SUPPLEMENTING OAI-PMH IN THE IMLS DIGITAL COLLECTIONS & CONTENT AGGREGATION

BY

JACOB GUY JETT

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Certificate of Advanced Studies in Library and Information Science
with a concentration in Digital Libraries
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

C.A.S. Committee

Professor Tim Cole, Chair
Professor Carole Palmer
Assistant Professor Jerome McDonough

ABSTRACT

The rate of adoption of OAI-PMH among the IMLS DCC (Digital Collections & Content) data providers remains a modest 23%. As a result, large quantities of item-level metadata records cannot be harvested into the DCC aggregation's item-level metadata repository. This thesis explores alternate methods of harvesting item-level metadata, either through the use of website HTML parsing technologies to capture metadata directly from webpages and permanently store it as xml files or through the use of broadcast metasearch technologies to provide additional links to information resources within the DCC's search results page. The nature of "collections" is also explored and a classification system based on the nature of the "items" within each collection is constructed in order to both better understand the contents of the DCC aggregate and to facilitate the prediction of experiment outcomes. While labor intensive with regards to the need to construct metadata standard crosswalks and retool harvesting code, website HTML parsing is found to be a powerful tool for both increasing the rate of item-level metadata repository growth and enhancing the choices for both aggregate users and collection developers. While broadcast metasearch experiments were inconclusive several emerging applications of broadcast metasearch technology may be promising methods of supplementing the contents of the aggregate's item-level metadata repository.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In September of 2002, the Institute of Museum and Library Services (IMLS) awarded the University of Illinois at Urbana-Champaign (UIUC) a research and demonstration project grant, the IMLS Collection Registry and Metadata Repository. The overall goal of this project was "to design, implement, and research a collection-level registry and item-level metadata repository service that will aggregate information about digital collections and items of digital content created using funds from IMLS National Leadership Grants." A key initial focus was to explore emerging technologies for accessing and harvesting metadata from disparate and dispersed sources. The original project proposal identified 4 approaches to meet its goal:

- Create a registry of IMLS NLG funded digital collections that were funded between the dates of 1/1/1998 and 9/30/2005.
- Design and implement a searchable item-level metadata repository using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as the metadata aggregation method.
- Assist NLG recipients in setting up OAI-PMH compliant data provider services.
- Research the costs and potential benefits of participation in these service processes.

(Proposal for an IMLS Collection Registry and Metadata Repository, 2002)

The initial phase of the project, which resulted in the IMLS Digital Collections & Content (DCC) aggregate (registry, repository, and portal), was successful and the project was extended in 2005 so that further studies of collection identity, metadata issues (including granularity, normalization, transformation, and enrichment), audience specific portal design, workflow issues, and knowledge diffusion could be conducted. The registry and repository membership was also expanded to encompass those collections that had been funded by IMLS Library Services and Technology Act (LSTA) grants and selected collections of digitized cultural heritage materials. (Proposal to Extend IMLS Collection Registry and Metadata Repository Project, 2005)

1.1. PROBLEM OVERVIEW

Based on results from the first 3 years of the project, the initial project coordinator, Sarah Shreeves, published a white paper detailing the barriers to interoperability between digital collections. She noted that only 22% of collections within the registry had been successfully harvested. She identified three broad categories of issues that prevented the remaining 77% of collections from being harvested. These were:

- Insufficient technical infrastructure to make implementation of OAI applications practicable.
- Insufficient metadata or metadata that was too poorly formed to make sharing practicable.
- Socio-economic factors specific to each individual institution and project.

(Shreeves, 2005)

In December 2009 the contents of the aggregate were again surveyed and it was discovered that this percentage had remained essentially unchanged (increasing to a little less than 23%, see Chapter 3.1).

Since the adoption of OAI-PMH by institutions that participate in the DCC aggregate remains relatively low, only a limited number of item-level metadata records can be harvested from a small pool of collections listed in the collection registry. From a collection development point of view it is very desirable to increase the number of collections from which items can be harvested into the DCC item-level metadata repository. This CAS project seeks to answer the research question, what are the pros and cons of developing additional means to aggregate data provider content (i.e. item-level metadata records)? Specifically:

- What are the potential benefits of supplementing OAI-PMH data harvesting?
- What is the potential for using traditional aggregation technologies and methodologies, like website HTML parsing (screen scraping), to harvest item-level metadata from data providers?
- Are certain types of data providers better suited to having their item-level metadata harvested using website HTML parsing?
- What is the potential for using broadcast metasearch technologies and methodologies to capture and present item-level metadata from data providers and present it to the aggregate service's users?

- Are certain types of data providers better suited to have their item-level metadata harvested and presented to users via broadcast metasearch?

In order to assess the potential benefits of supplementing OAI-PMH, the contents of the DCC aggregate are first surveyed in detail in order to assess the coherency of the aggregate as a whole and the accuracy of the records within the both the collection registry and the item repository. This survey will also aid in determining if certain types of data providers are better suited to harvest by either website HTML parsing or broadcast metasearch.

Both website HTML parsing and broadcast metasearch are built on the same essential principles. Web-pages are retrieved and algorithmically parsed for specific information. The primary difference in the two technologies is that the data retrieved by the website HTML parsing is permanently stored as xml files for addition to the item-level metadata repository while data retrieved by broadcast metasearch is integrated into the search results but is not stored in the item-level metadata repository. Website HTML parsing allows retrieved metadata to be normalized, indexed, and classified, which facilitates the retrieval of item-level metadata records by users (Stern, 2009).

1.2. DESCRIBING THE DCC'S CONTENTS

In order to make the best assessment of the potential benefits of supplementing OAI-PMH, it is necessary to understand what is already in the DCC aggregate, both in its collection-level metadata registry and in its item-level metadata repository. There are two primary barriers to adequately articulating the contents of the DCC aggregate. These barriers are: the wide scope of the term "collection" and the heterogeneous nature of both the DCC aggregate's collection-level and item-level records. Overcoming these barriers is important to any discussion of the coherency and accuracy of the aggregate's contents.

1.2.1. DESCRIBING "COLLECTION"

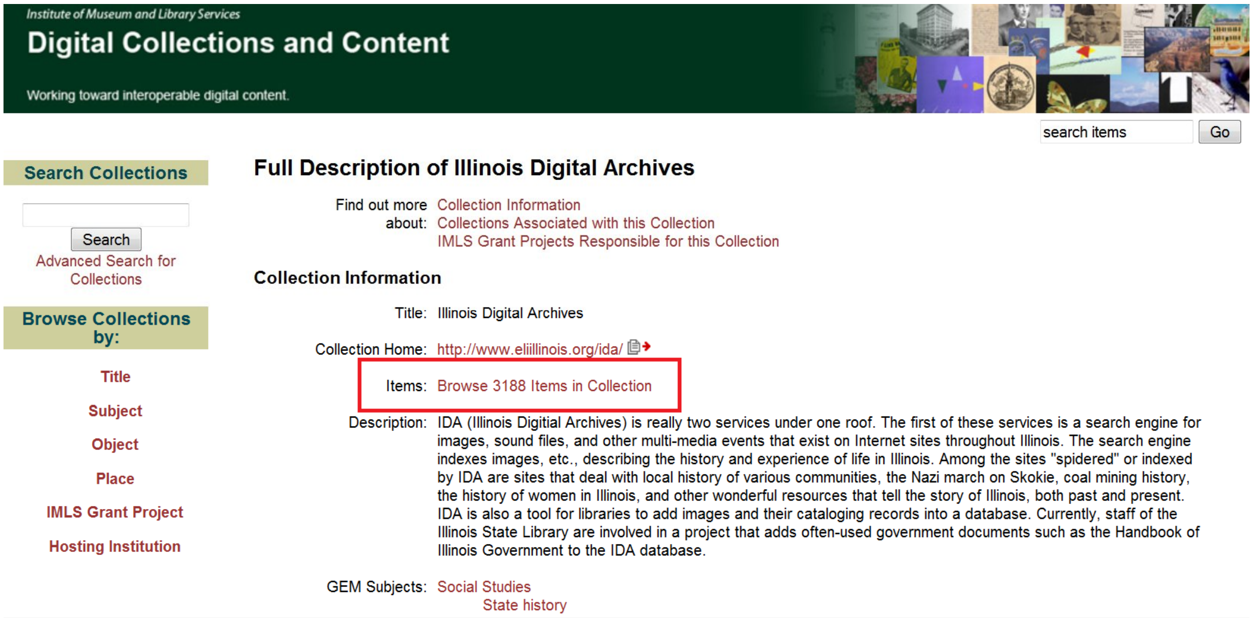
I noted above that as of December 2009 item-level metadata records had been harvested for only about 23% of the collections in the DCC's collection registry. This may or may not be an accurate account of the ratio of harvested collections to un-harvested collections. One key problem facing the DCC aggregate and similar digital projects is the concept of "collection."

As Palmer et al. noted in their 2006 ASIS&T Conference paper, “among digital content developers there is little agreement on what constitutes a collection;” however, accepting an anything goes definition of “collection” is going to prove to be a barrier to understanding the content of the DCC aggregate. What is needed is a strict definition of what is meant by “collection.” In simplest terms we could conceptualize a “collection” as a container for an arbitrary accumulation of things (items) which are related to one another along one or more axes (e.g. they were made by the same creator, they are of the same item type, they contribute information to the same topical body, etc.).

The United Kingdom’s Research Support Libraries Program articulated a very general definition of “collection” as a “term…[that] can be applied to any aggregation of individual items, where those items may be physical or digital” (Johnston & Robinson, 2003). The officially articulated use of “collection” with regards to the collections described by the records in the DCC registry added the additional criteria that to be considered a “collection,” the aggregation of items must also be: cohesive, searchable as a distinct collection, and available through a unique point of entry (Cole & Shreeves, 2004). The DCC definition allows for the existence of sub-collections within collections as a kind of individual digital object.

Unfortunately these criteria do not narrow what is meant by collection; if anything they broaden the meaning, making it even less useful for most practical purposes. In numerous instances it was observed during the initial survey of the DCC aggregate’s contents (below) that these principles are applied in an inconsistent manner. Sometimes there are collection records for both sub-collections and collections and the subsequent item-level records are associated with both collection levels, leading to item-level over-counts. It is also clear that there are super-collections and other aggregations present in the DCC’s aggregate whole (See Appendix B). These super-collections and other aggregations are represented as though they were collections. This has lead to situations where item-level records are associated with the grandparent super-collection but not their parent collection and contributes to the lack of accuracy regarding the percentage of collections actually harvested using OAI-PMH.

A good example of this problem is the Illinois Digital Archive. The Illinois Digital Archive (IDA) itself is a super-collection, consisting of multiple collections. Some of the item-level metadata harvested from the IDA is associated with a parent collection in the DCC aggregate, but other item-level metadata is associated directly with the super-collection (see Figure1.1).


Institute of Museum and Library Services
Digital Collections and Content
Working toward interoperable digital content.
search items
Go
Search Collections
Search
Advanced Search for Collections
Browse Collections by:
Title
Subject
Object
Place
IMLS Grant Project
Hosting Institution

Full Description of Illinois Digital Archives

Find out more about: Collection Information
Collections Associated with this Collection
IMLS Grant Projects Responsible for this Collection

Collection Information

Title: Illinois Digital Archives

Collection Home: http://www.eliillinois.org/ida/

Items: Browse 3188 Items in Collection

Description: IDA (Illinois Digitial Archives) is really two services under one roof. The first of these services is a search engine for images, sound files, and other multi-media events that exist on Internet sites throughout Illinois. The search engine indexes images, etc., describing the history and experience of life in Illinois. Among the sites "spidered" or indexed by IDA are sites that deal with local history of various communities, the Nazi march on Skokie, coal mining history, the history of women in Illinois, and other wonderful resources that tell the story of Illinois, both past and present. IDA is also a tool for libraries to add images and their cataloging records into a database. Currently, staff of the Illinois State Library are involved in a project that adds often-used government documents such as the Handbook of Illinois Government to the IDA database.

GEM Subjects: Social Studies
State history

Figure 1.1 Illinois Digital Archives DCC Collection Record

For the purposes of the analyses carried out below, I will be defining and using a stricter set of definitions with regards to "collection." These definitions are based upon observations of the perceived roles of a "collection" and the nature of its contents. These definitions are intended to facilitate discussion of harvesting activities with respect to aggregation services and to build a more accurate accounting of the quantity of collections' items aggregated into the DCC via OAI-PMH in particular. They are not meant to dictate the final word on what a "collection" is or is not beyond the context of this paper.

1.2.2. HETEROGENEOUS METADATA

Finally, the metadata records in the DCC's registry and repository describe a heterogeneous set of data resources. Some of them describe collections while others describe items. The mixture of item types alone is cause for difficulty when comparing and contrasting records. Further difficulty occurs because the institutional source for each set of item level records uses a different standard for the creation of those records. These problems are further compounded by the often inconsistent application of local metadata standards during the record creation process.

