MODELING CULTURAL COLLECTIONS FOR DIGITAL AGGREGATION AND EXCHANGE ENVIRONMENTS

KAREN M. WICKETT[1]

ANTOINE ISAAC[2]

KATRINA FENLON[1]

MARTIN DOERR[3]

CARLO MEGHINI[4]

CAROLE L. PALMER[1]

JACOB JETT[1]

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

CIRSS

[1]Center for Informatics Research in Science & Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana Champaign

[2]Europeana Foundation

europeana
think culture

[3]Institute of Computer Science (ICS)
Foundation for Research and Technology - Hellas (FORTH)

[4]Instituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche

INSTITUTE of Museum and Library SERVICES

Partially funded by IMLS National
Leadership Grant
LG-06-07-0020

CIRSS TECHNICAL REPORT 201310-1

OCTOBER, 2013

TABLE OF CONTENTS

1. INTRODUCTION

Curated collections are the essence of memory institutions. Libraries, archives, and museums, in particular, curate many collections, unified by their material nature or intellectual content, or both. Curation means that the institution is responsible for creating and caring for the collections—for selecting, augmenting, preserving, documenting, and researching items that keep memories relevant to humanity and that pertain to a range of disciplinary interests. With the advent of information technologies, cultural heritage institutions have been moving paper-based documentation, and increasingly intellectual content as well, into digital formats managed by IT systems. Consequently, collections have become a fundamental feature of digital information organization systems in this sector. Collection structures and descriptions provide a variety of useful functions for users and managers of digital libraries, including technical capabilities for retrieval and evaluation of content, especially within large digital environments that aggregate many collections.

Collection structures provide the organizational and intellectual context important to researchers, and collection descriptions provide information needed by users for interpreting the relevance and significance of individual items for their purposes. Collections are also important representations of institutional identity for the organizations that invest in digitization and curation to provide public access to their special materials. Moreover, with public access to digital materials, individuals can now also build collections drawn from any number of institutional collections. Fostering a deeper level of user engagement with large digital aggregation systems is a promising area for further technical development.".

This report presents the results of a collaboration between members of the IMLS Digital Collections and Content (DCC) project and developers of the Europeana Data Model (EDM) to construct a formal extension of EDM that explicitly accommodates representation of collections and collection/item relationships. The goal is to enhance the representation facilities of EDM, and to make EDM conducive to representing collection-level data from DCC and other digital content providers. Here we report on the outcomes of the collaboration – use cases, requirements, and recommendations for modeling collections in exchange and aggregation environments – prefaced by a short section covering background on the foundational DCC and Europeana initiatives and an overview of related work in the field.

2. BACKGROUND

As recent technological innovations in web architecture introduced new methods of linking content and engaging users with digital materials and one another (Heath and Bizer, 2011), the Digital Collections and Content (DCC) and Europeana initiatives were working independently on developing large-scale cultural heritage aggregations for public access. The two 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 provider. Both groups have made progress on the problems associated with harvesting and integration of content from many diverse institutions as well as functionality for users to search, browse, and engage with content.

Synergies between the two initiatives were first explored in a one-day workshop held in Crete in May 2011 in conjunction with the European Semantic Web Conference, resulting in ideas on adapting the DCC data representation approach to be compatible with EDM and possible ways to extend EDM, with the aim of supporting international interoperability. A second three-day working meeting, held at the University of Illinois at Urbana-Champaign, March 7-9, 2012, resulted in an outline and detailed plan for production of this white paper. The coordinated data modeling effort is intended to advance interoperability between the two resources and with other aggregations, such as the Digital Public Library of America.[1] The advances also have the potential to support faceted information retrieval, topic modeling and other clustering techniques exploiting linked data and RDF, and utilization of relationships between collection-level and item-level representation to enhance functionality for users and developers of large-scale aggregations.

2.1 IMLS DCC

The IMLS Digital Collections and Content project (DCC) is a collaboration between researchers at the Center for Informatics Research in Science and Scholarship and the University Library at the University of Illinois at Urbana-Champaign, funded by the Institute for Museum and Library Services (IMLS). Originally proposed in response to an IMLS RFP in 2002, the resource was initially conceptualized as a collection registry combined with a repository for item-level metadata, to provide a single point of access to all the collections digitized with funding from IMLS. Starting in 2007, the DCC expanded its scope beyond IMLS funded content and continued research and technical advances on metadata, interoperability, aggregation workflows, collection evaluation, subject access, and usability. The DCC is now among the largest and most diverse cultural heritage digital aggregations in the country. At present the aggregation contains collection-level and item-level metadata records representing cultural heritage objects and collections for nearly 1500 cultural heritage institutions, large and small, across 46 states and 3 U.S. territories, with 1737 digital collections and over 1.2 million items.

One significant outcome of the DCC has been a data structure that supports representation of collection entities and the contextual information provided by collection-level description. The DCC collection-level schema was originally adapted from the Research Support Library Programme (RSLP) collection-level schema[2] and has since been aligned the Dublin Core Collections Application Profile.[3] This architecture has proven to be vital to how users identify and understand individual digital objects and

[1] http://dp.la/

[2] http://www.ukoln.ac.uk/metadata/rslp/schema/

[3] http://dublincore.org/groups/collections/collection-application-profile/

how they comprehend the nature of content available to them within a large aggregation, as well as for retaining the identities of special collections and their institutions in a large digital aggregation on the open web.

The Collection/Item Metadata Relationships (CIMR) group was formed as part of the DCC to investigate logical relationships between collection-level and item-level metadata and to explore how automated processes and tools can make the most of both types of metadata to improve access and use of digital content (Renear, et al. 2008; Wickett, et al. 2010). One primary result of the CIMR project and the continued research reported by Wickett (2012) is a method for expressing relationships between collection-level and item-level descriptions as *propagation rules* along with a framework for organizing rules according to their logical features. These categories and the inference rules can be used to supply detailed semantics for metadata vocabularies at the collection and item levels and to aid in the construction of collection-level metadata records from item-level information.

2.2 EUROPEANA AND EDM

Europeana brings together the digitized content of Europe's galleries, libraries, museums, archives, and audiovisual collections. Currently Europeana gives integrated access to over 26 million books, films, paintings, museum objects, and archival documents from some 2,400 content providers. The content is drawn from every European member state. Europeana.eu is a search portal that provides an interface to this wealth of resources in 29 European languages. Europeana, which receives its main funding from the European Commission, is committed to providing a platform for culture that is accessible for all. In addition to the portal, it works on providing core services, such as an API based on fully open metadata.

The Europeana Data Model (EDM)[4] is the schema underlying Europeana's data ingest, management, and publication. EDM has been developed and maintained by the Europeana community. It aims to standardize representation of heterogeneous records while supporting (1) the description of digital resources and data ingestion processes separately from those for the description of original cultural objects, (2) the retention of complete item descriptions from data providers, (3) data enrichment by Europeana and third parties, leading to multiple records for the same object, (4) the description of complex objects, (5) linking objects to other resources (concepts, places, persons…) related to them, potentially described by third-parties.

EDM prominently features three classes of resources:

- Provided Cultural Heritage Objects or CHOs (*edm:ProvidedCHO*) denote the original objects—either physical (e.g. a painting, a book, etc.) or born-digital (e.g. a 3D model), which are the focus of description and search in Europeana. The choice in granularity of description chosen for the ProvidedCHO belongs to the data provider, within the limits of relevance set by Europeana.
- Web Resources (*edm:WebResource*) represent digital representations of the provided CHOs, published on the web.
- Aggregations (*ore:Aggregation*) group the Provided CHO and the Web Resource(s) into one bundle, where information on the aggregation process is also recorded (e.g., the provider of the data).

EDM also defines contextual resources that can be used to provide more information related to the object (e.g., *edm:Agent*, *edm:Place*, *edm:Concept*, *edm:TimeSpan*).

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

Note that in EDM *ore:Aggregations* are also used as context to create perspectives on CHOs ("proxies") that carry provider-specific data on these objects, thus allowing one to separate it from data on the same object from other providers (including Europeana). Therefore *ore:Aggregation* is primarily used in the model to serve as an organizing construct for repository managers and to aid in interoperability, by providing assistance for harvesting or integration.

While many of Europeana's data providers maintain collection-level entities or descriptions (e.g. The European Library[5] and the European Film Gateway[6]), Europeana itself does not make use of or preserve collection-level information. The primary goal of this paper is to examine the technical requirements for preserving, reconstructing, and building collection-level entities within the Europeana context.

[5] http://www.theeuropeanlibrary.org

[6] http://www.europeanfilmgateway.eu/

3. RELATED WORK

From the institutional perspective, collections are imbued with significance and paramount to the role of cultural heritage institutions in society. Archives, libraries and museums have their own disciplinary methods for managing collections, and there are national laws regulating some of the responsibilities and handling of physical collections, as with laws for sites and monuments records to protect immobile cultural heritage. The ICA (International Council of Archives) maintains a set of ISO standards (ISAD G, ISDF, ISDIAH, ISAAR). IFLA provides international cataloguing rules and other recommendations for library collections. Museums tend to follow SPECTRUM,[7] the prescription of collection management processes from the British Collections Trust. The three international organizations ICA, IFLA and ICOM have never engaged in any common definition of what a collection is. The very terms "archive", "library", or "museum" often are synonyms for the physical collections themselves. Recently, however, the intellectual commonalities behind the diverse materiality of collections has become more obvious in their digital representation, and there is new interest in multidisciplinary knowledge exchange on the role and importance of the collections construct.

The introduction of digital resources into library catalogs was an opportunity to examine how collection development and management functions were addressed in the library domain (Buckland, 1995; Atkinson, 1998). More generally, with digitization came an opportunity to reconceptualize the collection beyond traditional notions rooted in physical proximity (Lee, 2000; Casserly, 2002), to evaluate the sufficiency of collection development and evaluation processes for digital resources (Covi and Cragin, 2004), and to redefine roles and responsibilities around collection management in digital environments (Kaczmarek, 2006). Yet, while digital content has grown and become increasingly accessible, and scholarly discourse on collections has intensified, a consistent definition of collection has not emerged (see, for example, Hill et al., 1999; Lee, 2000; Currall et al., 2004; Wickett et al., 2010).

Despite the lack of a widely agreed-upon definition of collection, it is clear that in many cases, collections themselves are the entities that meet the information needs of researchers. For example, Zavalina (2010) found clear transaction log evidence of searches specifically for collections in the IMLS DCC aggregation. While these collection-level searches were less than half as common as item-level searches, nearly one third of queries (880 out of 2740 queries in a 12-week sample) were performed to find collections rather than items. Without any explicit representation of collections as individual objects that can be searched for directly, users cannot reliably find and identify collections. Representing collections as entities in aggregations allows these carefully curated groupings to maintain their identity and to be indexed and retrieved as coherent objects. Studies of how collections are used have demonstrated how the environment of a collection aids the information seeking process (Lee, 2000) and the need for user-centered flexibility in collection structures (Lee, 2005).

The creation of collections is an important activity performed by scholars as part of their research process. In the digital era, and especially in the humanities, these collections are of value to larger research communities and are now becoming scholarly products in their own right (Palmer, 2004). Such "personal collections" may be built with similar criteria as those professionally created by memory institutions, but they may also have a much more speculative nature. Scholars may travel long distances to track down a source of importance in a distant archive, or collect items only loosely relevant to a context or concept that is unfolding in an area of interest (Brockman, et al. 2001; Palmer, 2005). More specifically, collections created by scholars for research purposes, while similar in their thematic nature to special collections in cultural heritage institutions, are distinguished by the "contextual mass" that results from how interrelated, diverse sources work together to support deep

[7] http://www.collectionslink.org.uk/spectrum-standard

inquiry in an area of research (Palmer, et al., 2010). Additionally, many student projects result in interesting collections following the project prescription, and more and more casual users of electronic media make use of the capabilities of IT services to exchange information in the form of collections.

Most museum curators and conservators privately maintain collections of documents or other objects that relate to a specific theme or activity, which are often referred to as "folders" (Low and Doerr 2010, Doerr et al. 1997). Low and Doerr studied the internal and external knowledge collection and transfer processes of several museums. They argue that digital representations of museum collections for research and public use should differ from the traditional institutional documentation practice and present the relevance of items under multidisciplinary views.

Collections are also powerful educational tools that can the meet information needs of educators and students. While humanities researchers have a long history of using archives, the availability of digital content has facilitated the use of primary sources in education, with more students being introduced to and interacting with archives, special collections, and digital exhibits. Since exhibits and collections offer interpretive content and showcase only carefully selected materials, they offer the student guidance through a topic and can frame the resources within a historical context (Gueguen, 2010).

