced Search Service and

Enhanced Technological Solutions for the European Digital Library (ASSETS), and Digital Library Interoperability, Best Practices and Modelling Foundations (DL.org).

Carole L. Palmer, Professor, Information School, University of Washington. Palmer works in the areas of data curation and digital research collections. Her research is aimed at advancing data services, especially for interdisciplinary inquiry, and she is particularly interested in optimizing the reuse value of small data and access to open data across disciplines. She holds a PhD in library and information science from University of Illinois at Urbana-Champaign and an MLS from Vanderbilt University.

Amy Rudersdorf, Assistant Director for Content, Digital Public Library of America. At DPLA, she is responsible for digitization partnerships and related workflows, metadata normalization and shareability, and community engagement to promote the DPLA as a community resource. Rudersdorf formerly served as the director of the Digital Information Management Program at the State Library of North Carolina. She was a Library of Congress National Digital Stewardship Alliance coordinating committee member and an active voice in the digital preservation community.

Megan Senseney, Senior Project Coordinator, Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Senseney is interested in the intersection between digital humanities and data curation, with a focus on scholarly information practices in the humanities. She holds an MS in library and information science from the University of Illinois, Urbana-Champaign.

Shenghui Wang, Research Scientist, OCLC Research. Her current research activities include text and data mining work as well as linked data investigations. Wang earned a PhD in Computer Science from the University of Manchester, a Master in Computer Application Technology at the University of Science and Technology of China (Hefei, China), and a Bachelor in Computer Science in Anhui University (Hefei, China).

Karen Wickett, Assistant Professor, School of Information, University of Texas at Austin. Wickett conducts research on the conceptual and logical foundations of information organization systems and artifacts. She is most interested in the analysis of common concepts in information systems, such as documents, datasets, digital objects, metadata records, and collections. She holds a PhD and MS in Library and Information Science from the University of Illinois at Urbana-Champaign and a BS in Mathematics from the Ohio State University.

Appendix B: Workshop Participants

1. Prof. Clément Arsenault, University of Montreal
2. Dr. Stephen Joe Asotie, University of Ibadan
3. Dr. Ian S. Brooks, University of Illinois at Urbana-Champaign
4. Prof. Katriina Byström, Oslo and Akerhus University College of Applied Sciences
5. Ryan Champagne, University of Pittsburgh
6. Prof. Sheila Corrall, University of Pittsburgh
7. Prof. Blaise Cronin, Indiana University
8. Femmy David Danyiwo, De Glowbiz Nig Ltd
9. Adam Abdelkader Diad, Algiers University
10. Louise Doolan, The British Library
11. Dr. Marian Dörk, Newcastle University
12. Dr. Yunfei Du, University of North Texas
13. Katrina Fenlon, University of Illinois, Urbana-Champaign
14. Dr. Jon Gant, University of Illinois, Urbana-Champaign
15. Prof. Anne J. Gilliland, University of California Los Angeles
16. Prof. Harriett E. Green, University of Illinois at Urbana-Champaign
17. Shah Hussain, University of Swat
18. Dr. Perla Innocenti, University of Glasgow
19. Ji Hei Kang, Florida State University
20. Sumi Kim, Seoul National University
21. Lesley Langa, University of Maryland
22. Prof. Gregory Leazer, University of California Los Angeles
23. Amalia Skarlatou Levi, University of Maryland
24. Dr. Michael John Mertens, Research Libraries UK
25. Dr. Steven Miller, Singapore Management University
26. Nathan Michael Moles, University of Toronto
27. Prof. Mitsuharu Nagamori, University of Tsukuba
28. Kaori Ochiai, University of Tsukuba
29. Johna Louise Picco, University of Illinois
30. Dr. Franziska Regner, Deutsche Forschungsgemeinschaft
31. Dr. Andrea Scharnhorst, DANS
32. Prof. Daniel Schiller, University of Illinois, Urbana-Champaign
33. Dr. Edward Francis Schneider, University of South Florida
34. Michael Svendsen, University of Copenhagen
35. Prof. Yasar Tonta, Hacettepe University
36. Dr. Yukiko Watanabe, Kyushu University
37. Stella Wisdom, British Library
38. Prof. Ying Ye, Nanjing University