Many researchers (Han et al., 2009; Dunsire, 2008; Jackson et al., 2008; Jackson, 2006; Shreeves et al., 2005a; Lagoze, 2004; Brogen, 2003) have noted problems resulting from the Dublin Core (DC) metadata standard's lack of expressivity and OAI-PMH's reliance upon DC as a de facto lingua franca. Others have noted how the use of Dublin Core leads to problems with the quality and interoperability of the metadata collected by aggregators (Weagley et. al., 2010; Han et. al., 2009; Hillmann, 2008; Shreeves, 2005; Hider, 2004; Ward, 2003). Further, Dunsire's 2008 paper on harvesting metadata from institutional repositories documents the need for the aggregator to interpret harvested metadata formats into a single format to facilitate the harvested metadata's use for information retrieval. He notes, "Community agreement on a single metadata structure richer than unqualified DC is likely to be hampered because there is wide variation in the scope of resources to be described within a local repository, leading to divergent functional requirements between the institution and the community."

It could be said that the primary problem is one of consistency of scope. Records at every level are not created using consistent terminology because they do not have consistent domains and ranges of descriptiveness, and so no consistent picture of the DCC's contents can be articulated by means of record analysis alone. Indeed, we can expect that even if item-level metadata available for harvest via OAI-PMH conforms to a metadata standard, the specific information within item-level records' elements is likely to differ substantially from institution to institution.

CHAPTER 2

PROPOSED SOLUTIONS

I propose three complementary methods to further facilitate expansion of the contents of the DCC's item-level metadata repository. First, I propose classifying the contents of the aggregate by interface type, content type, and object type. Second, I propose expanding the methods by which item-level records are harvested into the DCC's item-level metadata repository. Third, I propose supplementing the item-level records in the item-level metadata repository with access to additional item-level resources through the means of broadcast metasearch.

2.1. CLASSIFICATION OF AGGREGATE CONTENTS

Having a clear understanding of the aggregate's contents would be useful for facilitating further aggregation activities. As I noted above, problems surrounding the notion of "collection" have made it difficult to articulate the exact nature of the aggregate's contents. In order to more clearly articulate the contents of the aggregate I classify those contents according to the object types in the collection registry and the content types contained within those objects (the items in the item-level repository).

An initial survey of the DCC's collection registry in December 2009 (see Appendix B) revealed that the "collections" within it are generally other aggregations of collections, thematic research collections, collections of homogenous content objects (e.g. Charles W. Cushman Photograph Collection), and a small number of miscellaneous objects (such as information retrieval (IR) portals).

The contents, that is to say the "items," of many of the collections, especially thematic research collections, are heterogeneous in nature; however, without delving into their specific nature, we can safely refer to these item-level objects as surrogates, as each specific item-level object is meant to take the place of a physical information resource. These specific items within the collections generally fall into one of five categories of surrogate:

- simple surrogates (e.g. a pdf file with no associated metadata record)

- compound surrogates (e.g. a digitized newspaper, where a single metadata record represents a large series of pdf or jpg files)
- complex surrogates (e.g. a single jpg file with an associated metadata record)
- integrated surrogates (e.g. a digitized photograph on Flickr (photograph and metadata integrated into a single HTML webpage) where the Flickr page itself is the digital information object)
- traditional surrogates (a catalog record containing metadata that describes a physical resource, such as a book)

These distinctions are important as the harvest of metadata records into the item-level metadata repository is a key feature of the DCC aggregate. It is important to distinguish between collections of simple surrogates which have no metadata records and those with complex surrogates where the metadata record is an important feature of the information objects in the collection.

For the purposes of exploring and gaining a deeper understanding of the contents of the DCC aggregate I use the following classification terminology (see Table 2.1). These criteria are based on both Johnston & Robinson's (2003) and Cole & Shreeves (2004) definition of a "collection." That is, I consider any arbitrary grouping of individual items that are cohesive, searchable as a distinct group, and available through a unique point of entry to be a collection.

Table 2.1: Digital "Collection" Classification Terminology

Term	**Definition**
Collection (generally)	any arbitrary grouping of individual digital items that are cohesive, searchable as a distinct group, and available through a unique point of entry
Thematic Research Collection	any collection that has been built according to the principles of contextual mass (Palmer, 2002)
Homogeneous Object Collection	any collection that is cohesive across one or more ubiquitous shared attributes (e.g. item type = photograph or creator = Charles Cushman)
Aggregate	any arbitrary accumulation of collections and their items that is constructed by harvesting those collections and items from other digital sources, which is, itself, cohesive and searchable but whose contents is available through alternate points of entry; a form of combined registry and repository

Table 2.1 (cont.)

Term	**Definition**
Collection Registry	any arbitrary accumulation of collections that are cohesive and searchable
Register	any arbitrary grouping of individual digital items that are cohesive, searchable, and available through a unique point of entry and whose item-level metadata is primarily transactional and/or provenance information
Repository	any arbitrary accumulation of items or item-level metadata that is searchable and available through a unique point of entry
Super-Collection	any arbitrary accumulation of collections and their items that is constructed by holding those collections and items natively (or organically), which is cohesive, searchable, and available through a unique point of entry; a form of repository
Sub-Collection	any collection which is a member collection of a super-collection

Generally, experimentation has shown that any "collection" whose website is organized such that has individual item-level records with persistent urls can be harvested algorithmically walking through the resulting web-pages and parsing the HTML for the pertinent metadata. I expand on and discuss both the experiment and experiment results later in this paper.

2.2. SUPPLEMENTAL HARVESTING

As has been stated before, the low rate of OAI-PMH adoption among data providers has presented a major obstacle to expanding the DCC's item-level metadata repository. In turn, the opportunities for DCC users to discover and remix information resources have also been constrained to just those collections for which item-level metadata could be harvested. To some extent, it has also limited the ability of aggregation collection developers to play a more active role in the item-level metadata harvesting process.

Recall that only around 23% of all "collections" in the DCC's collection registry have had item-level metadata harvested. Supplementing OAI-PMH through alternate means of harvesting item-level metadata would permit the DCC's item-level metadata repository to encompass a greater percentage of the collections in the collection registry. It would increase the quantity and, potentially, the types of information resources that the DCC's users can interact with. Finally, it would empower the DCC's collection developers by allowing them greater choices in deciding which collections to harvest item-level metadata from and at which pace and order the collections should be harvested in.

Providing means to supplement OAI-PMH metadata harvesting would greatly facilitate the activities of collection developers as it would greatly increase the pool of resources that can be chosen for inclusion in the DCC's item-level metadata repository. In this paper I review the results of my experiments with the website HTML parsing in order to examine the methodologies and potential benefits of supplemental metadata harvesting techniques.

When harvesting item-level metadata using website HTML parsing, a list of web-pages containing metadata records with persistent urls is first constructed by harvesting extant indexes of the website or by algorithmically indexing the website's webpages. The harvester then algorithmically crawls across the individual web-pages, parsing and normalizing the pertinent metadata and storing it as an xml file. These xml files can then added to the item-level metadata repository exactly as those xml files that had been harvested using OAI-PMH.

2.3. SUPPLEMENTAL ITEM-LEVEL ACCESS

Even though supplemental harvesting is expected to greatly expand the quantity of item-level metadata records that could be harvested into the item-level metadata repository, there will still be many collections whose item-level metadata records cannot be harvested. These harvesting shortfalls may occur for several reasons. The collection, if it is a newspaper archive for example, may consist solely or primarily of compound surrogates. Compound surrogates are generally not too desirable for harvest since they often have but a single item-level metadata record for a very large number of individual images (often hundreds or even thousands for newspapers). Alternately, the collection to be harvested may primarily consist of simple surrogates (e.g. digitized documents with no metadata). These are also undesirable for harvesting since most item-level metadata repositories require metadata in order for a digital surrogate to be retrievable via the repository's IR (information retrieval) interface. Finally, there are also likely to be collections whose content is deemed too specialized for addition to the item-level metadata repository even though the overall nature of the collection merits its inclusion in the collection registry. These collections represent those collections for which broadcast metasearch may be a more economical means of providing users item-level metadata access and include specialized collections such as botanical image collections.

Even if a collection is such that harvesting item-level metadata is impossible or inappropriate, it may still be possible to provide item-level access to an aggregate's end users by using broadcast metasearch. Broadcast metasearch operates by submitting a search query to a target website's IR portal and then retrieving the search results web-page. The results web-page is then parsed for the pertinent results information. This information can then be fed back to aggregate users in a variety of ways. A link reporting the number of available information resources can be provided on the aggregate's IR results web-page or the parsed results can be integrated directly into the aggregate's IR results web-page. In this way the quantity and variety of information resources available to an aggregate's end users can be further enriched. In Chapter 6, I review the results of my experiments employing broadcast metasearch as a means to supplement the contents of the DCC's item-level metadata repository.

CHAPTER 3

EXPERIMENT OUTLINE

After an initial survey of the contents of the DCC aggregate characterized both the type of collection (e.g. thematic research collection, super-collection, etc.) and the type of surrogates (e.g. compound surrogates, complex surrogates, etc.) in those collections (see Appendix B). Once this had been completed the sum of the potential quantitative outcomes of harvesting un-harvested collections could be computed. From these calculations 25 collections were selected based on the perceived value that the harvesting of or access to item-level records would add to the aggregate. These collections were examined further to determine their actual suitability for experimentation and, if found suitable, to determine which experimental track they should be assigned to by classifying them according to their home collection website's organization scheme.

3.1. INITIAL SURVEY OF AGGREGATE CONTENTS

As of 12/15/2009 the DCC collection registry contained records for 341 collections. Of these, only 78 collections had had item-level metadata harvested. Of the remaining 263 collections, 5 of them were clearly not collections except with regards to the loosest possible interpretation of the original grant proposal (i.e. they are NLG funded digital projects). Examining the collection records it quickly becomes clear that they suffer from quality problems that are similar in nature to those observed in item-level records (Hillmann, 2008; Jackson et. al., 2008; Jackson, 2006; Shreeves et. al., 2005a). Specifically, the metadata contained within them is not consistent from record to record. Such useful quantitative metrics as "size" are often blank, unknown or contain erroneous data. As the collection-level records are created in an uncontrolled manner using a combination of survey completion by collection administrators and manual review of the collection portal by the DCC's project coordinator (Benevento, 2005), their heterogeneous nature is not unexpected.

Turning back to the issue of collection "size," there are no listed values for 83 of 341 DCC collection records at this time (just over 24% of all collection records, see Appendix A). Of these 83 records, 8 of them have actually had item-level records harvested and hence, actually have known collection sizes (see Table 3.1). It is also clear that if we classify the collections within the collection registry a very interesting picture emerges. Six of the collections are actually other aggregations of digital collections, which is to say that each of these 6 collections are super-collections that harvest item-level metadata from other metadata repositories and are not the primary hosts for the item-level metadata in their own item-level metadata repositories. Forty of the collections are ordinary super-collections, containing 2 or more sub-collections and which are the primary hosts for the item-level metadata held in their sub-collections. Further distinctions between the objects being described by the collection-level records became clear during the deeper survey (see Chapter 4).

Table 3.1: DCC collections with harvested items but no "size" metadata

Collections with missing "size" metadata	Number of items actually harvested
American Natural Science in the First Half of the Nineteenth Century	349
Arizona-Sonora Documents Online	2,887
Flora and Fauna of the Great Lakes Region: A Multimedia Digital Collection	26,300
Kinetic Models for Design Digital Library	787
Linking Florida's Natural Heritage	582
Maine Music Box	18,708
Photohio.org	23,545
WPA TVA Archaeological Photograph Archive	9,873

Examining the remaining "collections" reveals that about 64% of all of the collections in the DCC aggregate are what I call homogeneous object collections (see Table 3.2 and Appendix B). That is, 64% of the collections in the DCC aggregate are built around one or more ubiquitous attribute values. Some of them, like the Charles W. Cushman Photograph Collection, are collections built around a shared item type and creator. Others are built around a shared topical value but are not thematic research collections because they were not built according to the principles of contextual mass (Palmer, 2002).

Table 3.2: Contents of the DCC aggregate according to collection class type

Class	Number	Percent of DCC Aggregate
Aggregation	6	1.8%
Catalog	4	1.2%
Collection (general)	7	2.0%
Homogeneous Object Collection	219	64.2%
Index	2	0.6%

Table 3.2 (cont.)