Thus, digital collections can take many forms, including interactive exhibits or online tours, with open-ended potential for the creation of new collections from multiple, distributed content providers (Palmer et al., 2006). This flexibility, however, calls into question what might qualify as a collection in the digital arena, to which degree representations of physical collections in digital form are also collections in their own right, and more generally whether the term is linguistically overloaded with multiple senses that are not reducible to a common core. One interpretation is that any set of resources meeting a set of criteria qualifies as a collection ("set membership ... is ... criteria-based") (Lagoze & Fielding, 1998). Geisler et al. (2002) proposed the development of "virtual collections" within digital libraries, which were conceptualized as sub-collections of digital library collections based on a common attribute or relation to a common subject. These approaches are not restrictive enough, however. For example, they could not necessarily distinguish a group of items retrieved through an online search from the kind of collections that are developed by libraries, archives, museums through systematic selection of items, or the research collections created by scholars, or the other collections intentionally crafted by individuals, groups, and organizations.

There have been many arguments in favor of the usefulness of collection description for institutional administration and for supporting scholarship (Brack et al., 2000; Sweet and Thomas, 2000). Collection descriptions are designed to provide a range of information specific to the collection as a whole, such as creator, location, formats, extent, audience, access rights, collection policy, provenance, etc., creating a context that aids scholars in identification, interpretation, and use of items within a collection. Collection-level metadata can re-contextualize orphaned items by providing access points that are lacking in item-level descriptions (Foulonneau et al., 2005). Contextual information may include an account of relationships between a set of documents or information about how archival records are organized. As noted by Duff and Johnson (2002):

> "The totality of the records provides information that no individual record can. Historians must comprehend the records in their context rather than as separate disembodied items. Without this context information, the historian could easily misinterpret the meaning or significance of the information in an individual record."

It is important to provide for the recording and presentation of contextual information, so that scholars may understand resources as being situated in a context that arises either from external (e.g., historical or geographic) associations or the provenance of the resource itself. Contextual metadata has long been recognized in the archival community as being central to facilitating access to documents in archival collections (e.g., Bearman, 1992).

Heaney (2000) developed an analytical model for describing collections that informed the creation of several schemas for collection description (Shreeves and Cole, 2003). Some of the most well-known and widely used schemas for cultural heritage materials that allow for the representation of collections are the Dublin Core Collections Application Profile (Dublin Core Collection Description Task Group, 2007), RSLP (Research Support Libraries Programme) (Powell, 2000), NISO Z39.91-200x (National Information Standards Organization), and Encoded Archival Description (Library of Congress, 2002).

Utilizing collection descriptions to full advantage for technical capabilities and user experience is an important area of research and development, especially for repositories that include resources from multiple sources. In particular, Lourdi et al. (2009) has developed an approach for the integration of collection descriptions from different schemas based on an ontology of cultural heritage materials. Metadata techniques being advanced by the Dryad project are also of interest. Although their content focus is quite different—data associated with published research—their aim of implementing metadata propagation and inheritance functionality (Greenberg, 2009) relates to approaches explored by the DCC for exploiting relationships between collection- and item-level metadata (Wickett, Renear, Urban, 2010). Additionally, contextual metadata can play a critical role in the preservation of digital objects (Beaudoin, 2012), and as we argue in the next section, collection-level information can serve as important contextual information for items.

The representation and description of resources in distributed information environments calls for clear distinctions between the various stewardship roles taken on by participating institutions. The Metadata Encoding and Transmission Standard (METS)[8] for objects in digital libraries includes fields that differentiate between certain stewardship roles involved in the maintenance and dissemination of digital objects. In particular, METS captures information about the many agents responsible for a METS document—those responsible for preparing metadata for encoding, for the document or collection being described, for preservation functions, and for dissemination functions. However, these fields attend more directly to recording information about metadata records than cultural resources themselves. The stewardship roles discussed in Section 6 have some overlap with the roles documented by METS, but are specifically designed to capture stewardship of collections in digital aggregations.

A number of digital library efforts to formalize objects and relationships also have important implications for collection data modeling. For example, one of the best-known formal models, Streams, Structures, Spaces, Scenarios, Societies (5S), provides a comprehensive mechanism for modeling every aspect of a digital library within a cohesive set of mathematical formalisms (Gonçalves et al., 2004). The digital library model developed by Meghini et al. (2010) is explicitly based on first order logic, addressing digital objects, descriptions of those objects, and the schemas from which descriptive terms are drawn. In addition, the CIDOC Conceptual Reference Model (CRM) provides an ontology that supports the integration of descriptions of cultural heritage objects from multiple sources (Doerr, 2003; CIDOC, 2010) and is also an ISO standard (ISO21127).

In 5S, collections are modeled as mathematical sets of digital objects. Although the model provides for explicit accounting of metadata describing digital objects and for a catalog of metadata that pertains to the objects in a collection, there is no explicit allowance for collection-level description. Gonçalves et al. (2004) does not discuss whether collections can themselves be treated as digital objects, but it does not appear to be the case since the authors suggest that the description of a collection happens only by

[8] http://www.loc.gov/standards/mets/

virtue of descriptions of the digital objects that are members of the collection. In contrast, Meghini and Spyratos (2007, 2010) attend to collections directly, arguing for a distinction between collection extension, modeled as a function that assigns a set of documents to a collection, and intension, given as a function that assigns a description (a set of terms) to a collection. The assignment of descriptive terms to collections, however, is simply via the terms assigned to members of the collections.

The CRM is intended to be comprehensive and applicable to a wide range of cultural heritage materials. For that purpose, it defines concepts that have been empirically recognized in relevant cultural heritage documentation as a common reference for information integration. The current version of the standard defines a concept of "E78 Collection" as aggregations of physical things. The model derives its concept of collection from the intentions of curators in creating collections, stating that they "are assembled and maintained (curated and preserved, in museological terminology) by one or more instances of E39 Actor over time for a specific purpose and audience, and according to a particular collection development plan" (CIDOC, 2010). In addition, "items may be added or removed from an E78 Collection in pursuit of this plan." Thus, it is clear that the CRM aims to meet intuitive expectations about the creation and maintenance of collections as information organization artifacts. However, it has not yet been verified if the CRM representation of collections as physical objects extends to other uses or senses of "collection" relevant to large-scale digital aggregation and exchange scenarios.

4. CHARACTERIZING THE CONCEPT OF *COLLECTION*

In this section we aim to constrain the concept of *collection*, to define more clearly the kinds of collections to which our modeling recommendations apply. Throughout this paper, our focus has been on cultural heritage collections, whether gathered by an institution, such as a library, archive, or museum, or by an individual for personal purposes. Even within these constraints, collective objects assume many forms: virtual exhibitions, mash-ups, portals, groupings of user-provided content, pinboards, bookmark lists, and even bibliographies may be considered examples of collections. For the purposes of the modeling requirements and recommendations discussed here, our concept of *collection* emphasizes (a) the collecting process, (b) the curatorial or intellectual intent behind a collection, and (c) the premise that while collections do not have substantive content beyond their items, they are meaningful information objects in their own right.

A collection is a group of objects gathered together for some intellectual, artistic, or curatorial purpose. In addition, we limit our attention to collections that satisfy the following constraints:

1. The collection has members that have been gathered together in the past or will be gathered together in the future.
2. Membership in a collection is determined by some criteria that fit the purpose or intentions of the collector.
3. The collection may be treated as an individual object for purposes of description, access, and curation.

This is a broad conception of collections, which may be further divided into more specific kinds of collections. These specific kinds may be defined according to the stewardship relationship a collector expects to take on with respect to the items in a collection, or according to the particulars of the criteria used to determine collection membership.

4.1 HOLDINGS COLLECTIONS AND REFERENTIAL COLLECTIONS

In the digital age, there are many important collections created by institutions and even individual scholars that do not imply that the creator has taken over ownership or custodianship of the items gathered into the collection. This is in contrast to the institutional stewardship relationship between collections and individual items frequently found in museum exhibits, special collections, archives, and general library collections that are produced and maintained by librarians, archivists, and curators.

Given the difference between what can be inferred by membership, it seems useful to distinguish between collections that do not comprise the items themselves, but only reference them, and collections that directly comprise their items.

- **Holdings Collection:** A collection of items in the custody or control of an organization or curator.
- **Referential Collection:** A collection referring to rather than directly holding its items.

The distinction between holdings collections and referential collections is relevant for determining rights over and access to the content brought together by a collection. It is also relevant for reasoning, and in general, membership in a holdings collections may be used to infer more facts about individual items than membership in referential collections. For example, an institution providing access to a collection will generally also be able to provide access to the individual items (where technically feasible), but a researcher providing access to a referential collection may not have the appropriate rights to give access to the items within that collection. In general, there are no reliable means to guarantee complete and exact long-term access to material that is referred to rather than held physically by the collector.

It is worth considering whether to make a parallel distinction between institutionally developed collections and collections developed by private individuals, i.e. whether "amateur" collections follow the same principles as "professional" ones. Private collectors are frequently much more "scientific" than commonly assumed, and they may differ from institutional collectors more in what they regard as relevance, rarity etc., than in the type of questions that motivate their collecting activities. These collectors may tend to collect a special category of things rather than related objects. Bekiari, et al. (2008) have shown that the most complex collections are in small museums, which are typically more bound to a local context than to a global theme.

Such a view, motivated by the behavior of collectors of physical collections, renders the distinction between a holdings collection and a referential collection less clear—except for questions of acquiring actual content. The more we restrict the intellectual form of what we call a referential "collection," the more likely reasoning based on unity criteria will be the same for physical holdings collections and referential collections.

Overall, the differences between physical holdings collections and referential collections are not significant enough at the general level to justify fundamentally different modeling approaches, or to exclude collections created by individuals for personal purposes from participating in the functional roles of collections in aggregation and exchange environments. The usefulness of collection descriptions in the scenarios described in the following section depends on the quality of description, which may be just as high for a referential collection as for a collection where items are held by the collecting institution.

4.2 UNITY CRITERIA

We refer to the criteria that determine whether an item is gathered into a particular collection as *unity criteria*. These criteria are a formulation of the decision-making process that guides the development of a collection and captures the curator’s intent. These criteria are relevant for use and interpretation of individual items within a collection. In addition, unity criteria could be used to support collection-level-to-item-level reasoning in cases where items that are gathered together can be characterized by criteria that allow for inferences about items based on their membership in a collection.

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”. Although these titles are suggestive of the potential relevance of a collection with respect to a cultural phenomenon, they may not provide reliable evidence in general. 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 (2009).

The British Museum's founding collection was the 71,000 books, antiquities, and natural specimens bequeathed to the nation by Sir Hans Sloane[9] in 1753. It is maintained as a collection within the museum’s holdings. The unity criterion is the “common collector”. By virtue of that, the collection membership is evidence of what Sir Hans Sloane (and possibly his contemporaries) had known and evidence of his research interest. Standard criteria are space-time constraints, culture constraints, and object type constraints. For instance, the Sir John Beazley Archive in Oxford contains the world's

[9] http://www.britishmuseum.org/about_us/the_museums_story/sir_hans_sloane.aspx

largest collection of photographs of *ancient Greek painted pottery* (combining an object type with a temporal and cultural constraint).

More formally, we can distinguish four general categories, similar to the types of relevance described in Bekiari, Doerr and LeBoeuf (2008), that may be used to determine whether an individual item is suitable for membership in a collection:

1. Nature: The individual construction or form of an item provides evidence or information about the context of its creation, or means that the item is likely to be of significant value over long periods of time.
2. Example: An item exemplifies a particular category or type of thing.
3. Witness: 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 or content to some person, object, place, event, or phenomena of interest.

Examples of items included in collections on the basis of their nature include fine art objects, scientific equipment, manuscripts, and unique archaeological finds. Items that meet unity criteria based on exemplification include objects such as natural history specimen, a set of ethnological material, and individual objects from an archaeological mass find. A collector might use unity criteria based on witness and historical presence to select relevant objects for a historical heirloom collection or for the curation of the personal library of a famous scholar. Aboutness criteria have shaped many familiar subject-based collections that feature items like busts of Roman emperors, inscribed stones, birth registers, letters, or literature.

The four categories provide an intellectual basis for determining collection membership, and an item may be included in a collection due to a combination of reasons rooted in the categories. The categories are not fully independent, but are intended to emphasize core aspects of decision-making about collection membership. For example, it could be argued that Example and Witness are variations of Nature. On the other hand, the kinds of inferences that may be drawn from the collection context provided by a collection based on Aboutness are likely to be distinct from those that arise from collections based on Nature, Witness, or Example.

The context of interest may be described by restriction to a particular time-span or place. It may be restricted further to a particular thing, actor, event or place; a type of things, actors, events or places; or any reasonable combinations of those. These restrictions constitute a major focus of collection-level attributes that may propagate to the item level or at least inform the item level. In addition, knowledge about the collector may allow for inferring relevant knowledge even without explicit collection criteria.