Appendix C: Workshop Schedule

Digital Collection Contexts
iConference 2014 Workshop

Berlin, Germany March 4, 2014
Room 1.406
Humboldt-Universität zu Berlin
Seminargebäude am Hegelplatz
Dorotheenstr. 24
10117 Berlin

Organized by:

- Carole, L. Palmer, Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois
- Antoine Isaac, Europeana Foundation
- Karen Wickett, School of Information, University of Austin at Texas

Abstract

This full-day workshop examines conceptual and practical aspects of collections and the context they provide in the digital environment, especially in large-scale cultural heritage aggregations. Collections will be considered in relation to the information needs of scholars, roles of cultural institutions, and international interoperability. The workshop aims to:

- Broaden the conversation across an international community
- Further the research and development agenda for digital aggregations
- Relate conceptual advances to implementation goals
- Identify realistic approaches for collection representation, contextualization, and interoperability at scale

Trends in interoperable content and open data raise important questions on how to represent complex objects, curated and dynamic collections, and context in ways that benefit users and collecting institutions. This workshop will provide a forum for international engagement on this important topic and provide iSchools the opportunity to build a community around our strengths in this important research area. Sessions will be led by European and North American experts from iSchools and projects developing large-scale digital cultural heritage collections.

We encourage participation from:

- Faculty and students from iSchools involved in research and education in information organization, cultural heritage, digital collections and archives, and metadata.
- System designers and developers interested in the creation of metadata schemas and promoting interoperable digital cultural heritage content.

Workshop Schedule

Opening Remarks 9:30-9:45
On digital collection contexts, Antoine Isaac and Carole L. Palmer

Conceptual Foundations Panel 9:45-11:00
Moderator: Carole L. Palmer

- The notion of collection: A retrospective overview, Hur-Li Lee, School of Information Studies, University of Wisconsin-Milwaukee
- Is collection modeling contextual modeling?, Karen Wickett, School of Information, University of Texas at Austin
- Unity criteria of collection contexts: Why are items together?, Martin Doerr, Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas
- On the logical foundations of digital collections, Carlo Meghini, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche

Breakout Session Organization 11:00-11:10

Break 11:10-11:25

Breakout Sessions 11:25-12:05
Topics will be finalized by panelists and attendees.

Discussion Report Back 12:05-12:30

Lunch 12:30-14:00

Practical Implications Panel 14:00-15:15
Moderator: Antoine Isaac

- Implementing collection contexts, and metadata issues related to normalization and shareability, Amy Rudersdorf, Digital Public Library of America
- The CENDARI Knowledge Framework: a dynamic research environment and a grown knowledge framework, Sheila Anderson, King's College London
- Hunting for semantic clusters in aggregation, Shenghui Wang, OCLC Research
- Supporting users' navigation and exploration of large digital collections: experiences from the PATHS project, Paul Clough, Information School, University of Sheffield

Breakout Session Organization 15:15-15:25

Break 15:25-15:40

Breakout Session 15:40-16:20
Topics will be finalized by panelists and attendees.

Discussion Report Back 16:20-16:45

Reflective Discussion 16:45-17:15

Closing Remarks 17:15-17:30
Carole L. Palmer

Background Reading

Modeling Cultural Collections for Digital Aggregation and Exchange Environments, a whitepaper developed by researchers from the Europeana Foundation and CIRSS that discusses functions of collections in cultural heritage aggregations and proposes a formal extension to the Europeana Data Model to explicitly accommodate representation of collections and collection/item relationships. A public release of the paper is available at http://hdl.handle.net/2142/45860.

For additional information about the workshop, please visit http://bit.ly/collectionsworkshop2014.