Class	Number	Percent of DCC Aggregate
IR Portal	3	0.9%
Monograph	2	0.6%
Multi-media Resource	1	0.3%
Super-Collection	40	11.7%
Thematic Research Collection	39	11.4%
Video Game	1	0.3%
Website	11	3.2%
Unknown	6	1.8%

Despite problems adequately articulating a strict definition for “collection,” the aggregate’s contents are mostly collections of one type or another. 77.6% of all collection records actually refer to objects easily understood to be collections. Further, 13.5% of collection records refer to aggregations or super-collections which contain collections; however, 7.1% of collection records refer to digital objects that are not actually collections but items, registers, or what could be best be interpreted as websites. The remaining 1.8% is listed as unknown because they were not functioning correctly during the course of this project. Of these, one (Digital Past) is known to have been a super-collection but at the time of writing is in the process of being consumed by another super-collection and no longer has a functioning point of entry for itself or any of its sub-collections.

The relationships of super-collections and aggregations with their associated sub-collections play an important role in defining the content of the registry. Deep examination of the collection records and the resources that they describe reveals that super-collections are not treated consistently within the aggregate. Some of the super-collections are decomposed into their separate sub-collections. Collection-level records exist for both the parent super-collection and some or all of its sub-collections. In one case, the Arizona Memory Project, the super-collection has almost entirely been decomposed into its constituent collections; collection-level records appear in the DCC aggregate for 81 of its 93 sub-collections. The lack of a collection-level record for the super-collection, the Arizona Memory Project, is quite noticeable. Its sub-collections make up a significant minority (23.8%) of the “collections” in the aggregate. Further, the percentage of all “collections” that are actually sub-collections of super-collections or aggregations is 51.3%. These numbers have additional implications for the experiment. Since so many collections are sub-collections, the number of targets that a harvester would need to harvest metadata from is significantly reduced. Conversely the potential benefits of harvesting are substantially increased since item-level metadata from multiple collections can potentially be harvested in one packaged harvest not unlike the functionality already provided by those super-collections that use OAI-PMH.

3.2. EXPERIMENT TARGET SELECTION

Examination of the collection-level metadata for the 263 “collections” for which it appears that item-level records have not been harvested reveals that there are 75 that also lack “size” metadata. Examining the collection-level metadata for the remaining 188 collections that do have “size” metadata reveals that there are over 2.7 million potential resources for which no means of item-level access exists at the level of the DCC IR portal.

The survey numbers indicate that supplementary methods of harvesting may present a very powerful remedy to OAI-PMH’s harvesting shortfalls. Since so many collections are members of super-collections, the actual number of unique targets is greatly reduced; however, many inconsistencies within the collection-level metadata remain. For the purposes of initial experimentation targets are selected using “size” criteria so that expected outcomes can be calculated.

25 collections from the 188 noted above have been selected in order to test the proposed solutions for improving the ratio of collections within the DCC registry for which item-level access can be provided (either through harvesting or via federated search). These collections have been selected as they are thought to be the largest collections (as purported by the data in the collection records) for which item-level metadata has not been harvested. If the quantities listed in “size” for these collections are correct then they represent access to approximately 2.6 million items. Many of these items should be directly complementary to items already harvested into the DCC’s item-level metadata repository. Several of the targets were selected to demonstrate the potential benefits and difficulties unique items represent (such as the quilts in the Quilt Index collection). A list of the target collections is (see Appendix C for full details):

- Beyond the Shelf: Serving Historic Kentuckiana Through Virtual Access

- Brooklyn Daily Eagle Online
- Colorado’s Historic Newspaper Collection
- Cuneiform Digital Library
- Dallas Museum of Art Collections
- Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois
- Exploratorium Digital Asset Management Collection (EDAM)
- Florida Folklife Collection
- GATT Digital Library: 1947-1994
- George Edward Anderson Collection
- HEARTH (Home Economics Archive: Research Tradition, and History)
- History at our Hands: The Ponce’s Historical Archive & Historical Museum Digitized (Coleccion Historia de Puerto Rico)
- John Brown/Boyd B Stutler Collection Database
- Main Memory Network
- Mind Models: Artificial Intelligence Discovery at Carnegie Mellon
- Montana Memory Project
- New York Public Library’s Picture Collection Online
- Plant Images at Missouri Botanical Garden
- PlantCollections
- Quilt Index
- TIDES: Teaching, Images, & Digital Experiences
- Upper Mississippi Valley Digital Archive
- Utah Digital Newspapers
- Vanishing Georgia
- Virtual Motor City: Images from the Detroit News

The target collections were first surveyed to see if item-level metadata could be harvested via OAI-PMH. For those that had not implemented OAI-PMH, they were further assessed to determine their appropriateness for experimentation using supplemental harvesting techniques or federated search techniques.

3.3. EXPERIMENTAL TRACKS

The experiment proceeded in three parts. In the first part the 25 target collections were surveyed to determine their appropriateness for participation in experiments for harvesting item-level metadata through website HTML parsing or for participation in experiments for providing access to item-level metadata through broadcast metasearch. Once the survey was completed, experimentation began to determine the methodologies and potential benefits of using website HTML parsing to harvest item-level metadata from appropriate targets. Once the harvesting experiments were completed, experimentation to determine the methodologies and potential benefits of applying broadcast metasearch to the appropriate targets. The methods, results, and implications of these experiments are discussed in the remaining chapters.

CHAPTER 4

SURVEY OF TARGET COLLECTIONS

To determine which target collection should be used for each experiment an even more thorough survey was made of the 25 target collections. The results of this survey actually precluded several collections from the experiments as they either had been or could be harvested using OAI-PMH or they were found to be both undesirable for both harvesting and access via metasearch.

4.1. METHODS

Each of the websites hosting the 25 collections was closely examined to determine three things:

1. Has the source data provider, in fact, implemented OAI-PMH since the initial collection record was added to the DCC's collection registry?
2. If OAI-PMH has not been added, can the collection be harvested via an alternate, automated method, such as website HTML parsing?
3. If it cannot be harvested or contains information objects that are not appropriate for harvesting, can item-level records be accessed via broadcast metasearch techniques (which also leverage screen scraping technologies)?

The survey methods were based on the same observational, naturalist methodology developed for the preliminary survey of the aggregate's contents. In addition to determining whether or not OAI-PMH had been implemented, the nature of each collection's information objects was observed and then classified as one of five types (Table 4.1).

As the goal of harvesting item-level metadata is that extant metadata records be harvested, in the context of these experiments, some types of surrogates are more desirable for harvest than others. For instance, a simple surrogate, like a pdf file, with no separate metadata file, is probably of little desirability for the type of harvesting in these experiments and, when limited to the scope of these experiments, probably of greater value when accessed using broadcast metasearch. Since the pdf file's home collection has already been optimized to take advantage of the metadata encoded directly into the pdf file, the broadcast metasearch method can exploit that optimization to the aggregate user's advantage. Additional workflows would have to be developed in order for an aggregation service to generate separate metadata files based on the metadata encoded within the pdf file. While in the same vein as the experimentation in this specific paper, scraping metadata from within pdf files is outside of the scope of these experiments but I conjecture that such an approach would alter the table (Table 4.1) presented below such that simple surrogates would be highly desirable for harvesting via pdf file parsing.

Table 4.1: Information Object Classes

Class	**Desirability for Harvest**	**Desirability for Metasearch**
Simple Surrogate (item with no metadata record)	Low	Moderate
Compound Surrogate (many items with one metadata record)	Moderate	High
Complex Surrogate (item with metadata record)	Very High	Very High
Integrated Surrogate (metadata integrated into item)	Very Low	Very Low
Traditional Surrogate (metadata record without item)	Low	Very High

Additionally, the website organization scheme of each of the 25 target collections was classified in order to determine if the collection was appropriate for experimentation and which experimental track each collection was appropriate for. The organization schemes were classified generally by whether or not they had an IR portal through which their contents could be searched and whether or not that portal was password protected (i.e. closed or open). The ability of users to easily browse through the individual metadata records was also noted (as browsable or not browsable; see Table 4.2). Each of the browsable classes has additional sub-classes that are dependent on whether or not the collection's contents could be or has been indexed.

Table 4.2: Collection Website Organization Schemes

Website Organization Scheme	**Harvestable via Website HTML Parsing?**	**Accessible via Broadcast Metasearch?**
Closed Search, Browsable (password protected access)	Yes (with authentication)	Yes (with authentication)
Closed Search, Not Browsable (password protected access)	No	Yes (with authentication)
Open Search, Browsable	Yes	Yes

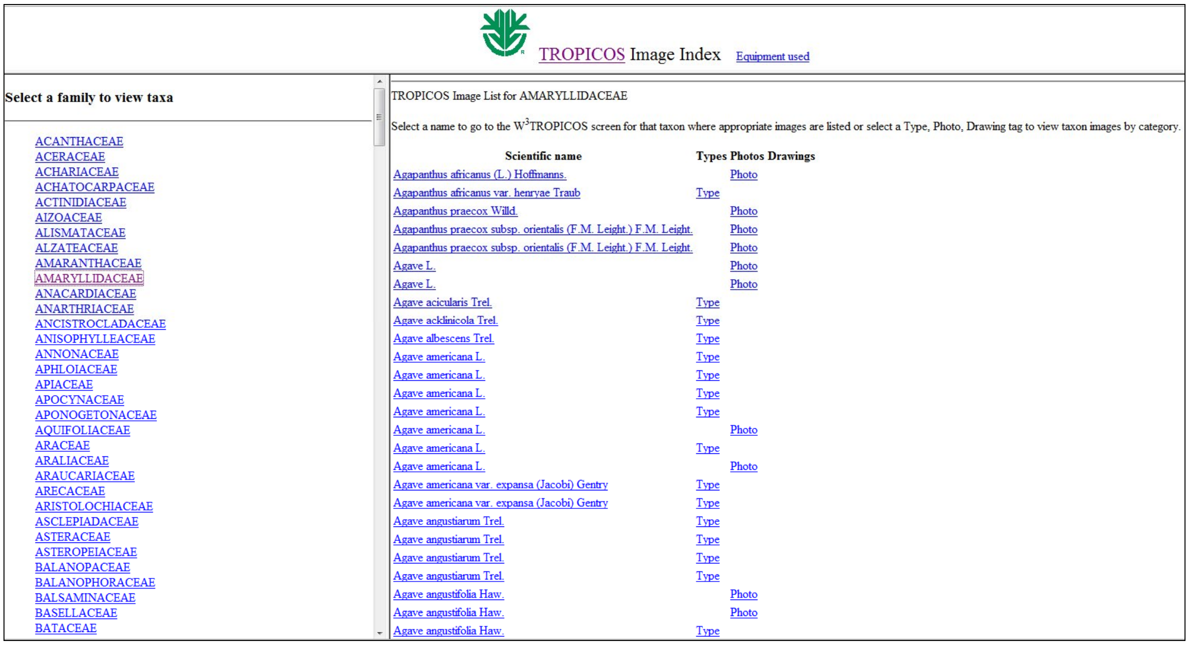
Table 4.2 (cont.)

Website Organization Scheme	Harvestable via Website HTML Parsing?	Accessible via Broadcast Metasearch?
Open Search, Not Browsable	No	Yes
No Search, Browsable	Yes	No
No Search, Not Browsable (an ordinary web-page)	No	No

4.2. RESULTS

Of the 25 collections selected for experimentation four of them either have already been harvested using OAI-PMH or can be (in the case of the Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois). One of the 25, Mind Models, is actually 2 separate collections which only have simple surrogates. 14 of the 25 contain primarily complex surrogates, often digitized photographs with records. Seven of the 25 collections contain primarily compound surrogates and three of these collections are newspaper archives. The remaining three collections possess mixed types. (See Appendix D for details.)

Some 84% of the target collection sites are classified as OSB-I; that is, no password required for access, searchable, browsable, and either indexed or indexable. Of the remaining four collections, three of the sites are open to the public and searchable but are not browsable or indexible. The final collection, Plant Images at Missouri Botanical Garden (Figure 2), is both browsable and indexible but provides no search interface for its users and is instead arranged as a relatively static, flat index.