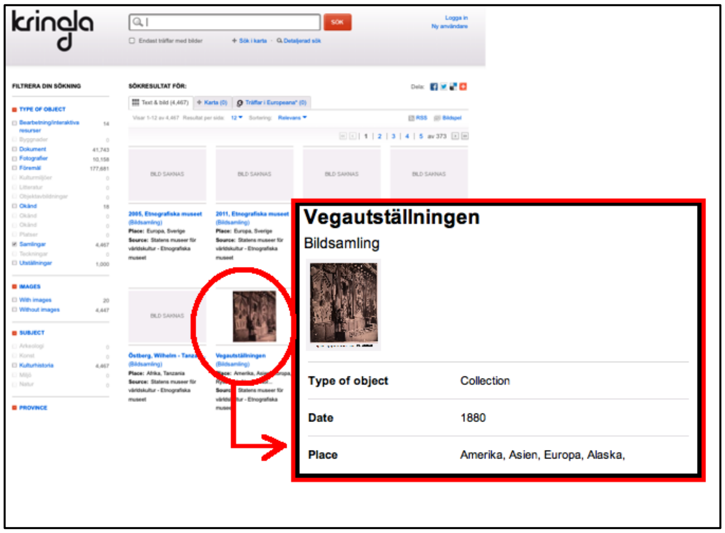
The distinction between referential collections and physical holdings collections, and the exploration of the dimensions that characterize unity criteria, are contributions to the development of a rigorous definition of the concept of *collection* to support the functional and intellectual roles of collections in digital aggregations and exchange environments that are discussed in the following section.

5. THE ROLES OF COLLECTIONS AND COLLECTION DESCRIPTION IN AGGREGATION SCENARIOS

There are many ways that collection-level entities and collection descriptions support users of digital aggregations, as well as the interests of content providers and the operational side of access services and collection development for aggregations. Here we present several selected examples, focusing primarily on the user experience.

5.1 REPRESENTING DATA PROVIDERS

Swedish Open Cultural Heritage (SOCH) is the Swedish national aggregator of cultural heritage collections[10] and currently provides more than one million item records to Europeana. The SOCH portal (see Figure 1) represents not only cultural heritage items but also collection objects (*samling*) that can represent collections or exhibitions.



Vegautställningen

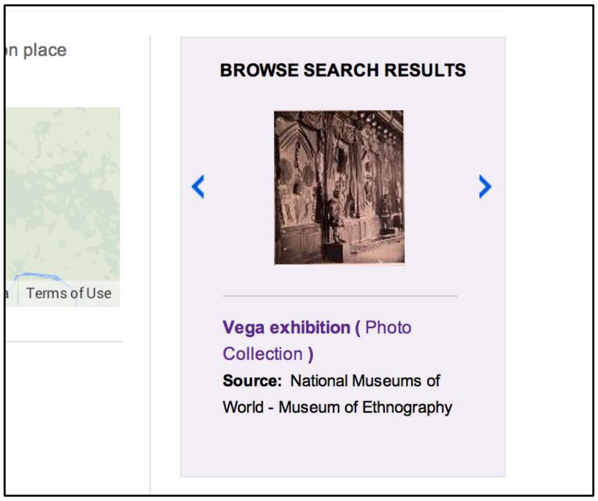
Bildsamling

Type of object	Collection
Date	1880
Place	Amerika, Asien, Europa, Alaska,

FIGURE 1: ACCESS TO CULTURAL COLLECTIONS

Clicking on a given collection-level object gives a detailed record. Thumbnails of items (or sub-collections/-exhibits), which can be expanded, are also displayed on the page. Because Europeana does not yet represent collections, items provided by SOCH lose collection context as item metadata is mapped to EDM. SOCH is one among many Europeana data providers that stand ready to benefit from collection-level representation in EDM.

[10] http://www.kringla.nu



Vega exhibition (Photo Collection)

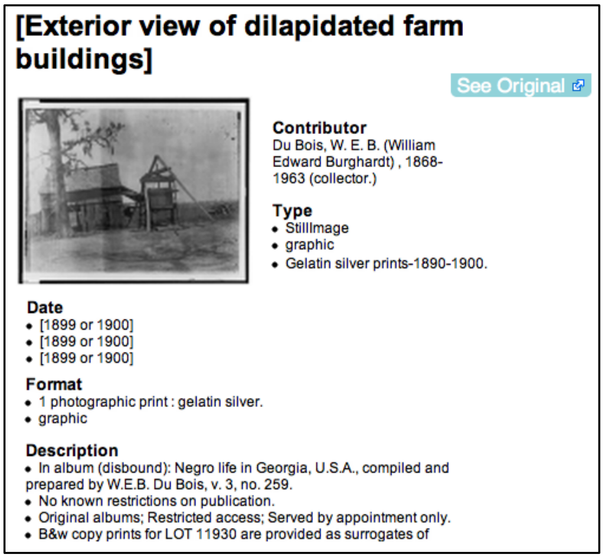
Source: National Museums of World - Museum of Ethnography

FIGURE 2: REPRESENTATION OF CONTENT PROVIDER

In the SOCH aggregation, the content provider is represented through the collection-level metadata with a link that allows users to access the collection and its individual items in the original context hosted by the content provider. In the case of Figure 2, the content provider is the National Museum of Ethnography. This kind of representation and linking increases the institutional presence of providers in aggregation systems and allows users to access content from both aggregations and institutional providers.

5.2 PROVIDING CONTEXT FOR ITEMS

Figure 3 shows an item record abstracted from its context in the IMLS DCC aggregation. Without collection-level information to accompany the item record, this historical photograph offers a compelling but rather uninformative image of a dilapidated farm structure. While the item's description field provides one obscure clue to the wider context of the item ("In album (disbound): Negro life in Georgia..."), only an unusually dedicated user might glean the implications of this statement or, alternatively, seek further evidence. Why should a user be interested in this photo, other than for its aesthetics or age? What is the significance of this picture? Why was it worth collecting?



[Exterior view of dilapidated farm buildings]

Contributor
Du Bois, W. E. B. (William Edward Burghardt) , 1868-1963 (collector.)

Type
- StillImage
- graphic
- Gelatin silver prints-1890-1900.

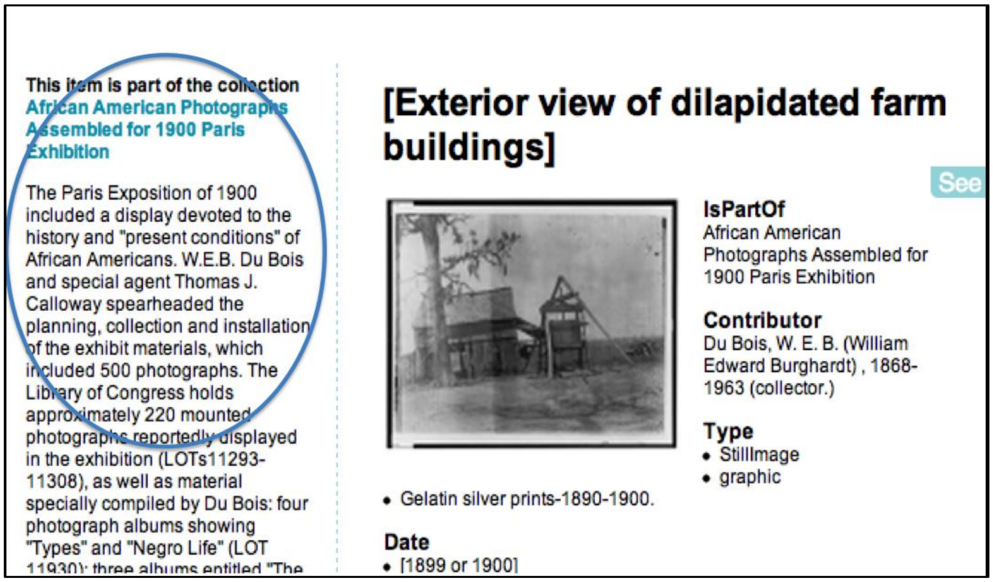
Date
- [1899 or 1900]
- [1899 or 1900]
- [1899 or 1900]

Format
- 1 photographic print : gelatin silver.
- graphic

Description
- In album (disbound): Negro life in Georgia, U.S.A., compiled and prepared by W.E.B. Du Bois, v. 3, no. 259.
- No known restrictions on publication.
- Original albums; Restricted access; Served by appointment only.
- B&w copy prints for LOT 11930 are provided as surrogates of

FIGURE 3: PHOTO WITH ITEM-LEVEL METADATA

Collection information, shown alongside the same photograph and record in Figure 4, reveals that the photograph is part of a cohesive exhibit, constructed for the Paris Exhibition of 1900 to depict the "history and present conditions of African Americans". Collection-level contextual information imbues the image with new significance. Information about how an item has been curated--- including why and by whom it was gathered into a collection---is a valuable function of collection description. Collection description also serves to augment information in an item record. In this same example, the item record suggests, by a parenthetical statement in the Contributor field, that W. E. B. DuBois' contribution to the item was as a collector (or curator), rather than photographer: "Du Bois, W. E. B….(collector.)". The collection record makes DuBois' contribution as a collector explicit. This kind of information is essential for situating a resource in context and fully understanding the sometimes limited or obscured information in item records.



This item is part of the collection
African American Photographs Assembled for 1900 Paris Exhibition

The Paris Exposition of 1900 included a display devoted to the history and "present conditions" of African Americans. W.E.B. Du Bois and special agent Thomas J. Calloway spearheaded the planning, collection and installation of the exhibit materials, which included 500 photographs. The Library of Congress holds approximately 220 mounted photographs reportedly displayed in the exhibition (LOTs11293-11308), as well as material specially compiled by Du Bois: four photograph albums showing "Types" and "Negro Life" (LOT 11930); three albums entitled "The

[Exterior view of dilapidated farm buildings]

IsPartOf
African American Photographs Assembled for 1900 Paris Exhibition

Contributor
Du Bois, W. E. B. (William Edward Burghardt) , 1868-1963 (collector.)

Type
- StillImage
- graphic
- Gelatin silver prints-1890-1900.

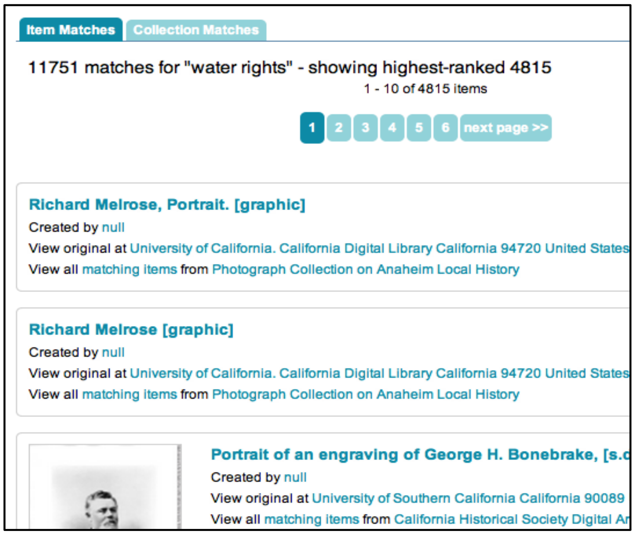
Date
- [1899 or 1900]

FIGURE 4: CONTEXT GIVEN BY COLLECTION DESCRIPTION

We can also observe the value of representing this collection-level information from a retrieval perspective. Added to indices used for search and retrieval, the text of collection descriptions increases the search system's recall of records relevant to a query. Using the photo in Figure 3 as an example, an item-level search for keywords "African Americans" would fail to return this artifact, and many of the others in this collection, because those terms do not appear in the item record. Incorporating collection-level description into the search index effectively expands the number of relevant terms to be matched against a query. In other cases, collection-level information could help narrow a search to increase precision by supplying terms to further refine queries.

5.3 MANAGEMENT AND PRESENTATION OF SEARCH RESULTS

A user interested in water rights in the American West, searching IMLS DCC for the phrase "water rights", will find nearly 5,000 item-level results (see Figure 5). Most of the results are highly specific in topic, such as biographies of historical figures with no obvious connection to the history of water rights. Few of the results, presented with snippets of the records, explicitly relate the item to the search for "water rights". Given results like these, how would the user begin exploring the available resources? This quandary stems from an inherent limitation of item records (absent collection records), rather than a limitation of the retrieval mechanism or interface design. Item records, by design, are highly specific in description; therefore it may be difficult to locate their relevance to, and position within, a broad, historical context, or even a long list of decontextualized search results.