TROPICOS Image Index Equipment used

Select a family to view taxa

ACANTHACEAE
ACERACEAE
ACHARIACEAE
ACHATOCARPACEAE
ACTINIDIACEAE
AIZOACEAE
ALISMATACEAE
ALZATEACEAE
AMARANTHACEAE
AMARYLLIDACEAE
ANACARDIACEAE
ANARTHRIACEAE
ANCISTROCLADACEAE
ANISOPHYLLEACEAE
ANNONACEAE
APHLOIACEAE
APIACEAE
APOCYNACEAE
APONOGETONACEAE
AQUIFOLIACEAE
ARACEAE
ARALIACEAE
ARAUCARIACEAE
ARECACEAE
ARISTOLOCHIACEAE
ASCLEPIADACEAE
ASTERACEAE
ASTEROPEIACEAE
BALANOPACEAE
BALANOPHORACEAE
BALSAMINACEAE
BASELLACEAE
BATACEAE

TROPICOS Image List for AMARYLLIDACEAE

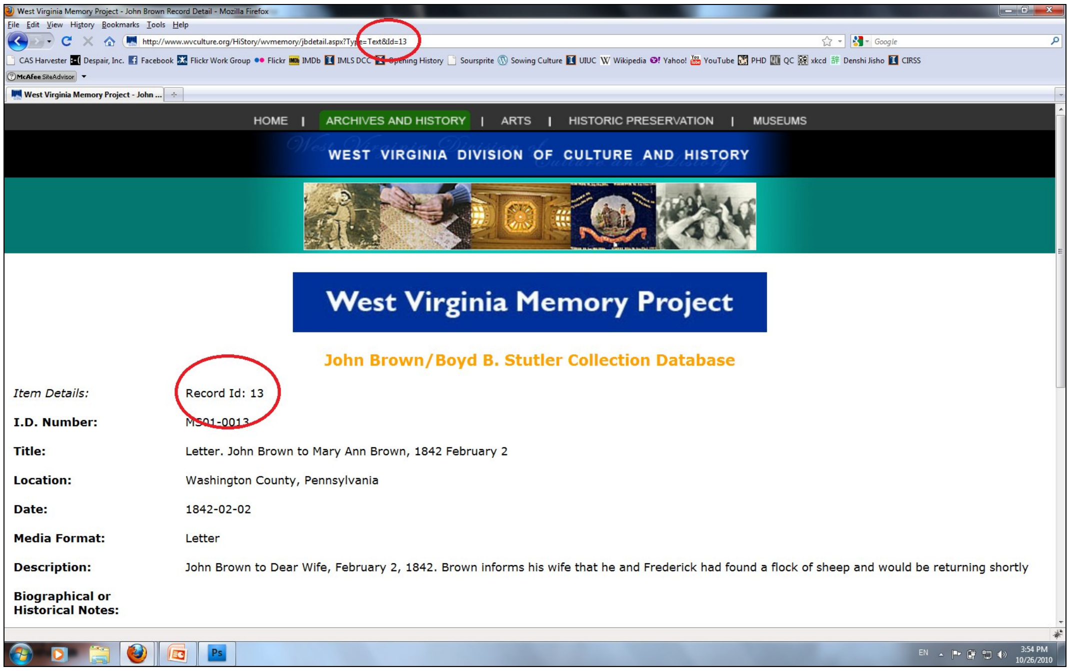
Select a name to go to the W^{3}TROPICOS screen for that taxon where appropriate images are listed or select a Type, Photo, Drawing tag to view taxon images by category.

Scientific name	Types	Photos	Drawings
Agapanthus africanus (L.) Hoffmanns.		Photo	
Agapanthus africanus var. henryae Traub	Type		
Agapanthus praecox Willd.		Photo	
Agapanthus praecox subsp. orientalis (F.M. Leight.) F.M. Leight.		Photo	
Agapanthus praecox subsp. orientalis (F.M. Leight.) F.M. Leight.		Photo	
Agave L.		Photo	
Agave L.		Photo	
Agave acicularis Trel.	Type		
Agave acklinicola Trel.	Type		
Agave albescens Trel.	Type		
Agave americana L.	Type		
Agave americana L.	Type		
Agave americana L.	Type		
Agave americana L.	Type		
Agave americana L.		Photo	
Agave americana L.	Type		
Agave americana L.		Photo	
Agave americana var. expansa (Jacobi) Gentry	Type		
Agave americana var. expansa (Jacobi) Gentry	Type		
Agave angustiarum Trel.	Type		
Agave angustiarum Trel.	Type		
Agave angustiarum Trel.	Type		
Agave angustiarum Trel.	Type		
Agave angustifolia Haw.		Photo	
Agave angustifolia Haw.		Photo	
Agave angustifolia Haw.	Type		

Figure 4.1. Plant Images at Missouri Botanical Garden

From here I break the target collections into three preliminary groups. Those collections which have been or can be harvested via OAI-PMH, those collections selected for experiments with supplemental harvesting techniques, and those collections selected for experiments with metasearch techniques (see Table 6 and Appendix D).

Since the purpose of these experiments is to supplement OAI-PMH and not compete with it, if I found that a collection had been harvested using OAI-PMH, could be harvested using OAI-PMH, or was scheduled to be harvested via OAI-PMH then, I eliminated it from the experiment. The remaining candidate collections were split up according to whether or not I could conceive of a method of harvesting their metadata records using website HTML parsing. Those that I could conceive of a method, such as the John Brown/Boyd B. Stutler collection which used simple numerical schemes within their resources' persistent urls (Figure 3), were selected for harvesting experiments with website HTML parsing. The remainder of the candidate collections was set aside for broadcast metasearch experiments. Many selection errors were made at this stage as it was revealed during the course of the experiments that many of the collections initially set aside for broadcast metasearch experiments could actually be harvested using website HTML parsing.


West Virginia Memory Project - John Brown Record Detail - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.wvculture.org/HiStory/wvmemory/jbdetail.aspx?Type=Text&Id=13
Google
CAS Harvester Despair, Inc. Facebook Flickr Work Group Flickr IMDb IMLS DCC Opening History Soursprite Sowing Culture UIUC Wikipedia Yahoo! YouTube PHD QC xkcd Denshi Jisho CIRSS
McAfee SiteAdvisor
West Virginia Memory Project - John ...
HOME | ARCHIVES AND HISTORY | ARTS | HISTORIC PRESERVATION | MUSEUMS
WEST VIRGINIA DIVISION OF CULTURE AND HISTORY
West Virginia Memory Project

John Brown/Boyd B. Stutler Collection Database

Item Details: Record Id: 13

I.D. Number: MS01-0013

Title: Letter. John Brown to Mary Ann Brown, 1842 February 2

Location: Washington County, Pennsylvania

Date: 1842-02-02

Media Format: Letter

Description: John Brown to Dear Wife, February 2, 1842. Brown informs his wife that he and Frederick had found a flock of sheep and would be returning shortly

Biographical or Historical Notes:

Figure 4.2. J.B./B.B.S. Collection Persistent URL Numbering Scheme

Table 4.3: Experimental Groups

Eliminated from Experiments	**Harvesting Experiments using Website HTML Parsing**	**Broadcast Metasearch Experiments**
Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois	Cuneiform Digital Library	Beyond the Shelf: Serving Historic Kentuckiana Through Virtual Access
George Edward Anderson Collection	GATT Digital Library: 1947-1994	Brooklyn Daily Eagle Online
HEARTH (Home Economics Archive: Research, Tradition, and History)	John Brown / Boyd B. Stutler Collection Database	Colorado’s Historic Newspaper Collection
Mind Models	Maine Memory Network	Dallas Museum of Art Collections
Montana Memory Project	New York Public Library's Picture Collection Online	Exploratorium Digital Asset Management Collection (EDAM)
	Plant Images at Missouri Botanical Garden	Florida Folklife Collection
	Quilt Index	History at our Hands: The Ponce's Historical Archive & Historical Museum Digitalized (Coleccion Historia de Puerto Rico)
	TIDES: Teaching, Images & Digital Experiences	PlantCollections
	Upper Mississippi Valley Digital Image Archive	
	Utah Digital Newspapers	
	Vanishing Georgia	
	Virtual Motor City: Images from the Detroit News	

4.2.1. CANDIDATE COLLECTIONS ELIMINATED FROM EXPERIMENTS.

Of the five collections found to have been harvested or be harvestable by OAI-PMH, one, the Montana Memory Project was discovered to be in the project's OAI-PMH harvesting queue. Another collection, Mind Models, was found to be unsuitable for the experiments. The remaining three collections all represent examples of some of the technical problems faced by the DCC aggregate. Two of them, the George Edward Anderson Collection and HEARTH have already actually been harvested. Their items appear in the DCC's sister aggregate, Opening History. It is not known why they appear there but not in DCC's item-level metadata repository. Finally, the Digital Archive of Historical Aerial Illinois Photography is a collection that is hosted by the Illinois Digital Archives.

Since the records are administered by the Illinois State Library, it seems likely that they may also be hosted at the Illinois Digital Archives site. IDA (the Illinois Digital Archive) is a digital initiative based at the Illinois State Library that administers digital collections from many of the state's public libraries and museums. More importantly there is a collection record for IDA in the DCC registry. In my preliminary analysis, I found that IDA was a super-collection with many sub-collections. Some of IDA's collections have already been harvested, so if this collection is in IDA it should either be one of the collections that have already been harvested or be easily harvested from IDA. Searching IDA, I readily find the Aerial Photograph collection (http://www.idaillinois.org/cdm4/browse.php?CISOROOT=%2Fisgs).

Examining the project's original data archive I find that this collection has not yet been harvested from IDA but, it can readily be harvested from IDA using OAI-PMH, as IDA has configured its CONTENTdm implementation to serialize collections' item-level metadata using OAI-PMH. Upon further examination I discover that the metadata records describe the individual html mosaics of links. The information objects in this collection are compound surrogates, and most closely resemble monographs, specifically photo albums in this case.

Finally, at first glance, Mind Models appears to be a super-collection with 2 sub-collections; however, there is no search interface and each collection has another unique entry point. So it would be more accurate to say that this is two separate collections. One collection is the Herbert

Simon Collection (http://diva.library.cmu.edu/Simon/) and the other is the Allen Newell Collection (http://diva.library.cmu.edu/Newell/). Each collection has a distinct search and browsing interface and users cannot search or browse across both collections simultaneously, so Mind Models is missing all of the structure and functionality that is the hallmark of a true super-collection. Although the highly structured browsing index should make it easy to carry out a harvest, the lack of metadata records makes it undesirable to harvest these collections as parsing the pdf files themselves is outside of the scope of these experiments. Further the narrow scope of the collections makes them less desirable for federated search targets, except for queries where their contents is likely to be relevant, such as topical searches for artificial intelligence. As such this website might be a viable target for the dark target broadcast metasearch methods discussed in chapter 6.

4.2.2. TARGETS FOR HARVESTING EXPERIMENTS USING WEBSITE HTML PARSING

I chose the targets for harvesting experiments primarily on the basis of their containing complex surrogates (i.e. one item with one metadata record). Generally these targets can be separated into two types: those with heterogeneous (home-grown) interfaces and those using CONTENTdm. Building website HTML parsers to collect metadata from many of the heterogeneous interfaces seems as though it will be easy to do as each of their items has a unique, persistent url by which it can be accessed. Implementing a website HTML parser for CONTENTdm looks somewhat more challenging. I have included the Utah Digital Newspapers collection here simply because it uses CONTENTdm as its interface. As a newspaper archive, Utah Digital Newspapers is arguably a more appropriate target for metasearch experimentation. This is due to the nature of the compound surrogates that appear in newspaper archives. Typically there are very few metadata records for large numbers of images which makes the utility of harvesting the item-level metadata somewhat dubious.

Finally the Cuneiform Digital Library appears to be a unique target amongst the ones included here since the contents of its item-level metadata are available for download in zip file format (http://cdli.ucla.edu/tools/cdlifiles/cdlicat_20090905.zip). Although primarily an aggregation of collections from 15 institutions, it would also be fair to interpret the Cuneiform Digital Library as a super-collection. It provides a single homogeneous style of access to the collections of multiple institutions that individually have heterogeneous access methods. It can be harvested and it can also be accessed via federated search techniques. The information objects in the Cuneiform Digital Library are uniformly complex surrogates, where there is a metadata record for every item. Hypothetically, harvesting the item-level metadata records for the Cuneiform Digital Library should be as easy as downloading the zip file.

A final interesting note regarding the Cuneiform Digital Library, the collection-level record in the DCC registry is internally inconsistent. It provides one figure for size in the description element and a much smaller one in the size element. This type of inconsistency has made it difficult to accurately project the benefits of harvesting some of the collections that are in the DCC registry.

4.2.3. TARGETS FOR BROADCAST METASEARCH EXPERIMENTS

I chose the targets for broadcast metasearch experiments primarily on the basis that they contain multiple types of surrogates (e.g. both complex and compound surrogates). Beyond the Shelf is a good example of this type of "collection." A good example of this is Beyond the Shelf (a.k.a. Kentuckiana). Kentuckiana is both harvestable and accessible via federated search techniques; however, since Kentuckiana is an aggregator and only provides collection-level records for three of its eight item-type "collections," it is probably best suited for item-level access via broadcast metasearch. Since search terms can be submitted in a manner that allows for searching simultaneously across all eight of its "collections" a broadcast metasearch solution seems viable.

Other targets in this category are newspaper archives like the Daily Eagle Online. To the user, the Daily Eagle's content is presented as a series of simple surrogates but since it runs on an xml repository it is probable that the entire database is a single compound surrogate; that is to say, there is a single metadata record that represents the contents of the entire database (the newspaper's entire print run). It is very probable that the entire xml repository could be harvested but, as this would be one extremely large item from the DCC aggregation's point of view it seems unlikely that harvesting this newspaper archive would be beneficial. Item level access to individual pages and articles seems potentially beneficial, especially to those doing historical research, so I include this target in my initial list of targets for federated search experiments. As search terms cannot readily be submitted to the database's IR (information retrieval) system using HTTP POST via the url, implementing federated search may be difficult.

4.3. DISCUSSION

Of the initial 25 targets selected, 4 have implemented OAI-PMH or are members of super-collections who have implemented OAI-PMH. Two of collections (HEARTH and the George Edward Anderson Collection) have already been harvested but are not appearing in one of the two pertinent web portals. This indicates the presence of some technical inconsistencies within the DCC's infrastructure.

There are additional sub-collections which have had item-level metadata records harvested but for which those item-level metadata records are not associating with the correct collections. Specifically sub-collections belonging to the Illinois Digital Archives super-collection are having their items associated with the super-collection in the DCC's web interface, which makes it appear as though no item-level records have been harvested for the pertinent collection. These problems have resulted in the loss of some of the collection-level context that the aggregation is attempting to preserve. This problem occurs for the following collections:

- Arthur, Once Upon a Time (relevant identifier: http://www.idaillinois.org/u?/apl)
- Coal Mining in Illinois, Machine vs. Man (relevant identifier: http://www.idaillinois.org/u?/ccpl)
- Oak Ridge Cemetery, Illinois Interment Records (relevant identifier: http://www.idaillinois.org/u?/linl3)
- Park Forest (relevant identifier: http://www.idaillinois.org/u?/pfpl)
- A University Goes to War, World War I Women (relevant identifier: http://www.idaillinois.org/u?/isu)

There are also additional sub-collections for Illinois Digital Archives which, like the Digital Archive of 1936-1941 Historical Aerial Photography, have collection-level records but for which no item-level records have been harvested via OAI-PMH. These collections are:

- William Hayes Collection, 1820-1860 (http://www.idaillinois.org/cdm4/browse.php?CISOROOT=%2Fspl)
- World's Columbian Exposition of 1893, and the Founding and Early History of The Field Museum (http://www.idaillinois.org/cdm4/browse.php?CISOROOT=%2Ffmnh)

Unfortunately a more detailed examination to discover the causes of these inconsistencies could not be made at this time. This problem should be explored in the future, but it may generally indicate that data aggregators should spend some time becoming familiar with the data that is being ingested into their collection registries and item-level metadata repositories. It seems likely that there may be some disconnect between the internal hierarchies of collection-level metadata, especially with respect to sub-collection/super-collection relationships.

Further, internal inconsistencies in many of the DCC collection level metadata records has made it difficult to quantitatively predict the benefits of supplementing OAI-PMH through alternate means of harvesting and access via federated search applications with much accuracy. In her 2008 paper, Diane Hillmann noted that consistency of the content within the fields is a mark of quality. Internal consistency between the assertions within a metadata record's elements should be the very first benchmark of quality. If a record is not internally consistent then, there is little hope that its content will meet other established standards of interoperability.

Finally, despite the fact that items from some collections are not being correctly associated with their parent collections, the problem occurs in less than 1% of collections. As such the 23% ratio for collections successfully being harvested using OAI-PMH appears to be [essentially] correct.

CHAPTER 5

HARVESTING ITEM-LEVEL METADATA VIA WEBSITE HTML PARSING

Website HTML Parsing was carried out on a local server using server side scripting techniques employing the VBScript language. While a wide variety of programming languages can be used to implement website HTML parsing (what Schrenk (2007) calls simple screen scrapers), VBScript was chosen primarily because of my past experiences using it in federated search applications. Broadcast metasearch resources rely on website HTML parsing, among other data aggregation technologies, to produce results (Mischo, 2004). The primary difference for these experiments is what happens to the data once it is parsed. In the website HTML parsing case the webpages are retrieved algorithmically, independent of user interaction. Once retrieved they are then parsed for the relevant data, which in turn, is saved as an xml file usually conforming to the qualified Dublin Core metadata standard.

5.1. METHODS

The key feature of harvesting metadata records via website HTML parsing is the metadata crosswalk. It has often been noted that the differences in metadata standards are the primary barrier to interoperability of the records of digital libraries, archives, and museums (Bountouri & Gergatsoulis, 2009; Chan, 2005). An aggregator operates by collecting sets of heterogeneous data that describe heterogeneous information objects that users need to find. In order to both facilitate users' ability to find information objects within the aggregate's item-level metadata repository and to create a consistent information retrieval experience for users, it makes sense for all of the metadata presented to a user to conform to a single standard. The need to crosswalk metadata in the DCC aggregate has to some extent been greatly reduced by the use of OAI-PMH as the metadata harvesting standard, as Dublin Core is the de facto metadata standard for use with OAI-PMH (Lagoze, 2004; Jackson et. al., 2008).

Parsing the websites of digitized collections presents a new dimension in crosswalking metadata. The metadata must be interpreted from a local display standard, which may or may not provide all of the metadata actually stored in the source record to a standard that can be exploited by the DCC aggregate's item-level metadata repository. For the initial experiment, the Dublin

Core standard was used as the standard for the harvester's output. All of the remaining experiments produced xml files conforming to the Qualified Dublin Core standard.

The crosswalks were hand constructed and based on direct observation of the metadata presented to each collection's users. On average, I examined a sample of approximately 50 random records (this means some smaller collections were examined completely) to determine the kinds of metadata fields the parser would need to find, collect, and map. For several sites I was able to use pre-existing metadata policies that clearly noted the types of metadata being used by the collection, as was the case with the Quilt Index. PastPerfect collections, in particular, often used a core set of metadata fields to describe items within their collections. When it seems useful or predictable, certain metadata fields were simply filled in using the parser's code (e.g. format). As it took an average of about an hour to construct each crosswalk, repositories that published their overall metadata standard, like Quilt Index and the Upper Mississippi Valley Digital Image Archive, provided a substantial

The heterogeneous targets presented a number of parsing challenges as each used unique HTML structures and in the case of the Quilt Index, a large amount of unique vocabulary such as:

- Quilter
- Top by
- Quilted by
- Construction [technique]

The homogenous targets were also challenging since they often used different mixtures of metadata elements between the sub-collections within the same supper-collection.

5.2. RESULTS

The initial list of targets for harvesting experiments included 12 of the 25 target collections described in Section 4. These targets were selected according to the criteria detailed in Section 4. They represented a mix of eight heterogeneous and three homogeneous (CONTENTdm) targets. In one case, the Cuneiform Digital Library, the collection catalog was available for download as a zipped FileMaker Pro database file.

As early CONTENTdm experiments seemed inconclusive, three additional target collections were added (Illinois Digital Archives, Long Island Memories, and Olympic Peninsula Community Museum). Further, as the harvesting experiments were generally successful, experimentation was expanded to encompass PastPerfect, a standard interface software client gaining popularity among the digital museum community. Three target collections were selected for the experiments (Lewis & Clark Trail Heritage Foundation, Longmont Museum and Cultural Center Photo Collection, and Center for Sacramento History). Finally, as the CONTENTdm experiments were expanded and experiments with PastPerfect were added, 4 of the heterogeneous collections were eliminated from the experiments. These collections were: the GATT Digital Library, New York Public Library's Picture Collection Online, Plant Images at Missouri Botanical Garden, and Vanishing Georgia.

5.2.1. HOMEGROWN (OR HETEROGENEOUS) COLLECTION WEBSITE ORGANIZATION SCHEMA

5.2.1.a. John Brown / Boyd B. Stutler Collection Database

Harvesting of the Collection Database began with an analysis of the surrogates within the collection. They are uniformly of two complex surrogate types, images with records or text with records. In the case of the text surrogates, a uniform set of metadata (short record) is always presented to the user, with the option to view the full (MARC) record. Rather than capture the full record, I decided to capture the short record metadata, as it is the default metadata being displayed to the users. My reasoning for this is threefold:

1. It remains unclear exactly how the richness of item-level metadata records relates to retrieval of the items they are attached to.
2. There are likely to be a number of unexplored ethical issues regarding this item-level metadata harvesting methodology.
3. Finally, as these experiments are proof of concept experiments, the short record seems both more appropriate and eminently sufficient for the task at hand.

The short record elements are: Record Id, I.D. Number, Title, Location, Date, Media Format, Description, Biographical or Historical Notes, and Text. The final element is not metadata but rather a section of the display page that has been set aside for an html surrogate of the source item (i.e. the full text of the document has been stuffed into the webpage, within a "metadata" element called "text"). The html surrogate is further supplemented by one or more jpg images of the source item. The image surrogates use all of the same metadata elements as the text elements; the primary difference between the two surrogates being that the content of the "text" element is a thumbnail of the jpg image surrogate. I note that this element is still called "text."

The J. Brown/B.B. Stutler Collection Database uses a static display page with uniform display elements that appear on every page. In cases where a record is missing some particular metadata element, the display element for that metadata appears but is followed by no content. The uri's for each surrogate are also very uniform, and despite the fact that there is no index for the collection, each page possesses a navigation widget that allows users to skip from the first surrogate in the collection to the last surrogate in the collection.

Based on observation, I concluded that the persistent urls are numerically sequential and constructed the experimental harvester accordingly. The size of the collection is also easily established as the text uri's range from http://www.wvculture.org/HiStory/wvmemory/jbdetail.aspx?Type=Text&Id=1 to http://www.wvculture.org/HiStory/wvmemory/jbdetail.aspx?Type=Text&Id=5062 and the image uri's range from http://www.wvculture.org/HiStory/wvmemory/jbdetail.aspx?Type=Photo&Id=1 to http://www.wvculture.org/HiStory/wvmemory/jbdetail.aspx?Type=Photo&Id=1599. The collection contains a total of 6661 surrogates, which is slightly less than 1/3rd of the 20,000 purported in the DCC's collection record. An explanation for this discrepancy may be that many of the text items also have multiple jpeg files attached to them (digital scans of the document's pages).

Table 5.1: J. Brown/B.B. Stutler to DC Crosswalk

Source Display Element	Dublin Core Element
Title	dc:title
Description	dc:description
Biographical or Historical Notes	dc:description>Notes:…<
*[Hardcoding: West Virginia Archives &amp; History]	dc:publisher
Date	dc:date
Media Format	dc:type

Table 5.1 (cont.)

Source Display Element	Dublin Core Element
*[Hardcoding: Text or Photo]	dc:type
*[Hardcoding: text/html or image/jpg]	dc:format
I.D. Number	dc:identifier>WV Memory Number:…<
[Hardcoding: surrogate uri]	dc:identifier
*[Hardcoding: isPartOf John Brown/Boyd B. Stutler Collection Database]	dc:relation
Location	dc:coverage
*[Hardcoding: http://www.wvculture.org/history/findinginformation.html]	dc:rights

As I stated above, the true challenge is in developing a crosswalk that adequately captures the metadata being presented to the J. Brown/B.B. Stutler Collection's users. The end goal is for the records harvested by these experiments to be used within the DCC aggregate, so the xml files are constructed to mimic the xml files that appear in the aggregate's original data folders, ready for ingestion into the aggregate's SQL database. For this initial experiment the metadata is mapped into simple Dublin Core format (see Table 5.1).

As can be seen in the table above, the short records are enhanced with administrative and object classification information that will be useful when comparing the records with other item-level records in the DCC aggregate. Specifically, information about the publisher, generic item type, item format, unique uri identifier, collection relationship, and rights information is added to each item-level record as it is re-constructed by the harvester.

5.2.1.b. Maine Memory Network

Like the J. Brown/B.B. Stutler Collection Database, the items in the Maine Memory Network are complex surrogates. Also like the J. Brown/B.B. Stutler Collection, the Main Memory Network's surrogates all have a unique, numerical uri. Unfortunately, while it is easy to physically count the number of potential items for harvest (there are 19,011 according to counts of the browse by item type indexes) the uri item id numbers do not correspond to a sequential scheme. This necessitated engineering a routine that could distinguish pages with content from those without.

The experimental harvester was programmed to crawl across a series of 42,000 uri's, starting with http://www.mainememory.net/bin/Detail?ln=1. The experimental harvester successfully harvested items with id numbers ranging from 72 to 33,627 and produced 25,209 unique xml files. This discrepancy in quantities of records may be due to the harvest capturing new records that had not yet been indexed or for which large amounts of metadata assertions cannot be made. I note that both the expected and actual counts vary from collection-level record in the DCC which records the collection size as 10,000 items.

Once again the crosswalk is the key feature. Unfortunately, the metadata displayed to Maine Memory Network's users is not as uniform as the J. Brown/B.B. Stutler Collection. It does generally breakdown into two broadly general interface views: well documented surrogates and poorly documented surrogates that may, in fact, be simple surrogates.