11751 matches for "water rights" - showing highest-ranked 4815

1 - 10 of 4815 items

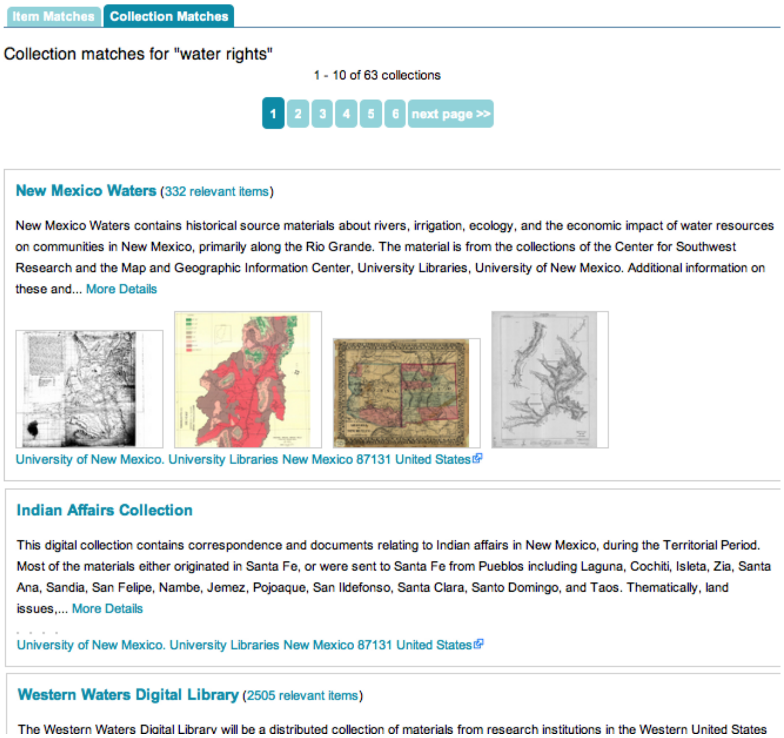
Richard Melrose, Portrait. [graphic]
Created by null
View original at University of California. California Digital Library California 94720 United States
View all matching items from Photograph Collection on Anaheim Local History

Richard Melrose [graphic]
Created by null
View original at University of California. California Digital Library California 94720 United States
View all matching items from Photograph Collection on Anaheim Local History

Portrait of an engraving of George H. Bonebrake, [s.c
Created by null
View original at University of Southern California California 90089
View all matching items from California Historical Society Digital Ar

FIGURE 5: ITEM LEVEL SEARCH RESULTS

Collection results, shown in Figure 6, augment the item results to provide a more intuitive view of the landscape of available resources in the aggregation and how they are organized into collections. Users can choose to filter an item-level search by collections devoted to specific thematic aspects of the "water rights" topic, such as water rights related to North American Indians and territorial struggles or collections specific to state or region.



Collection matches for "water rights"

1 - 10 of 63 collections

New Mexico Waters (332 relevant items)

New Mexico Waters contains historical source materials about rivers, irrigation, ecology, and the economic impact of water resources on communities in New Mexico, primarily along the Rio Grande. The material is from the collections of the Center for Southwest Research and the Map and Geographic Information Center, University Libraries, University of New Mexico. Additional information on these and... More Details

University of New Mexico. University Libraries New Mexico 87131 United States

Indian Affairs Collection

This digital collection contains correspondence and documents relating to Indian affairs in New Mexico, during the Territorial Period. Most of the materials either originated in Santa Fe, or were sent to Santa Fe from Pueblos including Laguna, Cochiti, Isleta, Zia, Santa Ana, Sandia, San Felipe, Nambe, Jemez, Pojoaque, San Ildefonso, Santa Clara, Santo Domingo, and Taos. Thematically, land issues,... More Details

University of New Mexico. University Libraries New Mexico 87131 United States

Western Waters Digital Library (2505 relevant items)

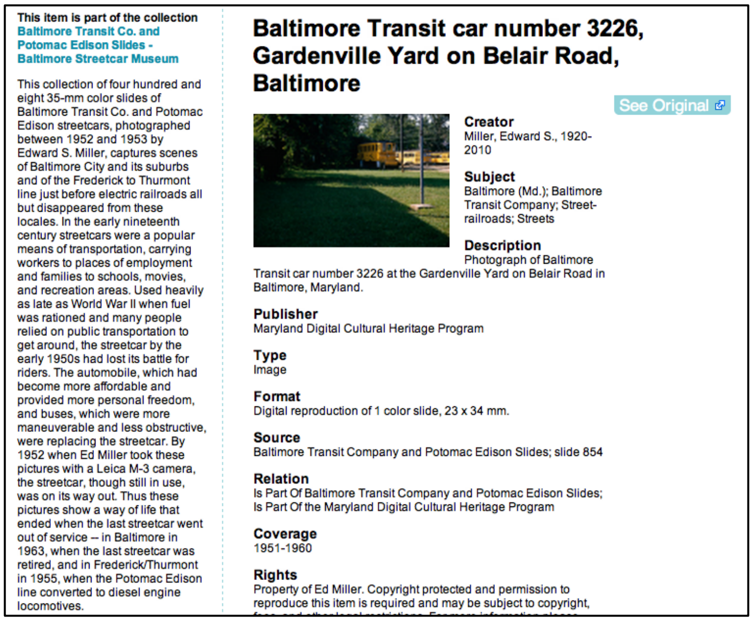
The Western Waters Digital Library will be a distributed collection of materials from research institutions in the Western United States

FIGURE 6: COLLECTION LEVEL SEARCH RESULTS

This use of collection descriptions, to supplement highly specific search results, is particularly important for systems that perform retrieval on the full text of items but do not display that text along with the results, whether due to intellectual property issues or technical constraints. Supplying elements of the collection description alongside item records in the course of search and retrieval can help orient a user and assist in moving more effectively through available search results.

5.4 ASSESSING RELEVANCE AND ACCESSIBILITY

IMLS DCC has a colorful collection of slides from the Baltimore Streetcar Museum. In Figure 7, collection-level contextual information augments a photograph of a streetcar in Baltimore. The collection description suggests that the collection as a whole may function as a coherent local history or educational resource: "these pictures show a way of life that ended when the last streetcar went out of service".



This item is part of the collection
Baltimore Transit Co. and Potomac Edison Slides - Baltimore Streetcar Museum

This collection of four hundred and eight 35-mm color slides of Baltimore Transit Co. and Potomac Edison streetcars, photographed between 1952 and 1953 by Edward S. Miller, captures scenes of Baltimore City and its suburbs and of the Frederick to Thurmont line just before electric railroads all but disappeared from these locales. In the early nineteenth century streetcars were a popular means of transportation, carrying workers to places of employment and families to schools, movies, and recreation areas. Used heavily as late as World War II when fuel was rationed and many people relied on public transportation to get around, the streetcar by the early 1950s had lost its battle for riders. The automobile, which had become more affordable and provided more personal freedom, and buses, which were more maneuverable and less obstructive, were replacing the streetcar. By 1952 when Ed Miller took these pictures with a Leica M-3 camera, the streetcar, though still in use, was on its way out. Thus these pictures show a way of life that ended when the last streetcar went out of service -- in Baltimore in 1963, when the last streetcar was retired, and in Frederick/Thurmont in 1955, when the Potomac Edison line converted to diesel engine locomotives.

Baltimore Transit car number 3226, Gardenville Yard on Belair Road, Baltimore

Creator
Miller, Edward S., 1920-2010

Subject
Baltimore (Md.); Baltimore Transit Company; Street-railroads; Streets

Description
Photograph of Baltimore Transit car number 3226 at the Gardenville Yard on Belair Road in Baltimore, Maryland.

Publisher
Maryland Digital Cultural Heritage Program

Type
Image

Format
Digital reproduction of 1 color slide, 23 x 34 mm.

Source
Baltimore Transit Company and Potomac Edison Slides; slide 854

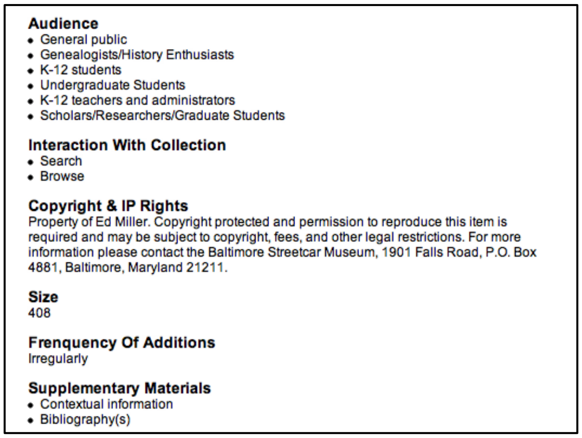
Relation
Is Part Of Baltimore Transit Company and Potomac Edison Slides;
Is Part Of the Maryland Digital Cultural Heritage Program

Coverage
1951-1960

Rights
Property of Ed Miller. Copyright protected and permission to reproduce this item is required and may be subject to copyright,

FIGURE 7: COLLECTION DESCRIPTION DISPLAYED WITH ITEM DETAILS

Certain elements in the collection record, beyond the description field, give more comprehensive context – not only about the provenance of items in the collection but also about the availability of items for different kinds of use. Certain aspects of item context, if shared across all items, are sometimes abstracted from item records into collection records to reduce redundancy.



Audience
- General public
- Genealogists/History Enthusiasts
- K-12 students
- Undergraduate Students
- K-12 teachers and administrators
- Scholars/Researchers/Graduate Students

Interaction With Collection
- Search
- Browse

Copyright & IP Rights
Property of Ed Miller. Copyright protected and permission to reproduce this item is required and may be subject to copyright, fees, and other legal restrictions. For more information please contact the Baltimore Streetcar Museum, 1901 Falls Road, P.O. Box 4881, Baltimore, Maryland 21211.

Size
408

Frenquency Of Additions
Irregularly

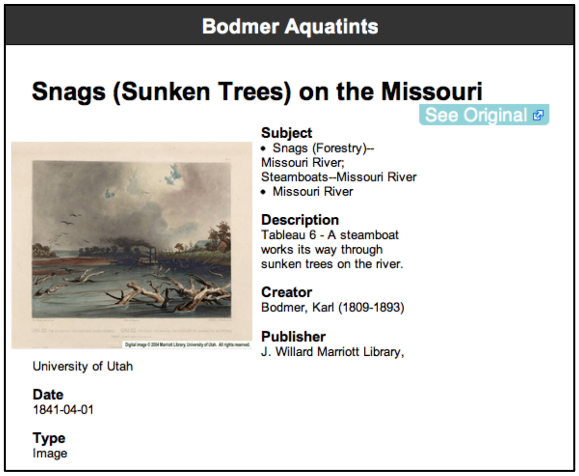
Supplementary Materials
- Contextual information
- Bibliography(s)

FIGURE 8: EXCERPT OF COLLECTION RECORD

The detailed collection record, an excerpt of which is shown in Figure 8, offers high-level guidance to users. Pragmatic properties of the collection record supply information about the rights for the collection as a whole, on options for interacting with the collection (such as searching or browsing), on potential audiences, on collection size and completeness, and on available supplementary materials. These properties allow the potential usefulness of any given resource to be fully exploited.

5.5 CONTEXT AND NAVIGATION

Bodmer Aquatints is a collection of watercolors, hosted at the University of Utah and accessible through the IMLS DCC aggregation.



Bodmer Aquatints

Snags (Sunken Trees) on the Missouri

Subject
- Snags (Forestry)--Missouri River; Steamboats--Missouri River
- Missouri River

Description
Tableau 6 - A steamboat works its way through sunken trees on the river.

Creator
Bodmer, Karl (1809-1893)

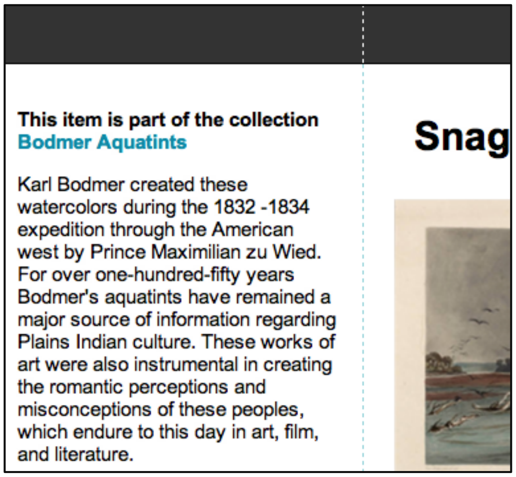
Publisher
J. Willard Marriott Library, University of Utah

Date
1841-04-01

Type
Image

FIGURE 9: ITEM VIEW OF AN IMAGE

Figure 9 shows the item-level information for one item from this collection. The item as it appears in this view could be imagined to satisfy various user interests, such as the history of river transportation, but information about the context of the item is not represented explicitly. The *dc:date* information suggests that the work has something to do with America in the age of western exploration ("1841-04-01"), but only a knowledgeable user could be expected to recognize the name of the creator and thereby infer the background of the piece. Adding the collection description to this view casts the item in a new light.