A typical, well-documented surrogate in the Main Memory Network collection displays the following metadata fields: Title, Contributor, Description, Creator, Creation Date, Subject Date, Town, Local Name, County, State, Media, Dimensions, Local Code, Collection, Object Type, LC Subject Headings, and Keywords. A poorly documented Main Memory Network surrogate often consists of only Title and Contributor. At the request of my project advisor, I altered the harvester code to create Qualified Dublin Core records.

Table 5.2: Maine Memory Network to qDC Crosswalk

Source Display Element	Qualified Dublin Core Element
Title	dc:title
Creator	dc:creator
LC Subject Headings	(multiple) dc:subject
Description	dc:description
*[Hardcoding: Maine Historical Society]	dc:publisher
Creation Date	dcterms:created
Media	dc:type
*[Hardcoding a format value based on object type value]	dc:format
*[Hardcoding uri]	dc:identifier
*[Hardcoding uri end integer]	dc:identifier >Maine Historical Society Item Number:…<
Local Code	dc:identifier>Local Idenitifier:…<
Collection	dcterms:isPartOf

Table 5.2 (cont.)

Source Display Element	Qualified Dublin Core Element
Subject Date	dcterms:temporal
State + (Town or City) + County	dcterms:spatial
*[Hardcoding: http://www.mainememory.net/aboutus/]	dc:rights

5.2.1.c. Quilt Index

Quilt Index presents a much greater challenge for harvesting than the previous two test cases as it does not use uri's with sequential or semi-sequential numerical numbering scheme. Quilt Index does have a complete item index, spread across a series of 5067 html documents. I compensate for this by adding a pre-harvester or harvester target indexer. The indexer first harvests the unique uri's listed in the item index and constructs a large text file which the item harvester uses to visit each item-level html page.

Both the Quilt Index's index page and the DCC collection level record state that the size of this collection is 50,669; however, during the harvest a number of errors occurring within the Quilt Index database result in the harvest only creating 45,452 unique item-level records. Each of these errors indicated that the relevant uri harvested from Quilt Index's index page no longer referenced a metadata record in Quilt Index's database. Also, unlike the previous two examples, the metadata used in Quilt Index pages are very specific to describing quilts. Both basic and full record display options are available to users. Quilt Index also provides documentation of its metadata categories that are used to describe the quilts (Quilt Index Core Fields, 2009; Quilt Index Comprehensive Fields Final, 2009).

The metadata documentation is very helpful for making crosswalk decisions. Unfortunately, Qualified Dublin Core lacks much of the expressiveness of the metadata standard that Quilt Index and its collaborators have created. As the basic record seems sufficient to enable the retrieval of the Quilt Index's surrogates within the DCC's repository, I decide to map the Quilt Index core fields into qualified DC (Table 5.3).

Table 5.3: Quilt Index core fields to qDC Crosswalk

Source Display Element	Qualified Dublin Core Element
Title	dc:title
Quilter	dc:creator
Top by	dc:creator
Quilted by	dc:creator
Subject	dc:subject
Pattern Names	dc:description>Pattern Names:…<
Construction	dc:description>Construction Technique:…<
Quilting Techniques	dc:description>Quilting Techniques:…<
Layout Format	dc:description>Layout Format:…<
Purpose or Function	dc:description>Purpose or Function:…<
Colors	dc:description>Colors:…<
Inscription	dc:description>Inscription:…<
History	dc:description>Historical Notes:…<
Other Notes	dc:description>Notes:…<
*[Hardcoding: Quilt Index]	dc:publisher
Others	dc:contributor
Date	dcterms:created
*[Hardcoding: Quilt]	dc:type
Type of Quilt Object	dc:type
*[Hardcoding: image/jpg]	dc:format
Quilt Size	dcterms:extent
Fabrics	dcterms:medium
*[Hardcoding: source uri]	dc:identifier
*[Hardcoding: id parsed from uri]	dc:identifier> Quilt Index Item Number:…<
Institutional Inventory	dc:identifier> Inventory Number:…<
Brackman Number	dc:identifier>Brackman Number:…<
Project Name	dcterms:isPartOf
Period	dcterms:temporal
Location Made	dcterms:spatial
*[Hardcoding: http://www.quiltindex.org/about.php#copyright]	dc:rights
Owner	dc:rights>Owner:…<

As can be seen on Table 5.3, I have attempted to preserve as much of the unique information contained in the Quilt Index's short records as possible by repeated use of dc:description. An attempt at preserving the context of the unique metadata fields has also been made by adding a level of unique markup in the form of semi-standard text labels within many of the DC elements.

5.2.1.d. Virtual Motor City

Like the Quilt Index, the Virtual Motor City collection uses an arcane identification scheme for its items. It also provides a full item index. Unlike the other collections, the item-level records for the Virtual Motor City's surrogates are quite heterogeneous; however, the collection does have a mapping table (http://dlxs.lib.wayne.edu/v/vmc/vmc-config.html) that suggests the types and kinds of metadata elements that each surrogate is likely to have.

The Virtual Motor City collection also presents a relatively novel browsing interface to users. It presents an html page with 20 thumbnails of items in the collection. Item records are dynamically displayed on the left as the user clicks on each thumbnail. The records themselves are embedded within the html of each page of the index, along with a unique uri for each surrogate.

I built the mapping (Table 5.4) based on observations of the embedded records. The collection lists two quantities describing its size (36,783 online images/media and 100,211 records). As images also appear in the larger quantity it seems likely that the collection consists of 36,783 complex surrogates and 63,428 traditional surrogates. The harvester successfully created 36,783 unique xml files, one for each of the complex surrogates in the collection.

Table 5.4: Virtual Motor City display fields to qDC Crosswalk

Source Display Element	Qualified Dublin Core Element
Title	dc:title
Historical Title	dc:title
Photographer	dc:creator
LC Subjects	dc:subject
Description	dc:description
Notes	dc:description>Notes:…<
Donor	dc:contributor
Date	dcterms:created
*[Hardcoding: Photograph]	dc:type
*[Hardcoding: image/jpg]	dc:format
Film Size	dcterms:extent
*[Hardcoding: parsed uri]	dc:identifier
Record ID	dc:identifier>VMC Record Number:…<
Decade	dcterms:temporal
Rights	dc:rights

One of the most noticeable things about VMC (Virtual Motor City) records is the lack of type and format information. The VMC collection records clearly leverage the human factor in IR system/user interactions. It is clear to users from the thumbnail that the source item for the surrogate is a photograph (derived from negatives or otherwise) and that the format of the surrogate is a jpg file. As both item type and item format are used both to describe the nature of the DCC aggregate's contents and for analysis, I have enriched the records by adding type and format information.

5.2.2. STANDARDIZED (OR HOMOGENEOUS) COLLECTION WEBSITE ORGANIZATION SCHEMA

5.2.2.a. CONTENTdm Collections

Initial experiments in harvesting proved that CONTENTdm collections would need to be indexed to build a list of target item-level uri's. All three of the initial targets were successfully indexed. Unfortunately, errors made during the coding process led to difficulties with two of the harvests. These have been since resolved. Overall, harvesting CONTENTdm via screen scraping is generally as trivial as harvesting from heterogeneous interfaces. Once again it is the metadata crosswalking decisions which are the key to the success or failure of the process.

A new dimension of difficulty is that CONTENTdm is often used for super-collections. Each collection within the super-collection presents a heterogeneous mixture of records and surrogates. Some collections may use more expressive metadata standards to display richer information to users. The details of these metadata crosswalks appear in the Appendices (see Appendices E, F, G, and H). In the case of the Upper Mississippi Valley Digital Image Archive, an established, generalized metadata mapping was published (http://www.umvphotoarchive.org/umvdia/about.html) which greatly facilitated crosswalk construction, as only one crosswalk was needed for all of its sub-collections.

The harvesting experiment successfully harvested the sub-collections of the entire Upper Mississippi Valley Digital Image Archive super-collection. This produced 8,090 unique xml files for 10 collections. Arguably one of the sub-collections is not a collection at all as it only has one "test" item. Due to initial problems with the remaining two initial targets, specific sub-collections from three other super-collections using CONTENTdm were targeted for website HTML parsing experiments. The Arther, Once Upon a Time sub-collection (488 items) was successfully harvested from the Illinois Digital Archives super-collection. The Wing Luke Asian Museum sub-collection (158 items) was successfully harvested from the King County Snapshots super-collection. Like the Upper Mississippi Valley Digital Image Archive, data dictionaries (http://content.lib.washington.edu/imls/kcsnapshots/tips-data.html) published by the King County Snapshots project greatly facilitated the construction of the crosswalk for the Wing Luke collection.

Attempts to harvest the Long Island Memories super-collection all failed as the scripts being used continuously timed out before the sub-collections could be indexed. These failures are probably due to the primitive (outmoded even) code being used for the experiments. Later experiments with one of the previous targets (TIDES) resulted in the successful harvest of the Cason Monk-Metcalf Funeral Directors sub-collection (15 items) from the TIDES super-collection.

5.2.2.b. PastPerfect-Online Collections

With the encouragement of my project advisor, I expanded experimentation, applying brute force harvesting techniques to PastPerfect-Online, a standard software package beginning to gain use in the digital museum community. PastPerfect-Online immediately has at least one advantage when employing the website HTML parsing method of harvesting; it is designed to be indexed by Google (PastPerfect-Online, 2008). Each PastPerfect-Online collection serves out an xml list of all of the unique item-level uri's within the collection.

Like CONTENTdm, metadata records are heterogeneous with respect to the kinds and types of metadata presented to the user. Relatively generic crosswalks based on observed terminology were developed for the harvests (see Appendices I, J, & K). 50 records were harvested from each of the PastPerfect-Online collections. All three of the test cases used relatively uniform nomenclature for metadata fields displayed to users. Capturing specific and deeper information, especially dates, can be challenging and requires building a significant body of samples to ensure that the crosswalk used by the harvester is robust enough to capture all of the desired metadata.

5.2.3. CUNEIFORM DIGITAL LIBRARY AND OTHER UNUSED TEST CASES

The Cuneiform Digital Library was a unique collection in comparison to the other collections in the harvesting experiment candidate pool and the DCC aggregate as a whole. It represented a novel collection of artifacts that are not represented by other collections in the aggregate, and it also provided a copy of its catalog database available for download. Ideally, harvesting the collection should be as simple as downloading the zipped database file and mapping its contents into Dublin Core xml files. Unfortunately, the library only has one copy of the Filemaker Pro database software that the Cuneiform Digital Library uses to hold its catalog. As insufficient time was available to use this software, only an xml dump of the database could be acquired.

Even with an xml dump of the original catalog, it should still be possible to build an xslt that will produce a series of Dublin Core xml files. Unfortunately the xml dump produced a file so large that (even when broken down into 6 parts) it proved intractable to open, read, and modify. The Cuneiform Digital Library's user interface is such that it should be possible to use brute force harvesting to iterate across its contents and build Dublin Core records; however, due to the unique nature of this collection's contents and project time constraints it was decided not to proceed with any harvesting experiments.

Additionally, four other heterogeneous collections that had been initially selected for harvesting experiments were also not harvested due to time constraints. It is probable that one of them (Plant Images at Missouri Botanical Garden) contains surrogates that are too specialized in nature to merit harvesting. I was also unable to perform the types of analyses that others (Weagley et. al., 2010; Han et. al., 2009; Jackson et. al., 2008; Jackson, 2006) have performed on metadata records within the DCC aggregate. The failure to perform a quality analysis is partially due to major renovations to the DCC's database structures, including a complete re-harvest of its collections, and partially due to my own failings to implement an alternate database structure in a timely manner.

5.3. DISCUSSION

Website HTML parsing has proven to be successful in as far as it can reliably access collection websites, parse the contents of records displayed to users, and produce xml files from that data. While it is impossible to say with any authority that xml files created by the experimental harvester are high quality metadata records, the crosswalks used for each harvest should be a good indicator of the quantity of metadata contained in each record along with the potential quality of the record (assuming the content within each record is internally consistent).

With website HTML parsing of CONTENTdm collections, there is also a greater incentive for aggregators to decompose super-collections into their component sub-collections. Decomposition of super-collections into their child collections entails the development of multiple crosswalks, which increases the amount of preparation time for each harvest. The benefit of this crosswalk development is that the aggregator can preserve contexts that are sometimes lost when local fields are not mapped into Dublin Core for serialization by OAI-PMH (Han et. al., 2009).

Similarly, context is lost when collection-level records are not created for the sub-collections within super-collections. This collection-level context is further damaged when items from these sub-collections are associated only with the parent super-collections whose actual items are more properly the sub-collections themselves. The sub-collections at greatest risk of contextual loss in this way are the thematic research collections. Every piece of a thematic research collection has been grouped into the collection either because it is a primary source or because it is a secondary source that supports a primary source already in the collection. When the items in the collection are disassociated from one another the contextual information gained simply by their mutual association is lost.