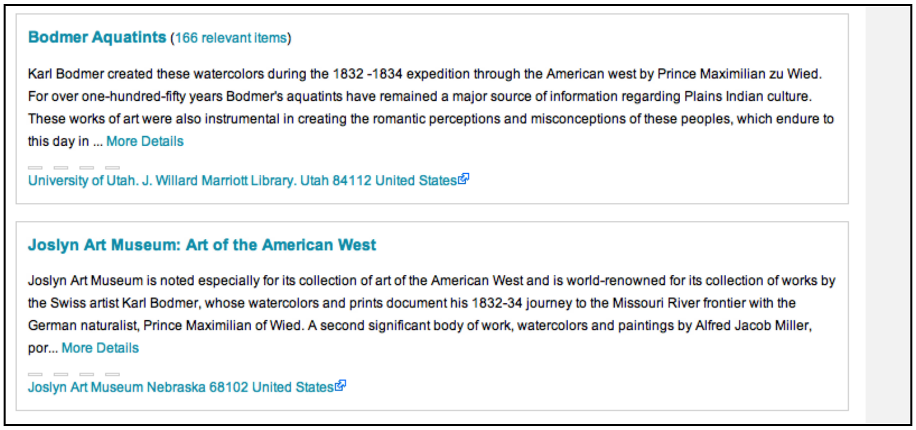
This item is part of the collection
Bodmer Aquatints

Karl Bodmer created these watercolors during the 1832 -1834 expedition through the American west by Prince Maximilian zu Wied. For over one-hundred-fifty years Bodmer's aquatints have remained a major source of information regarding Plains Indian culture. These works of art were also instrumental in creating the romantic perceptions and misconceptions of these peoples, which endure to this day in art, film, and literature.

FIGURE 10: COLLECTION INFORMATION PROVIDES CONTEXT FOR ITEMS

Figure 10 displays the collection description in the sidebar of the item view, drawn from the collection record associated with and linked to this item: "Karl Bodmer created these watercolors during the 1832-1834 expedition through the American west by Prince Maximilian zu Wied. For over one-hundred-fifty years Bodmer's aquatints have remained a major source of information regarding Plains

Indian culture.” The collection information shows that the item has historical importance in influencing popular conceptions about American Indian life, potentially increasing the audience for the river image.



Bodmer Aquatints (166 relevant items)

Karl Bodmer created these watercolors during the 1832 -1834 expedition through the American west by Prince Maximilian zu Wied. For over one-hundred-fifty years Bodmer's aquatints have remained a major source of information regarding Plains Indian culture. These works of art were also instrumental in creating the romantic perceptions and misconceptions of these peoples, which endure to this day in ... More Details

University of Utah. J. Willard Marriott Library. Utah 84112 United States

Joslyn Art Museum: Art of the American West

Joslyn Art Museum is noted especially for its collection of art of the American West and is world-renowned for its collection of works by the Swiss artist Karl Bodmer, whose watercolors and prints document his 1832-34 journey to the Missouri River frontier with the German naturalist, Prince Maximilian of Wied. A second significant body of work, watercolors and paintings by Alfred Jacob Miller, por... More Details

Joslyn Art Museum Nebraska 68102 United States

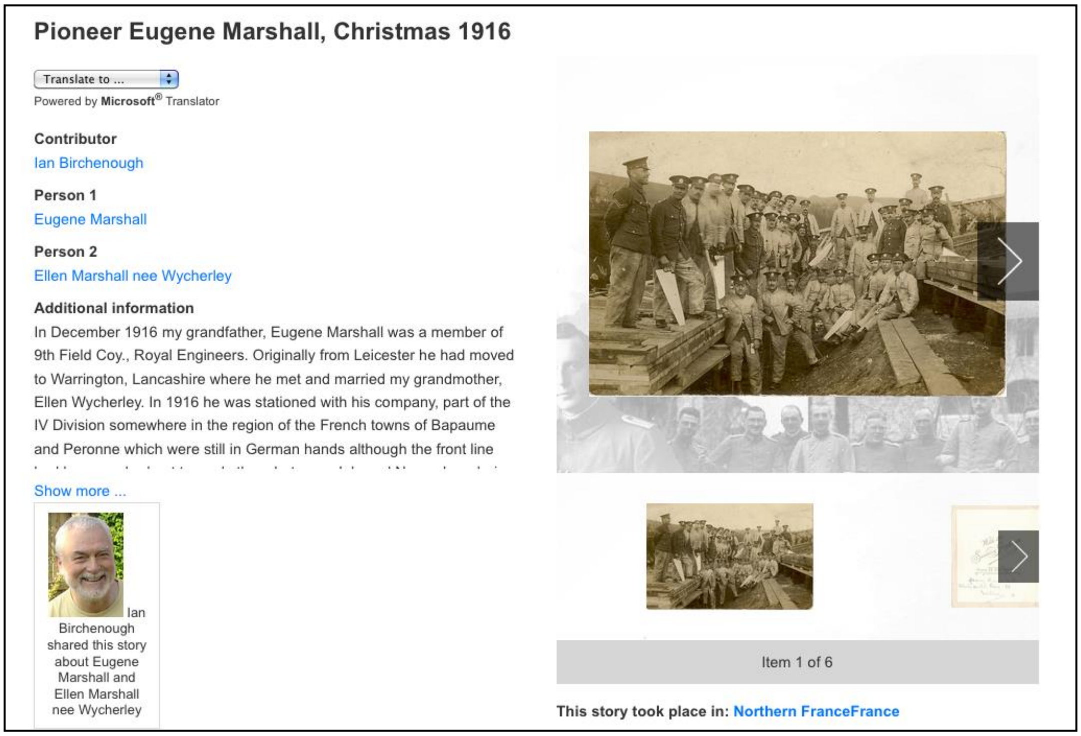
FIGURE 11: COLLECTION LEVEL SEARCH RESULTS SUPPORT NAVIGATION BETWEEN RELATED COLLECTIONS

A collection itself, rather than any particular item, may satisfy a researcher's information need. To this end, collection-level description is obviously critical. This is true in the Bodmer example. A collection-level view of search results shows two related collections: the Bodmer Aquatints at the University of Utah and the Joslyn Art Museum Collection, which also holds Bodmer aquatints (see Figure 11). By comparing collection-level records, a researcher interested in these watercolors can choose which, for example, offers higher quality digital representations or more authoritative curation. A collection's relevance to any given information need may be dependent on contextual information provided by the collection description, such as the collection's accessibility, provenance, and authority.

5.6 CONTRIBUTION OF COLLECTIONS BY USERS

Europeana 1914-1918[11] allows users to contribute personal collections of their own World War I memorabilia and the stories that surround them. As part of Europeana’s “user-generated content” program, users contribute “stories,” which comprise digital representations of objects from the personal collection (digitized by the project), a metadata record for each artifact, and a narrative description. Compared to institutionally generated (non-personal) collections, the 1914-1918 contributions tend to be very small (between one and ten objects). And unlike other collections we have explored, stories or artifacts may be sequentially ordered. Figure 12 displays one story from 1916, told in a sequence of six digitized photographs with accompanying metadata. The metadata documents the story contributor and people referenced by the story, along with the narrative text. Stories are searchable by full text and browsable by a controlled subject vocabulary, with subjects such as “Western front” and “Aerial warfare”.

[11] http://www.europeana1914-1918.eu/



Pioneer Eugene Marshall, Christmas 1916

Contributor
Ian Birchenough

Person 1
Eugene Marshall

Person 2
Ellen Marshall nee Wycherley

Additional information
In December 1916 my grandfather, Eugene Marshall was a member of 9th Field Coy., Royal Engineers. Originally from Leicester he had moved to Warrington, Lancashire where he met and married my grandmother, Ellen Wycherley. In 1916 he was stationed with his company, part of the IV Division somewhere in the region of the French towns of Bapaume and Peronne which were still in German hands although the front line

Ian Birchenough shared this story about Eugene Marshall and Ellen Marshall nee Wycherley

This story took place in: Northern FranceFrance

FIGURE 12: A USER-CONTRIBUTED COLLECTION

Allowing users to contribute individualized collections from their own content or to build collections from existing repositories of resources is an exciting area for development for digital aggregations. Although user-created collections within an aggregation will require specific kinds of technological support, these collections are likely to have the same basic requirements for representation and modeling as large collections supplied by institutional data providers.

6. REPRESENTATIONAL REQUIREMENTS TO SUPPORT ROLES OF COLLECTIONS AND COLLECTION DESCRIPTION IN AGGREGATIONS

The following requirements for modeling collections arise from the roles of collections as illustrated in Section 5:

1. Models must treat collections as individual resources within the aggregation and allow for the representation of properties of the collection.
2. Models must be prepared to represent collection membership as a property that stands between resources. When item-level representation is available, items should be explicitly linked to collection-level entities.
3. It is necessary to have a set of properties (i.e. a schema) designed to describe collections in ways that support users and administrators. Properties essential to the use of collections in digital libraries and aggregations include:
    a. Institutions or individuals that have participated in the stewardship of resources, including those responsible for: holding physical resources, gathering items together into collections, hosting digital versions of resources, and creating descriptions of resources. For that purpose, we explicitly distinguish the following roles :
        i. creation of original items
        ii. collection of the items (following some policy)
        iii. preservation of the items within a collection
        iv. creation of a digital representations
        v. provision of access
    b. Properties that can be used to reflect the contextual information implied by collection membership, including topical or subject properties, properties related to the purposes a collection was created to serve, and properties about the intended audience for a collection.
4. To the extent possible, property values in metadata should be identifiers of other resources that the system can make actionable.

6.1 REPRESENTING COLLECTIONS AS RESOURCES

For aggregations to meet information needs of users and administrators, it is necessary to have models that treat collections as first-class entities.

Above we have discussed some of the ways users benefit from collections when searching and interacting with digital aggregations. In addition, the direct representation of collections as individual objects offers important functionality for curators and administrators of aggregations. It provides a baseline unit for effective recording of usage statistics for analyzing demand and impact of specific types of content. Collection-level analyses are meaningful for conveying to stakeholders how end user communities value an institution's digital content and where future investments might best be directed.

In addition, for large aggregations of digital content, such as Europeana or state and regional level digital libraries in the U.S., retaining collections is critical to preserving the identity and branding of individual institutions as their materials become part of very large aggregations and readily accessed, linked to, and potentially remixed on the Web. In particular, collection representation allows institutions with smaller collections to better showcase their content and demonstrate the reach and impact of their digitization and curatorial efforts.

The requirement to represent collections as individual objects is related to the requirements to represent properties of collections and to represent collection membership. The collection object provides an entity to which collection-level properties can be attached. Similarly, collection objects allow the representation of collection membership by supplying an entity that can stand in a membership relationship with the resources that are the members of a collection.

6.2 REPRESENTING COLLECTION MEMBERSHIP

Item-level entities must be explicitly linked to collection-level entities to support a number of the potential roles of collections in large-scale aggregations. This requirement may be satisfied through a number of technical approaches; what is essential in terms of modeling is that the collection membership relationship be represented in a way that is available for inference and for indexing. This kind of linking will not be possible in aggregation or exchange scenarios where items are not given individual representation, but where items are available, the explicit representation of the membership relationship is critical to support inference between the collection and item levels and for the roles of collections illustrated in Section 5.

The representation of collection membership in aggregations assists in management and presentation of search results. Collection-level entities can be used to narrow search scope, increase result set precision, and improve the users' ability to discover and gain access to individual items. It can bring items in small collections to the surface that might otherwise be buried or inaccessible if discoverable only through item level access. Small collections can be included in a balanced presentation of the available resources through sampling rather than flooding a results list with items from a few overpoweringly large collections. In this way collection representation can act as an equalizing force in search. Also, as seen in the example of searching for "water rights" in a large aggregation presented in the Section 5, viewing the collection information for member items retrieved in a search helps users understand and quickly navigate between resources when relevance might otherwise be difficult to assess.

Browsing remains a key search and retrieval strategy for end users. Collection-level entities support browsing in large-scale aggregations by providing an additional organizational layer. Foulonneau et al. (2005) found that presenting collection-level entities as distinct search results facilitated the browsing behavior observed in humanities scholars through directly linking collection-level descriptions with the descriptions of their member item-level objects. This kind of display is only a possibility when the collection membership relationship is explicitly represented and available for integration into the display of search results.