CHAPTER 6

ACCESS VIA BROADCAST METASEARCH

The initial plan was to investigate, compare and contrast 3 emerging options for broadcast metasearch services:

- integrated search results
- dark target background searches
- supplemental resources

Each of these three options applies the outcome of a broadcast metasearch in a different manner.

Integrating the results of a broadcast metasearch directly into the aggregator's search results is the most straightforward of these three options. One imagines that this is the type of "federated searching" that David Sterns is referring to in his 2009 Online article on harvesting item-level metadata as a method to mine large-scale databases. Under the integrated search results option, a query is run against interfaces of several of the collections' for which the DCC has collection-level metadata records. The results are aggregated, sorted, and combined with the DCC's search results.

The use of dark targets as a broadcast metasearch method has been pioneered by several metasearch grants at the University's Grainger Engineering Library. Under the guidance of the grants' principal investigator, Bill Mischo, the grant staff at Grainger added unique search functionality to the UIUC Library's Easy Search database search engine. When a user enters a query into the Easy Search IR portal the Easy Search system not only runs the query against a standard set of database targets to retrieve results but also runs the query against a small set of additional database targets, "dark targets." When the IR system's logics determine that the results retrieved from the standard set of database targets is too small then, it displays the results of the dark target search to the user.

The final broadcast metasearch option, accessing supplemental resources is directly based on the standard functionality of the UIUC Library's Easy Search broadcast metasearch system. As mentioned, Easy Search takes a user's search query and runs it against a standard set of databases. It then collects the quantity of results each database produces from the user's query and returns those numbers to the user as links to the databases' results. Using broadcast metasearch to access supplemental resources works in the same manner except that the result links are displayed to the user beside the regular IR system results. Some experimentation with this method has already been carried out on an experimental version of the DCC's sister site's (Opening History) IR portal.

Unfortunately due to time and technology constraints, only the integrated search results experiment was completed. The dark target federated technique is used to increase search result quantities when normal search results return few or no results. For the specific case of the DCC aggregate, the dark target technique was intended to investigate the benefits of providing access to topically specialized collections for which little or no item-level metadata has been or can be harvested (e.g. PlantCollections™). The supplemental resource technique is best used for those resources that are undesirable for harvest but which are still likely to provide resources that are relevant or even pertinent to query string provided by the user. The use case that was intended to be explored was access to newspaper archives as a supplemental resource.

6.1. METHODS

Eight target collections out of the 25 collections in the sample were initially selected for inclusion in the proof-of-concept broadcast metasearch experiments. Of these eight, it quickly became clear that one of them, History at our Hands, was going to be too difficult to implement within the timeframe of the experiments. As the collection's IR interface is built on proprietary software, it presents some significant software engineering challenges.

The experimental pool was further reduced due to lessons learned from the harvesting experiments. When I initially surveyed the Dallas Museum of Art Collections, it was not clear to me how to go about harvesting the items in the collections; therefore, I added it to the candidate list for broadcast metasearch experiments. Since that time many successful harvesting experiments have been carried out. Experiences in building applications with which to index a collection's website have now made it quite clear how to go about harvesting the collections at the Dallas Museum of Art, which are quite appropriate for inclusion in the DCC's item repository. The Dallas Museum of Art website is simply a larger, slightly more complex indexing problem than the initial harvesting experiment candidates. As such, it seems superfluous to include the Dallas Museum of Art in the broadcast metasearch experiments.

The remaining six collections were divided between the three techniques. As both the Brooklyn Daily Eagle Online and Colorado's Historic Newspaper Collection are newspaper archives, they were to be the subjects for the supplemental resource federated search experiment. PlantCollections™, was to be the subject for the dark target experiment. The final three remaining target collections (Beyond the Shelf, Exploratorium Digital Asset Management Collection, and Florida Folklife Collection) were successfully used as the subjects of the integrated search results experiment.

VBScript was used to construct relatively simple applications for the broadcast metasearch experiments. Algorithms supplied by Josh Bishoff, a federated search services researcher at Grainger Engineering Library, were used to calculate the cosine and Jaccard coefficient similarity measurements that, in turn, were used to score each title against the query string as a pseudo-relevance ranking measure for the search results.

Cosine and Jaccard coefficient similarity measurements are basic methods of using vector spaces to score documents against each other (Manning, Raghavan, & Schütze, 2008; Paijmans, 2002). For this experiment, only the "document" titles were scored.

The experiment was carried out using two ASP programs. One program was used to aggregate and sort the search results from the three targets and, one was used to proxy the DCC website and to integrate the sorted, aggregated results into the regular DCC results. To increase the ease of search result integration the sorter program also performed a search in the DCC aggregate.

As the goal of the experiment was only to prove that concept of integrating search results from both the aggregate's IR portal and a broadcast metasearch application was possible, only up to the first three search results were taken from each of the three experimental collections' results pages. The first three results from the DCC page were also ingested into the sorter. They were then ranked using an average of the sum of their cosine and Jaccard coefficient measurements. Finally they were sorted by rank using a primitive bubble sort algorithm and written out to the web. From there, the proxy page reads the results and writes them over the first three results on the proxied DCC page.

6.2. RESULTS

The results of the integrated search results experiment were decidedly mixed. While the crude bubble sorting algorithm worked reliably, the value of relevance ranking needs to be explored further. As a full examination of results sorting technology was beyond the scope of this project, the ordering of the results was not deeply scrutinized. A much bigger stumbling block proved to be the overall speed of the application which was completely dependent on the relatively slow response times of the DCC page, both for the sorter and for the proxy page. The use of broadcast metasearch in and of itself did not appear to be the weak link, slowing down the production of the results page to the user. Some of the response speed issues are likely due to the primitive nature of the code used and also due to time losses from asp’s compile on the fly nature. As previously stated the dark target and supplemental resource experiments were not carried out due to a lack of time.

6.3. DISCUSSION

In addition to the questionable results of the relevance ranking algorithm, the appropriateness of the combination of collections was also somewhat questionable. The DCC and Kentuckiana are both aggregators that provide users very diverse types of information objects. In contrast to this, the Florida Folklife Collection is primarily a digitized photograph collection. Its contents are homogenous with regards to item type and it displays very little information to users unless they click features such as “details.” The Exploratorium presents yet a third, radically different collection type. It is a collection a museum exhibits and, unlike the other collection and the aggregates, its primary audiences are primary and secondary school students and teachers. Like Florida Folklife it provides users with very little metadata at the level of its search results page.

Despite parsing the results out of their native html wrappers and building new DCC-style wrappers around them, it is often clear that non-DCC resources within the results list differ substantially from the DCC's resources. Of the three experimental subjects, only the Kentuckiana aggregate's results are able to provide the same level of information as DCC results.

Results from Florida Folklife usually have creator metadata so the results are very similar to DCC results, but they also typically include date information. Unfortunately date information cannot be effectively inserted into the DCC's display space as there is no case for it in the DCC site's CSS. Attempts to insert the date metadata frequently led to display problems with the proxy site. In contrast to this, results from the Exploratorium lack creator and date metadata, but instead provide free text descriptions of the resources. I was also unable to effectively insert this description metadata into the DCC display space as it caused even worse display problems than the Florida Folklife dates.

Because the DCC results page is designed for the display of specific types of metadata, large amounts of context are lost when integrating results from websites that conform to different display standards. This context loss is directly analogous to the types of context losses seen by Han et al. in their 2009 paper discussing the loss of context from the local use of unique fields in CONTENTdm collections.

CHAPTER 7

CONCLUSIONS & RECOMMENDATIONS

The complex nature of aggregating heterogeneous resources is a major obstacle that OAI-PMH has attempted to rectify. As has been seen (Shreeves, 2005), its adoption rate has been and remains low among the population of DCC data providers. The use of supplemental harvesting methods has the potential to provide a powerful set of tools for aggregation collection developers.

7.1 CONCLUSIONS

The supplementation of OAI-PMH can facilitate the growth of an aggregator's item-level metadata repository. There are several prices for this increase in growth rate. While some contexts that are often lost with OAI-PMH can be preserved, it is possible that other contextual information can be lost using the supplementary method. Overall, context loss due to metadata interoperability issues remains a problem. Additionally, the supplementary method requires that the aggregator make larger investments of staff time. On the positive side, the aggregator can clearly increase the pace at which new collections of item-level metadata are ingested into the item-level metadata repository as they no longer have to wait for the data provider to deploy OAI-PMH.

Whereas traditional views of the interoperability of OAI-PMH metadata place many of the burdens for assessing what metadata to include for harvesting on the shoulders of the data providers (Beisler & Willis, 2009), the supplementary approach moves much of this burden to the aggregator. While this can sometimes aid in the preservation of contextual information that is sometimes lost when data providers fail to map unique local fields to a transmission standard such as OAI Dublin Core, loss of context is still a present danger when building metadata crosswalks. The supplementary method can also be vulnerable to context loss because the staff preparing the metadata crosswalks will not be as intimately familiar with the source metadata as the data provider.

Supplementary harvesting can permit the inclusion of unique collections into an aggregate; however, harvest developers must take special care in mapping unique types of data in order to maximize the interoperability of the records and avoid context loss. In one particular experimental case, the Quilt Index, large amounts of data specific to the description and retrieval of quilts had to be repeatedly mapped to Dublin Core's description element. For example:

Table 7.1: Using qualifiers within Dublin Core elements

Source Element	Qualifier
Pattern Names	dc:description>Pattern Names:…<
Construction	dc:description>Construction Technique:…<
Quilting Techniques	dc:description>Quilting Techniques:…<
Layout Format	dc:description>Layout Format:…<
Purpose or Function	dc:description>Purpose or Function:…<
Colors	dc:description>Colors:…<
Inscription	dc:description>Inscription:…<
History	dc:description>Historical Notes:…<
Other Notes	dc:description>Notes:…<

In each case qualifying language was added to the Dublin Core description element in order to preserve the context of the information being mapped into that element. While the use of qualifying terminology within the Dublin Core elements remains controversial, quilts turned out to be an excellent case demonstrating how different the language necessary to describe them and aid in their automated retrieval is from more traditional library-oriented items, such as monographs or photographs. In a large-scale cultural heritage aggregation like the DCC, the addition of culturally significant but unique items could provide an interesting set of additional resources for historians and similar cultural heritage scholars.

Further, adoption of supplementary methods of harvesting means that aggregators are no longer left waiting for data providers to serialize their metadata, like beggars waiting for a handout. Rather, aggregators can take a much more active role in building their aggregation; they can go forth like hunter-gatherers and collect the fruits of the collections which best benefit their users. The aggregation service provider is left in a position to better articulate the role that the aggregation plays in the larger community of data and service providers and create value-added services and content for its users.

The price that aggregators must pay for this increase in both the flexibility and power in item-level harvesting is the increased need for the human and technological resources that make supplementing OAI-PMH possible. Aggregators that harvest metadata via OAI-PMH already normalize the heterogeneous metadata received during the harvesting process (Shreeves, et. al, 2003; Arlitsch & Jonsson, 2005). As a result, some structures for standardization and enrichment of metadata already exist; however, in order to supplement OAI-PMH additional standardization structures need to be implemented. In addition, aggregators have an opportunity to make some hard choices about metadata normalization that may result in loss of context but improve users' odds of locating the harvested items.

For instance, when a collection known to consist solely of photographs is being harvested, then the aggregator may wish to simply normalize all of the records as dc:type = photograph and dc:format = image/jpeg or apply similar standardized vocabulary. Application of standardized vocabulary can make it easier for the aggregator to articulate to the user exactly what is within the aggregate's item-level metadata repository, but a user looking specifically for photographs produced by the calotype process would lose out since this type of information was probably lost during harvest due to normalization decisions made prior to harvesting.

Finally, if item-level metadata cannot be harvested from a collection, alternate methods, such as broadcast metasearch may provide a last ditch method of providing item-level access to aggregate users. Again, while there is some potential for this final type of federated search functionality, the loss of context that occurs when results are interpreted from one display standard to another is quite troubling. This method of item level access is not as strong as actually harvesting the item-level display metadata, which can preserve and even enrich source metadata. While still untested, it is hoped that the use of other broadcast metasearch options, such as dark target background searching and accessing supplemental resources will prove to be beneficial for topically specific collections or collections with difficult to harvest compound objects, such as digitized newspaper archives.

7.2 RECOMMENDATIONS

In closing I have three primary recommendations for improving the DCC aggregate.

1. Design and implement additional metadata quality control measures to help insure that collection-level metadata is of the highest quality possible.
2. Continue experimentation with harvesting methods to supplement OAI-PMH in order to better determine the exact nature of the costs and benefits of using supplemental harvesting methods.
3. Expand experimentation with broadcast metasearch options to assess if and how they can enrich the experience of the average DCC user.