For navigation between resources, historians need to move easily between items and collections. This finding has been further corroborated by usability studies conducted by the developers of American Social History Online aggregation of digital collections (as reported by Zavalina, 2010). Study participants, consisting of history faculty members and doctoral students remarked that they were most interested in being able to navigate from item-level search results to more hierarchically structured and contextually comprehensive collection-level metadata which better supports browsing tasks. This movement from the individual items to the collection level requires direct representation of collection membership.

The degree to which one can know a collection's unity criterion that "explains" membership to that collection may vary considerably. In many cases observation of a collection leads to easy determination of its unity criterion. But the accounts of collections reviewed in Section 3 are inconclusive regarding whether an item can become a member of a collection simply by adding its name (or URI) to a list, even if the person adding the item has no particular knowledge of that item.

Moreover, it is unclear whether collection membership is based on the fact that someone has collected the items following some conscious criteria and has managerial control over them, such as ownership, copyright, preservation, etc. As such, explicitly modeling all unity criteria requires a syntactic structure to describe constraints, which though technically feasible, would be too complex an issue for the scope of this paper.

Many inferences that can be made on the basis of a membership link will depend on the understanding of what it means to be a collection (e.g., on the assumed relationships between the creator of the collection and the items in a particular collecting context). Hence, expressing all possible inferences remains out of our reach for this work.

6.3 REPRESENTING PROPERTIES OF COLLECTIONS

Many of the uses of collections in large-scale digital aggregations that we have described rely on collection-level description. Collections must be described according to a collection-level schema for users to find and identify them as information objects, or for the managers of aggregations to use them to represent the contributions of data providers. Whenever a resource is accessed by a user, the contextual information should be readily available and some elements of the context may be directly presented to the user, depending on the specific access function.

Properties that can be used to reflect the contextual information provided by collection membership are particularly useful in large-scale digital aggregations. These properties include topical or subject properties, properties related to the purposes a collection was created to serve (e.g. an exhibition on African-American life as in the DuBois example described in the previous section), and properties about the intended audience for a collection.

Context in general is the totality of the associations in which a resource participates either as a subject or an object. This entire set of relationships includes the link between a resource and the collection to which it belongs, and collections themselves can carry properties that characterize the context of individual objects that are members of the collection. This is a very broad view of contextual information, where any statement about a resource may contribute to the context. Nevertheless, some collection properties are more intuitively contextual (e.g., the collection’s accrual policy, the institution that hosts the collection’s digital objects and metadata records, the institution that stewards the physical collection from which the digital collection was derived, the existence and whereabouts of supplementary materials for the collection, etc.). They do not directly describe the resource itself but give additional or supplemental information about the resource, perhaps in terms of its origin or source.

6.4 REPRESENTING PROPERTY VALUES WITH IDENTIFIABLE RESOURCES

Our final requirement concerns the data model's ability to express collection-related data. Addressing the various cases gathered in Section 5 requires representing properties of collections using not only string values but also connections between collections and other type of entities, which become first-order resources with their own identity: a person, an organization, a place, a concept, and so on.

This is especially crucial for attaching several properties to these entities in an appropriate way. For example, to state that a concept has a label in English and another in French, which should be distinguished from the labels for other topics associated with a collection. Similarly, using identifiable entities for persons allows the association of a person to a date (e.g. the year of their birth) that is not the date at which the collection was created.

Providing such resources with a clearly defined identity is also a first step towards enabling the re-use of such resources (and their data) across different collection descriptions. A given concept should be described and assigned a shareable identifier. This allows different collections to seamlessly re-use the concept by just referring to this identifier. For these reasons, the use of identifiable resources for property values is a core design concept in the Europeana Data Model.

7. COLLECTION REPRESENTATION AND DESCRIPTION

In the preceding sections we have demonstrated the usefulness of collections and collection description, and outlined a set of general modeling requirements that support this usefulness. In this section we present corresponding modeling strategies for large-scale digital libraries and aggregations that seek to incorporate collections.

Our first design decision, prior to discussing which types of resources and attributes are needed to represent collections, is to adhere to the core Europeana Data Model (EDM). Specifically, EDM extensively relies on the RDF modeling principles of using identifiable entities and properties for representing information about resources. Answering the requirement expressed in Section 6.4, this choice supposes the provision of identifiers (especially, web identifiers) for any resource worthy of description, and the description of these resources as distinct entities, which precisely matches our general design requirements.

In the following, we discuss the classes and properties needed to represent collection data. Our approach is twofold: (a) build on progress made on collection representation in the IMLS DCC project (discussed in Section 2.1); (b) systematically align with the existing EDM classes and properties, or when such alignment is not possible, present new candidates as extension to the EDM. At the time of writing this report, EDM does not provide for expressing collections as resources with distinct properties and relationships. An EDM extension to this effect is desirable so that it can express data that meets the requirements presented in Section 6.

7.1 DEFINING THE CLASS OF COLLECTIONS

To support the potential functions of collection structures and collection description discussed above, collections must be treated as "first-class" entities within an aggregation. The problem addressed here relates to a specific case of the design principle committed to representing resources with unambiguous identifiers and describing the classes to which resources belong. From a digital library perspective, this modeling strategy means that collections are treated as individual resources within the repository; a collection will be assigned an identifier and can be given a description that reflects properties of the collection itself.

The first option for modeling collections in EDM is to represent collections as instances of the *Provided Cultural Heritage Object* (*edm:ProvidedCHO*) class. As defined, this class "comprises the Cultural Heritage objects that Europeana collects descriptions about." Generally, the instances of this class are the main focus of digitization and access efforts. In the Europeana context of operation, the collection would be embedded in an *ore:Aggregation* , which bundles the collection with any digital representations (see Section 2.2). This approach adopts the standard object modeling methodology within Europeana, and treats collections in the same manner as any composite object that appears in the Europeana object space.

The use of the *ProvidedCHO* class for representing collections will allow collections to be incorporated into an aggregation with the existing techniques and mechanisms that repositories using EDM currently use. This will allow clear distinctions between collections and digital "representations" of them (e.g. web pages or sites) and will support different levels of descriptions for a collection, if needed.

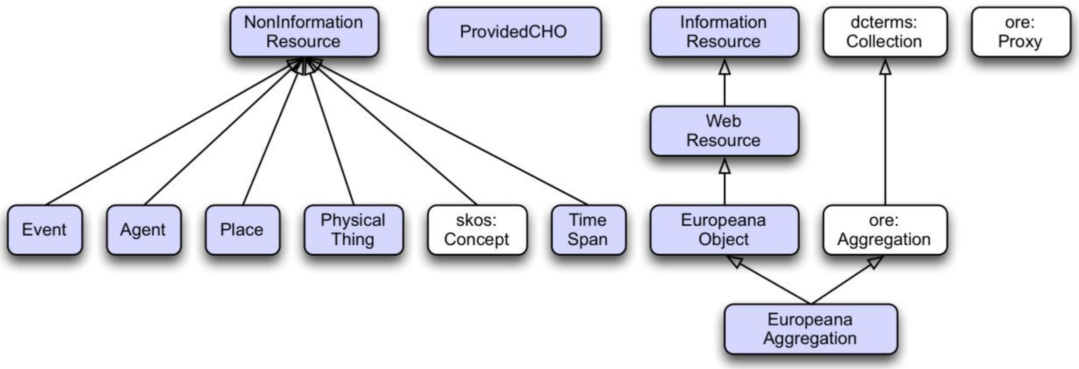


FIGURE 13: THE EDM CLASS HIERARCHY. CLASSES INTRODUCED BY EDM ARE SHOWN IN LIGHT BLUE RECTANGLES. CLASSES IN WHITE ARE RE-USED FROM OTHER SCHEMAS

However, *edm:ProvidedCHO* is a class of functional nature. That is, it enables the characterization of a resource from the perspective of its usage in a data aggregation and access service like Europeana, but does not provide meaningful information about the exact nature of the resource. In particular, a collection simply typed as Provided CHO would be difficult to distinguish from its item-level members, also typed as Provided CHOs.

One option for a class to specifically represent collections in digital aggregations is to adopt the view on collections developed by the Dublin Core Collection Description Task Group (DCMI, 2007) and use the *dcmitype:Collection* class. This class is an element of the DCMI Type Vocabulary provided by the Dublin Core Metadata Terms[12], and it is defined as "an aggregation of resources."

Such a vague definition will, of course, accommodate collections as we have conceptualized and discussed them. However, it does not seem optimal to directly re-use *dcmitype:Collection* in our context. As seen in Fig. 12, *ore:Aggregation* is a subclass of *dcmitype:Collection*. This means that in a system that use subclass reasoning, a query for resources of type *dcmitype:Collection* would also bring all resources of type *ore:Aggregation*, which is not ideal given the intensive use of that class in EDM for entities that are not collections, as noted in Section 2.2. In fact, the very general definition of *dcmitype:collection* includes any given set of resources, a scope that is considerably broader than the one of intentionally created or curated collections. It seems thus preferable to use a subclass of *dcmitype:Collection* that carries more specific semantics.

The next question is whether such a sub-class should be *ore:Aggregation* itself, or a sub-class of it. A first objection relates to the conceptual complexity of such a construction in the EDM context. If, as discussed earlier, we treat collections as EDM ProvidedCHOs, collections would be aggregations embedded in other aggregations (as other Provided CHOs submitted to Europeana). While not trivial, such a use of *ore:Aggregation* seems possible. In particular, at first sight it does not conflict with the basic definition of Aggregations as given in the ORE specifications: a member of this class is a "set of other Resources."[13]

However, this definition has to be interpreted in the context defined by the entire ORE documentation and requirements. ORE is designed for the packaging and interchange of web information objects, and is not focused on providing for the direct representation or description of a "real-world" resource. As another introduction to *ore:Aggregation* describes the class: "In order to be able to unambiguously

[12] http://purl.org/dc/dcmitype/Collection
[13] http://www.openarchives.org/ore/1.0/datamodel#Aggregation

refer to an aggregation of Web resources, a new Resource is introduced that stands for a set or collection of other Resources."[14]

ORE Aggregations are thus very technical constructs. The expectation seems to be that someone creates a resource of type *ore:Aggregation* with the specific purpose of creating bundles of web resources already in mind. It is quite different from other classes meant to describe entities that have their own existence outside the Web, such as the class *edm:Agent* – a person existing independently from a web data representation scenario. We are looking to represent collections in the context of information organization with a focus on supporting scholarship and general discovery purposes. That is, these collections are created with a meaningful (curatorial or scholarly) purpose in mind, not simply groups of objects brought together for administrative or technical purposes.

For these reasons, we recommend the introduction of a new class specifically for collections as they function to support the processes and needs of the users and the creators of collections. Note that we do not argue here for creating a sub-class of *dcmitype:Collection* that is disjoint with *ore:Aggregation*. Some collections (e.g., Flickr groups and galleries) are "born-digital" and clearly owe their existence to the Web environment. They may thus qualify as ORE Aggregations. We leave the door open to having a collection typed as *ore:Aggregation*, when it serves an application's purpose.

7.2 THE COLLECTION MEMBERSHIP RELATIONSHIP

The requirements analysis presented in Section 6 has shown that for collections to play their expected role in digital library aggregation and exchange environments, it is necessary to represent collection membership as a property that stands between resources. This property can then be used to explicitly link item-level entities to the collection-level entities of which they are members.

The DCMI Metadata Terms[15] defines *dcterms:hasPart* as "a related resource that is included either physically or logically in the described resource", and *dcterms:isPartOf* as "a related resource in which the described resource is physically or logically included." These terms are appropriate for representing collection membership, since it is easy to see that an item is logically included in a collection that it has been gathered into. In addition, as argued in Wickett, Renear and Furner (2011), the collection membership relationship aligns closely with a proper parthood relationship, if it is assumed that collection membership is a transitive relation.

However, it may be that these parthood relations are too general for the representation of collection membership in digital library aggregation and exchange environments. There are many kinds of parthood relations that may be represented with *dcterms:hasPart*. For example, pages are parts of books, and volumes are parts of series, and these seem like semantically distinct relationships from collection membership. It is perhaps most accurate to characterize collection membership as a particular kind of parthood.

A strategy that maintains connection to the commonly used Dublin Core property while maintaining specialized semantics for collection membership is to define a new property, *isGatheredInto* specifically for collection membership as a sub-property of *dcterms:isPartOf.* The sub-property relationship means that every instance of *isGatheredInto* implies a corresponding instance of *dcterms:isPartOf.* This connection from the specialized collection membership relation to the more general parthood relation provides better interoperability between different applications.

[14] http://www.openarchives.org/ore/1.0/primer#Nutshell

[15] http://dublincore.org/documents/dcmi-terms/

7.3 COLLECTION DESCRIPTION AND REQUIREMENTS

The IMLS DCC Collection Description Metadata Schema[16] is a data structure aligned with the Dublin Core Collections Application Profile[17] that is designed for representation of collections in a digital cultural heritage aggregation. As noted in Section 2.1 this schema was an early outcome of the DCC project (Shreeves and Cole, 2003), and it functions to maintain interoperability with item descriptions and to represent collection membership.

Here we analyze how this early schema fits the requirements discussed in Section 6, laying out the connections between the collection-level properties and the roles and requirements developed in previous sections, and briefly outlining potential inferences from item-level properties to collection-level properties.

Propagation rules between collection-level and item-level metadata properties can be used as a basis for the construction of collection records or for making inferences about items. These rules are logical conditionals that allow facts about collections to be inferred from descriptions of items, or vice versa (Wickett, 2012). In the following, the possibilities for constructing collection-level values from item-level ones, are noted for each property included in the analysis. Since the full development of propagation rules and associated techniques for inference is on-going, the notes in the table below are intended largely to suggest the kinds of inferences that can be supported.

The section also presents how the DCC collection schema compares with the existing EDM schema. The alignment with EDM is realized by (i) mapping the DCC AP fields onto the available properties used by EDM[18], and (ii) specifying the classes of resource the statements using these properties should be attached to. Namely, following the pattern presented in section 2.2, a record following the proposed schema is expected to result in:

- an *edm:ProvidedCHO* instance that represents the collection as an intellectual creation, independently from its digital realization(s), as discussed in 7.1.
- an *ore:Aggregation* instance that bundles the ProvidedCHO together with digital representations as created by a digitization and/or data aggregation process, thus representing the true “digital” context of a collection.

The following paragraphs present how the data represented in a DCC AP record would distribute over these two resources.

Organization and access for resources in digital aggregation and exchange environments frequently involves many different stewardship roles that are not always described explicitly. To avoid conflation between these roles, the property tables that follow also include an analysis of how properties that represent the institutions that have participated in stewardship map to the specific stewardship roles given in Section 6:

A. creation of original items
B. collection of the items (following some policy)
C. preservation of the items within a collection
D. creation of a digital representations
E. provision of access

[16] http://imlsdcc.grainger.uiuc.edu/CDschema_elements
[17] http://dublincore.org/groups/collections/collection-application-profile/
[18] http://pro.europeana.eu/edm-documentation

This analysis provides a clearer picture of how the institutions and individuals that participate in the stewardship of collections can be represented in metadata.

7.3.1 COLLECTION IDENTITY PROPERTIES

These properties are used to assign and manage properties of and relationships to individual collections. The collection identity properties align with the requirements above to treat collections as individual entities and explicitly represent collection membership, and to record information about the properties of collections.

Relevant representational requirements met: collections as individual objects, collection description

In EDM, the collection identity properties below are expected to be attached to a resource instance of edm:ProvidedCHO that represents the collection object itself, in terms that are independent from a specific digitization or access mediation.

DCC Schema Element	Proposed Definition	EDM Property	Relevant user requirements	Propagation notes
dc:title	Name of digital collection	dc:title	Collections meet information needs	Does not propagate
dcterms:alternative	Alternative name	dcterms:alternative	Collections meet information needs	Does not propagate
dc:identifier	Unique key for collection	dc:identifier	Collection repository management	Does not propagate

While this mapping is straightforward for *dc:title* and *dcterms:alternative*, the alignment of *dc:identifier* raises some issues. Namely, its values in DCC can sometime reflect preoccupations that are related to web representation purposes, e.g., using the URL of a digital collection on a website as its identifiers. We argue that such cases should be treated as *views* of a collection, not identifiers, and represent a practice that should be discouraged, even in the DCC context (e.g., the element *cld:isLocatedAt* in 7.3.2 would be a better choice for such values).

Identity properties do not propagate between the collection level and item level. Generally, the name or identifier of an individual item does not provide information about a collection that the item has been gathered into.

7.3.2 ACCESS PROPERTIES

These properties support access to collection-level entities. They function to support the roles of both collections and items in those collections, especially in scholarly research practices.

Relevant representational requirements met: collection description, hosting institution

DCC Schema Element	Proposed Definition	EDM Property	Relevant user requirements	Propagation notes
cld:isLocatedAt	Collection URL (home page)	edm:isShownAt	-Collections meet information needs -Organization and navigation -Situating or interpreting a resource in a	Does not propagate

			context	
imls:interactivity	Indication of how a user can interact with a collection		-Search and retrieval -Situating a resource in context	Values may propagate from item level when items have a common value. Source properties: dc:description
dcterms:accessRights	A statement of any access restrictions	edm:rights if the value is among the ones listed at http://pro.europeana.eu/available-rights-statements, otherwise keep dcterms:accessRights	-Organization and navigation -Search and retrieval -Collection repository management	Values may propagate from item level. Source properties: dcterms:accessRights dc:rights edm:rights

These access properties pertain to digital representations of collections intended for access and presentation via the web. Therefore, in the context of EDM, they should be attached to the instance of *ore:Aggregation* that specifies the digital collection context of a given collection (represented as an instance of *edm:ProvidedCHO*). For example, *edm:isShownAt* would be used to record the URL for a landing page for a collection.

The distributed nature of content hosted in many digital aggregation and exchange environments means that access properties frequently do not propagate between the collection and item levels. In particular, a property expressing network locations where digital collections can be found with a browser (e.g. *cld:isLocatedAt*) will not imply that items can be found in the same network location since items may not have any digital representation or may be hosted elsewhere. However, information about interactivity and rights may propagate from the item level to the collection level.

7.3.3 AGGREGATOR CONTEXT PROPERTIES

These properties record information essential to the operation of data creation and aggregation systems, such as information about a funded project that conducted the digitization of a collection of materials or the institution that makes a digital representation of a collection available on the web.

Since these properties are specific to aggregation operations, they can be attached to the *ore:Aggregation* resource in EDM.

Relevant representational requirements met: hosting institution, source data

The DCC AP features the following properties:

DCC Schema Element	**DCC Schema Definition**	**Relevant user requirements**	**Propagation notes**
imls:project	The project(s) which created the digital version of the collection. (Links to the Project Entity)	-Administrative -Situating a resource in context - Stewardship Role D: actor that converted the collection into a digital resource without intellectual	Does not propagate

		contribution to its composition.	
cld:isAccessedVia	Services beyond the URL that provide access to the collection data, such as an OAI data provider or Z39.50 target	-Organization and navigation -Search and retrieval	May propagate to item-level.
dc:publisher	The institution that makes the collection available on the web. Links to an Institution Entity.	-Administrative -Situating a resource in context - Stewardship Role E: actor who makes the content available on the web)	Does not propagate
dc:contributor	The institution(s) that has contributed content to the digital version of the collection. Links to an Institution Entity.	-Administrative -Situating a resource in context - Stewardship Role C: actor who preserves the original collection.	Does not propagate
imls:managedBy	The person who has responsibility for the collection. Link to the Administrator Entity. (contact point on provider side)	-Administrative -Situating a resource in context Some individual in the institution that has Role C.	Does not propagate

An important question is whether these elements can be aligned to the EDM properties representing the same sort of context information for Europeana objects in general: *edm:provider* and *edm:dataProvider*.

EDM Property	**EDM Definition**	**Propagation notes**	**Example and relevant stewardship role**
edm:provider	The name or identifier of the organisation who delivers data directly to an aggregation service (e.g. Europeana)	May propagate to item level. Propagation will be between instances of ore:aggregation. Target properties: edm:provider	The Linked Heritage project is an edm:provider for digital objects from the The Arts and Theatre Institute in Prague Stewardship Role E: actor who makes content available on the web.
edm:dataProvider	The name or identifier of the organisation who contributes data indirectly to an aggregation service.	May propagate to item level. Target properties: edm:dataProvider	The Arts and Theatre Institute in Prague is an edm:dataProvider via the Linked Heritage project Stewardship Role C, actor who preserves the original collection.

The analysis of stewardship roles shows that the actor that is responsible for the care and preservation of the items within a collection (Role C) is clearly represented with *dc:contributor* or *edm:dataProvider.* Similarly the actor that provides digital representations on the web (Role E) is clearly represented with *dc:publisher* or *edm:Provider.* In contrast, the actor who creates a digital version of a collection (Role D) is not clearly represented. Frequently, the creation of a digital representation of an item will be considered part of the preservation process, so the same institution that plays Role C will also be playing Role D (and it could sometimes play role E, as well). This information is not directly represented in the properties defined by either DCC or EDM, but we do not propose an extension since it is unlikely that a user will need specific information about the creation of a digital version for their information needs.

The two approaches to aggregation should continue to use their own strategies and properties, which are tailored to meet their specific requirements for representing, displaying, and exchanging administrative and contextual information.

7.3.4 COLLECTOR CONTEXT PROPERTIES

These properties reflect aspects of the creation of a collection via the gathering together of individual items. They represent the intent of a curator or scholar with respect to the collection, and facts about the collection process. Generally these properties align with the requirement to represent properties related to the purposes a collection was created to serve and the provenance of the collection itself, including institutions that have participated in the stewardship of resources, that hold physical resources and host digital versions of resources, and have created descriptions of resources.

Relevant representational requirements met: collection description

In EDM all collector context properties should be attached to the instance of *edm:ProvidedCHO* that represents the original collection.

DCC Schema Element	**Proposed Definition**	**EDM Property**	**Relevant user requirements**	**Propagation notes**
dc:creator	Entity that gathers items together following implicit or explicit criteria or accrual policy	dc:creator	-Creating personal collections -Situating / interpreting a resource in a context Stewardship Role B: The Actor applying the relevant collection criteria/ accrual policy.	Does not propagate
dc:type	The nature or genre of the resource (i.e., dcmitype:Collection)	dc:type	- Collection repository management	Does not propagate.
dc:rights	Information about rights held in and over the collection	edm:rights if the value is among the ones listed at http://pro.europeana.eu/available-rights-statements, otherwise keep dc:rights	-Organization and navigation -Collection repository management	Values may propagate from item level when items have consistent values. Source properties: dcterms:accessRights dc:rights edm:rights
dc:relation	Any other collection(s) associated with or that complements the current collection	dc:relation	-Organization and navigation -Situating a resource in context	Values may propagate from item level. Source properties: dc:relation edm:isDerivativeOf edm:isAnnotationOf

				edm:hasMet dcterms:isReferencedBy
dcterms:extent	The number of items within the collection	dcterms:extent		Value may be constructed from counting *isGatheredInto* arcs with resource as subject.
dcterms:provenance	A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation.	dcterms:provenance	-Collection repository management -Situating a resource in context	Values may propagate from item level. Source properties: dcterms:provenance dc:source
dcterms:accrualPolicy	A statement of the collection development policy for the collection		-Collection repository management -Collections meet information needs -Situating a resource in context	Does not propagate
dcterms:accrualPeriodicity	A statement of how often the collection is updated		-Collection repository management -Situating a resource in context	Does not propagate
imls:supplement	Additional materials included alongside the collection that explain, incorporate, or otherwise make use of the collection. For example, may be used for finding aids, or material that describes a collection.	edm:isRelatedTo dc:relation dcterms:isReferencedBy edm:isSimilarTo	-Situating a resource in context	Does not propagate
dcterms:audience	The primary audience(s) of the items		-Collections meet information needs -Situating a resource in context	Values may propagate from item level. Source properties: dcterms:audience dc:description
dcterms:abstract	A summary of the content and topics of the collection.	dc:description	-Search and retrieval	Values may propagate from item level, especially given universal or consistent item-level values. Source properties: dc:description

				dc:subject dc:coverage dcterms:spatial dcterms:temporal
[no corresponding IMLS DCC property]	A 'key object' from the collection be it a masterpiece, or a good exemplar. (This is a sub-property of the inverse of the *isGatheredInto* collection membership property)	edm:highlight	-Organization and navigation. Can especially be useful to provide representative objects for a collection page or snippet on portal.	Does not propagate

Note that we cannot suggest a trivial mapping for *imls:supplement*. The analysis of the definition and existing data reveals that it could match either *edm:isReferencedBy*, *edm:isSimilarTo*, *dc:relation* and *edm:isRelatedTo*. If a provider can assign the relevant properties on a case by case basis, they should select the most specific property that fits each of them (*edm:isReferencedBy* and *edm:isSimilarTo* are sub-properties of *dc:relation*, itself a sub-property of the very general *edm:isRelatedTo*).

7.3.5 SECONDARY COLLECTOR CONTEXT PROPERTIES

These properties are used to describe relationships between collections and to reflect the embedding or inclusion of one collection-level entity into another collection-level entity. In EDM all secondary collector context properties should be attached to the instance of *edm:ProvidedCHO* that represents the original collection.

Relevant representational requirements met: collection description: sub-collections, source data

DCC Schema Element	**Proposed Definition**	**EDM Property**	**Relevant user requirements**	**Propagation notes**
dcterms:isPartOf	Any other collection(s) that contain the current collection	dcterms:isPartOf	-Organization and navigation -Situating a resource in context	Propagation to sub-collections not recommended due to risk of conflation between direct inclusion and inclusion inherited through propagation.
dcterms:hasPart	Any other collection(s) contained within the current collection	dcterms:hasPart	-Organization and navigation -Situating a resource in context	Propagation to super-collections not recommended due to risk of conflation between direct inclusion and inclusion inherited through propagation.
dc:source	The physical collection(s) from which the current digital collection is derived	dc:relation	-Organization and navigation -Situating a resource in context	Value may propagate from item level, given physical source and membership. Source properties: dc:source edm:isRepresentationOf

7.3.6 ITEM-RELATED PROPERTIES

These properties indicate attributes of the particular items that have been gathered into a collection. These properties align with the requirements to describe collections in ways that support the users and administrators of digital library aggregations. In a fully implemented linked data scenario where both collections and items are fully described, the representation of these properties at the collection-level would not be necessary, as the information could be retrieved from the item level through the membership relation. However, in cases where item descriptions may not be directly available or may not be expressed as structured metadata, it is necessary to have properties available at the collection level to record information about items that is relevant to the discovery and use of a collection. Further, collection level properties may not strictly hold for all items, but nevertheless be a good approximation that is relevant for recall.

Relevant representational requirements met: collection description: sub-collections, source data

In EDM all of the item-related properties should be attached to the instance of *edm:ProvidedCHO* that represents the original collection.

DCC Schema Element	**Proposed Definition**	**EDM Property**	**Relevant user requirements**	**Propagation notes**
cld:itemType	Genre or nature of objects or resources in the collection		-Collection repository management -Collections meet information needs	Values may propagate from item level. Source properties: dc:type
cld:itemFormat	The format (media type, physical or digital) of the items in the original collection. This may be information about a physical object (for physical items) or a digital media type (for born-digital items). This property refers to the ProvidedCHO, not a web representation		-Collection repository management -Collections meet information needs	Values may propagate from item level. Source properties: dc:format edm:hasType
[no corresponding IMLS DCC property]	The creator of one or more items in the collection.	edm:itemCreator	-Collections meet information needs - Stewardship Role A: Creator of the items (such as "Monet") regardless if he/she/it is identical with the creator of the collection	Values may propagate from item level. Source properties: dc:creator dc:contributor edm:hasMet
dc:language	If text, the language(s) in the collection	dc:language	-Collections meet information needs	Values may propagate from item level. Source properties:

				dc:language
cld:dateItemsCreated	A range of dates over which the individual items within the collection were created		-Collections meet information needs -Situating / interpreting a resource in a context -Search and retrieval	Values from item level may be aggregated into a collection-level value. Source properties: dc:date dcterms:created
dc:subject	Terms that describe the overall topical content of the items in the digital collection.	dc:subject	-Collections meet information needs - Collection description	Item-level values may be used to identify a generalized subject value for a collection. Source properties: dc:subject dc:description
dcterms:spatial	A place(s) or area(s) associated with most or all of the items in the digital collection.	dcterms:spatial	-Search and retrieval	Item-level values may be aggregated or used to identify a generalized collection-level value, especially given universal or consistent presence. Source properties: dcterms:spatial
dcterms:temporal	A time period(s) associated with most or all of the items in the digital collection.	dcterms:temporal	-Search and retrieval	Item-level values may be aggregated or used to identify a generalized collection-level value, especially given universal or consistent presence. e.g. collection-level range may be constructed from item values. Source properties: dcterms:temporal

These item-related properties are a particularly promising target for techniques using propagation rules to construct collection-level property values from item-level metadata. For example, *cld:itemType* values for a collection could be constructed from item-level metadata that includes *dc:type* values with a system that takes each distinct type value identified for items within a collection and produces a statement for the collection-level property.

8. CONCLUSIONS

In this report we have argued for the representation and description of collections in large-scale digital cultural heritage aggregations and exchange environments. We have presented a discussion and given examples of the functional roles collections can play in aggregations in terms of representation of contributing institutions, improvement of user search experiences and the provision of contextual information about cultural heritage resources, as well as the value for administrative functions. The requirements for collection representation and description were developed to directly support these roles in aggregation and exchange environments. Finally, the modeling recommendations are designed to meet each of the requirements.

Additionally, the report contributes clarification of the fundamental nature of collections in the digital cultural heritage domain. Specifically, the introduction of *unity criteria* as the binding force that brings the members of a collection together into a single entity provides a greater understanding of the ontological nature of collections and may be used to support inferences about collections and collection description. Additionally, the distinction between physical holdings collections and referential collections separates the functions scholars and other users expect collections to play from traditional assumptions about ownership and stewardship of digital resources.

Some of the most exciting aspects of representing collections in digital aggregations arise from the potential to propagate data between collection descriptions and item descriptions. This kind of propagation could be used to populate collection descriptions or to validate data in an aggregation. This report only outlines possibilities in this area, and work on specific propagation rules and implementation techniques is ongoing.

The representation of collections in digital cultural heritage aggregations and exchange environments has the potential to serve a range of intellectual, administrative, and functional roles. Given the modeling strategies discussed here, we are now in a strong position to put collections and collection description to work in aggregations to realize the many benefits for users and developers of digital content and access systems.

REFERENCES

- Atkinson, R. (1998). Managing traditional materials in an online environment: Some definitions and distinctions for a future collection management. *Library Resources & Technical Services*, 42(1), 7-20.
- Bearman, D. (1992). A User Community Discovers IT Standards. *Journal of The American Society For Information Science*, 43(8), 576-578.
- Beaudoin, J. (2012). Context and Its Role in the Digital Preservation of Cultural Objects. *D-Lib Magazine*, (11/12).
- Bekiari, C., Charami, L., Doerr, M., Georgis, C. and Kritsotaki, A. (2008). Documenting Cultural Heritage in Small Museums. In *Annual Conference of the International Documentation Committee of the International Council of Museums (CIDOC) 2008*. Available at: http://www.cidoc-crm.org/docs/SmallMuseums_CIDOC2008.pdf
- Bekiari, C., Doerr, M. and Le Boeuf, P.(2008) FRBRoo, a Conceptual Model for Performing Arts. 2008 Annual Conference of CIDOC, Athens. Available at: http://www.ics.forth.gr/_publications/drfile.2008-06-42.pdf
- Brack, E., Palmer, D., and Robinson, B. (2000). Collection level description-the riding and agora experience. *D-lib Magazine*, 6(9).
- Brockman, W., Neumann, L., Palmer, C., and Tidline, T. (2001). *Scholarly work in the humanities and the evolving information environment.* Digital Library Federation.
- Buckland, M. (1995). What will collection developers do? *Information Technology And Libraries*, 14(3), 155-159.
- Casserly, M. (2002). Developing a concept of collection for the digital age. portal: *Libraries and the Academy*, 2(4), 577-587.
- CIDOC (2010). *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC.
- Covi, L. M. and Cragin, M. H. (2004). Reconfiguring control in library collection development: A conceptual framework for assessing the shift toward electronic collections. *Journal of the American Society for Information Science and Technology*, 55(4), 312-325.
- Currall, J., Moss, M., and Stuart, S. (2004). What is a collection? *Archivaria*, 58, 131-146.
- Dublin Core Collection Description Task Group (2007). Dublin core collections application profile. http://dublincore.org/groups/collections/collection-application-profile/
- Doerr, M. (2003). The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3): 75.
- Doerr, M., Fundulaki, I., and Christophidis, V. (1997). The specialist seeks expert views: Managing digital folders in the AQUARELLE project. *Museums and the Web: An International Conference, Los Angeles, CA, March 16 - 19, 1997.* Available at: http://www.museumsandtheweb.com/mw97/speak/doerr.html
- Duff, W. M., and Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behaviors of historians in archives. *Library Quarterly*, 72(4), 472-496.
- Foulonneau, M., Cole, T., Habing, T. G., and Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*., pages 32-41. ACM Press.
- Geisler, G., Giersch, S., McArthur, D., and McCelland, M. (2002). Creating virtual collections in digital libraries: benefits and implementation issues. *Joint Conference on Digital Libraries 2002, Portland, OR*, pp. 210-218.
- Gonçalves, M. A., Fox, E. A., Watson, L. T., and Kipp, N. A. (2004). Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22:270-312.
- Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, 47(3), 380-402.

- Gueguen, G. (2010). Digitized Special Collections and Multiple User Groups. *Journal of Archival Organization*, 8(2), 96-109. doi:10.1080/15332748.2010.513324
- Heaney, M. (2000). An analytic model of collections and their catalogues. *UK Office for Library and Information Science*.
- Heath, T. and Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. HTML version accessible at http://linkeddatabook.com/book
- Hill, L., Janee, G., Dolin, R., Frew, J., and Larsgaard, M. (1999). Collection metadata solutions for digital library applications. *Journal of the American Society for Information Science*, 50(13), 1169-1181.
- Kaczmarek, J. (2006). The complexities of digital resources: Collection boundaries and management responsibilities. *Journal of Archival Organization*, 4(1), 215-227.
- Lagoze, C. and Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib magazine*, 4(11).
- Lee, H. (2005). The concept of collection from the user's perspective. *The Library Quarterly*, 75(1), 67-85.
- Lee, H. (2000). What is a collection? *Journal of the American Society for Information Science*, 51(12), 1106-1113.
- Lourdi, I., Papatheodorou, C., and Doerr, M. (2009). Semantic integration of collection description. *D-Lib Magazine*, 15(7/8), 1082-9873.
- Low, J. T. and Doerr, M. (2010). A Postcard is Not a Building - Why we Need Museum Information Curators. In *Proc. of the CIDOC 2010 Conference.*
- Lynch, C. (2002). Digital collections, digital libraries and the digitization of cultural heritage information. *First Monday*, 7(5-6).
- Meghini, C. and Spyratos, N. (2007). Viewing collections as abstractions. *Digital Libraries: Research and Development*, pp. 207-217.
- Meghini, C. and Spyratos, N. (2010). Unifying the concept of collection in digital libraries. *Advances in Intelligent Information Systems*, pp. 197-224.
- Meghini, C., Spyratos, N., and Sugibuchi, T. (2010). Modelling digital libraries base on logic. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, pp. 2-13. Springer-Verlag.
- M'kadem, A. and Nieuwenhuysen, P. (2010). Digital access to cultural heritage material: case of the Moroccan manuscripts. *Collection Building*, 29(4), 137-141.
- Moss, M. and Currall, J. (2004). Digitisation: taking stock 1. *Journal of the Society of Archivists*, 25(2), 123-137.
- Palmer, C. L. (2004). Thematic research collections. In *A companion to digital humanities*. Blackwell, Oxford.
- Palmer, C. L., Knutson, E., Twidale, M., and Zavalina, O. (2006). Collection definition in federated digital resource development. In *Proceedings of the 69th ASIS&T Annual Meeting (Austin, TX)*.
- Powell, A., Heaney, M., and Dempsey, L. (2000). Rslp collection description. *D-lib Magazine*, 6(9), 1082-9873.
- Palmer, C. L., Zavalina O. L., and Fenlon K. (2010). Beyond size and search: building contextual mass in digital aggregations for scholarly use. Proceedings of the 73rd ASIS&T Annual Meeting, Pittsburgh, Pennsylvania.
- Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008). Collection/Item metadata relationships. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*.
- 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

- Shreeves, S. L. and Cole, T. W. (2003). Developing a collection registry for IMLS NLG digital collections. In *Proceedings of the 2003 international conference on Dublin Core and metadata applications: supporting communities of discourse and practice---metadata research & applications* (DCMI '03).
- Sukovic, S. (2008). Information Discovery in Ambiguous Zones of Research. *Library Trends*, 57(1), 72-87.
- Sweet, M. and Thomas, D. (2000). Archives described at collection level. *D-Lib Magazine*, 6(9).
- Wickett, K. (2012) Collection/Item Metadata Relationships. *Doctoral dissertation, University of Illinois at Urbana-Champaign.* Available at: http://hdl.handle.net/2142/42198
- Wickett, K. M., Renear, A. H., and Urban, R. J. (2010). Rule categories for collection/item metadata relationships. In *Proceedings of the 73rd ASIS&T Annual Meeting (Pittsburgh, PA)*.
- Zavalina, O. (2010) Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use. *Doctoral dissertation, University of Illinois at Urbana-Champaign*. Available at: http://hdl.handle.net/2142/16620
- Zavalina, O.L. (2011) Contextual metadata in digital aggregations: application of collection-level subject metadata and its role in user interactions and information retrieval, *Journal of Library Metadata*, 11(3/4), 104–128.