My primary reasoning for the first recommendation is the great difficulty I had in accurately calculating the quantitative benefits (in terms of numbers of item-level metadata records harvested) of supplementing OAI-PMH. There is a need to rectify some of the collection records so that they more accurately describe the collections they are representing. A clearer, more hierarchical view of the collections, super-collections, aggregations, and other websites in the DCC's collection-level registry would have substantially aided in the selection of candidates for these experiments.

Because OAI-PMH has experienced such low adoption among the collections within the collection-level registry and because there is no trending data to indicate that the adoption rate is changing, OAI-PMH should be supplemented with additional methods to harvest item-level metadata into the aggregate's item-level metadata repository. Recall that 23.8% of all of the collections within the aggregate are part of a single super-collection, the Arizona Memory Project. Harvesting just this one super-collection would vastly increase the numbers of item-level resources directly available to the aggregate's users.

Finally, supplementation of item-level access through the use of broadcast metasearch access to newspaper archives would also increase the options available to aggregate users without unduly increasing the computational and storage burdens already experienced by the DCC's technical infrastructure. Exploring the use of this type of access for some of the older, more topically specific collection-level resources within the aggregate would also be worth exploring and could aid in some of the experimentation currently occurring with topically specific portals.

REFERENCES

- Arlitsch, Kenning, and Jeff Jonsson. "Aggregating Distributed Digital Collections in the Mountain West Digital Library with the CONTENTdm™ Multi-site Server." *Library Hi Tech* 23.2 (2005): 220-32. Web. 17 August 2010. <www.emeraldinsight.com/journals.htm?articleid=1509054&show=pdf >.
- Beisler, Amalia, and Glee Willis. "Beyond Theory: Preparing Dublin Core Metadata for OAI-PMH Harvesting." *Journal of Library Metadata* 9.1/2 (2009): 65-97. Web. 18 August 2010. <http://www.informaworld.com/smpp/section?content=a914010918&fulltext=713240928>.
- Benevento, Jenny. *Template for Collection Registry Information*. Tech. Urbana: Grainger Engineering Library, 2005. Web. 30 July 2010. <http://imlsdcc.grainger.uiuc.edu/3yearreport/docs/colltemplatewpjvb.pdf >.
- Brogan, Martha L. "A Survey of Digital Aggregation Services [2003]." *Digital Library Federation*. 1 July 2010. Web. 13 Sept. 2010. <http://www.diglib.org/pubs/dlf101/dlf101.htm>.
- Cole, Timothy W., and Sarah L. Shreeves. "Search and Discovery across Collections: the IMLS Digital Collections and Content Project." *Library Hi Tech* 22.3 (2004): 307-22. Print.
- "Data Dictionaries - King County Snapshots." *UW Libraries Digital Collections*. Web. 22 Aug. 2010. <http://content.lib.washington.edu/imls/kcsnapshots/tips-data.html>.
- Dunsire, Gordon. "Collecting Metadata from Institutional Repositories." *OCLC Systems & Services: International Digital Library Perspectives* 24.1 (2008): 51-58. Web. 18 August 2010. <www.emeraldinsight.com/journals.htm?articleid=1674218&show=pdf>.
- Han, Myung-Ja, Christine Cho, Timothy W. Cole, and Amy S. Jackson. "Metadata for Special Collections in CONTENTdm: How to Improve Interoperability of Unique Fields Through OAI-PMH." *Journal of Library Metadata* 9.3 (2009): 213-38. Print.
- Hider, Philip. "Australian Digital Collections: Metadata Standards and Interoperability." *Australian Academic Research Libraries* 35.4 (2004): 1-7. Print.

- Hillmann, Diane I. "Metadata Quality: From Evaluation to Augmentation." *Cataloging & Classification Quarterly* 46.1 (2008): 65-80. Web. 20 August 2010. <http://ecommons.cornell.edu/bitstream/1813/7899/1/Metadata_Quality_rev.pdf >.
- Jackson, Amy. *Preliminary Analysis of Item-level Metadata Harvested*. Tech. Urbana: Grainger Engineering Library, 2006. Web. 30 July 2010. <http://www.ideals.illinois.edu/bitstream/handle/2142/720/Item-levelmetadata.pdf?sequence=2>.
- Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. "Dublin Core Metadata Harvested Through OAI-PMH." *Journal of Library Metadata* 8.1 (2008): 5-21. Print.
- Johnston, Pete, and Bridget Robinson. *Collections and Collection Description*. Issue brief no. 1. Bath: UKOLN, 2002. Web. 30 July 2010. <http://www.ukoln.ac.uk/cd-focus/briefings/bp1/bp1.pdf>.
- Lagoze, Carl (2004). The relationship between OAI-PMH and Dublin Core: Required, recommended or other? In *Proceedings CERN Workshop on Innovations in Scholarly Communication: Implementing the Benefits of OAI* (OAI3), CERN (Geneva, Switzerland), February 12-14, 2004. Web. 30 July 2010. <http://eprints.rclis.org/archive/00001000/>.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schu□tze. *Introduction to Information Retrieval*. New York: Cambridge UP, 2008. Print.
- Mischo, William H., and Mary C. Schlembach. "Metasearch Technologies in Reference Work, OAI, and Search Navigation Assistance." In *Proceedings of American Society for Engineering Education Annual Meeting 2004, Engineering Education Research*: New Heights, Salt Lake City. Print.
- Paijmans, Hans. "A Taxonomy of Vector-based IR Models (II): Similarity Functions." 2002. Web. 30 July 2010. <http://paai.uvt.nl/Paai/Onderw/V-I/Content/similarities.html>.
- Palmer, Carole, Ellen Knutson, Michael Twidale, and Oksana Zavalina. "Collection Definition in Federated Digital Resource Development." In *Proceedings of the ASIST Annual Meeting*. ASIS&T Annual Meeting - 2006, Texas, Austin. Vol. 43. 2006. Print.
- Palmer, Carole L. "Thematic Research Collections." 26 Nov. 2002. Web. 30 July 2010. <http://people.lis.illinois.edu/~clpalmer/bwell-eprint.pdf>.

- "Proposal for an IMLS Collection Registry and Metadata Repository." *IMLS Digital Collections and Content*. 2003. Web. 30 July 2010. <http://imlsdcc.grainger.uiuc.edu/grantproposal.asp>.
- "The Quilt Index Comprehensive Fields Final." Quilt Index, 16 Feb. 2009. Web. 30 July 2010. <http://www.quiltindex.org>.
- "Quilt Index Core Fields." Quilt Index, 2009. Web. 30 July 2010. <http://www.quiltindex.org>.
- Shreeves, Sarah L. *Barriers to Metadata Sharing via the OAI Protocol*. Tech. Urbana: University of Illinois at Urbana-Champaign, 2005. Web. 30 July 2010. <http://imlsdcc.grainger.uiuc.edu/3yearreport/docs/BarriersToInteroperability.pdf >.
- (a) Shreeves, Sarah L., Ellen M. Knutson, Besiki Stvilia, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole. "Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections." *Proceedings of Association of College and Research Libraries (ACRL) 12th National Conference*. Association of College and Research Libraries (ACRL) 12th National Conference, Minneapolis, MN. 2005. 223-37. Web. 18 August 2010. <https://www.ideals.illinois.edu/bitstream/handle/2142/145/shreeves05.pdf?sequence=2>.
- Shreeves, Sarah L., Thomas G. Habing, Kat Hagedorn, and Jeffery A. Young. "Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting." *Library Trends* 53.4 (2005): 576-89. Web. 20 August 2010. <http://deepblue.lib.umich.edu/bitstream/2027.42/59513/1/Shreevesetal_576-589_LT_53_4.pdf>.
- Shreeves, Sarah L., Joanne S. Kaczmarek, and Timothy W. Cole. "Harvesting Cultural Heritage Metadata Using the OAI Protocol." *Library Hi Tech* 21.2 (2003): 159-69. Web. 18 August 2010. <http://www.emeraldinsight.com/journals.htm?articleid=861365&show=html>.
- Stern, David. “Harvesting: Power and Opportunities Beyond Federated Search.” *Online* 33.4 (2009): 35-7. Web. 12/9/2010 <http://web.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=2&hid=15&sid=2323c2b9-8882-45a9-96ea-ced10cf04550%40sessionmgr12>.

- "Upper Mississippi Valley - About." *Upper Mississippi Valley Digital Image Archive.* Web. 22 Aug. 2010. <http://www.umvphotoarchive.org/umvdia/about.html>.
- Ward, Jewel. "A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative." *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries.* Houston, TX. 2003. Web. 13 Sept. 2010. < http://delivery.acm.org/10.1145/830000/827196/p315-ward.pdf?key1=827196&key2=1638044821&coll=GUIDE&dl=GUIDE&CFID=101688966&CFTOKEN=86285728 >.
- Weagley, Julie, Ellen Gelches, and Jung-ran Park. "Interoperability and Metadata Quality in Digital Video Repositories: A Study of Dublin Core." *Journal of Library Metadata* 10.1 (2010): 37-57. Print.

ACKNOWLEDGMENTS

I would like to thank the members of my committee, Tim Cole, Carole Palmer, and Jerome McDonough, without whose advice this project would not have been possible. Thanks to William Mischo for his keen insights into broadcast metasearch and the uses of HTML parsing. A big thank you to Josh Bishoff for providing the Jacard and Cosine ranking algorithms for my broadcast metasearch experiments. I would also like to thank DCC project coordinator Katrina Fenlon and team member Mike Tang for feedback on harvester output.

Appendix A: Compiled list of collections for which "size" is unknown and no item-level metadata has been harvested.

100 Years of Oklahoma Governors
Abraham Lincoln Historical Digitization Project
Ada Lois Sipuel v. Board of Regents University of Oklahoma, 1948-
ArtsConnectEd II
Audio-Video Barn
Boston Streets: Mapping Directory Data
Building A Globally Distributed Historical Sheet Map Set
Centennial Exhibition Digital Collection
Chicago Botanic Garden Plant Evaluation Website
Civil Rights Digital Library
Civil Rights in Mississippi Digital Archive
Civil War Newspapers Project
Clifford K. Berryman Collection
CLIOH: Cultural Digital Library
Cornell University Collection of Political Americana
Creating Communities
Cuban Heritage Digital Collections
Digital Dress: Historic Costume Collection$_1$
Digital Video Library Toolkit for Museums and Libraries with Limited Resources
Dorothea June Grossbart Historic Costume Collection
Exit Art Digital Archive
Fairchild Tropical Botanic Garden Palm Collection
Federal Publications about Oklahoma
Fenian Brotherhood Collection
Georgia Legislative Documents
Getting the Message Out: National Political Campaign Materials: 1840-1860
Hawaiian Language Newspapers
Henry Ford Historic Costume Collection
Hoagy Carmichael Collection
Illinois Alive! The Heritage and Texture of a Pivotal State During the First Century of Statehood (1818-1918)
Images of the Catholic Diocese of Tucson
Integrated Finding Aid to Walt Whitman Manuscripts
John Bloomfield Jervis Papers
Louisiana Gumbo: A Recipe for Empowerment
Making Sense of Modern Art
Mark Twain's Mississippi
Masterworks Online
Meadow Brook Hall Historic Costume Collection
Medieval manuscripts in the Digital Age: New Paths to Old Books in the Free Library of Philadelphia
Murder & Mayhem: The Strange Saga of Winnie Ruth Judd
New Jersey Digital Highway
Object of History
Oklahoma Authors
Oklahoma Crossroads
Oklahoma Image
Oklahoma Resources
Oklahoma State Government Publications
Olasee Davis Newspaper Articles
Oneida Community Collection in the Syracuse University Library
Our Americas Archive Partnership
Photograph Connoisseurship Resource
Photomuse
Publishers' Bindings Online, 1815-1930: The Art of Books (PBO)
Raid on Deerfield: The Many Stories of 1704
Revolutionary City: Developing a Virtual Reality Model of Williamsburg in 1776
Rochester Images
Rome-Turney Radiator Company Records Collection
Shuffle Along: The Eubie Blake Collection

Appendix A (cont.)

Sound Model: Collaborative Infrastructure for Digital Audio
Southeast Asia Visions
Territorial Kansas Online
Texas Heritage Online
TIMEA (Travelers in the Middle East Archive)
Timepieces$_3$
Timothy Vedder Letters
T-RACES: a Testbed for the Redlining Archives of California's Exclusionary Spaces
Tulsa Race Riot Documents
University of the Virgin Islands Research Reports and Occasional Papers
Virgin Islands Funeral Memorial Booklets
vPlants: A Virtual Herbarium of the Chicago Region
Walt Whitman Archive
WebERA
Western New York Suffragists: Winning the Vote
William Gedney: Photographs and Writings
Worklore: Brooklyn Workers Speak

Appendix B: IMLS DCC Complete Collection Survey

Collection	Class	NOTES
100 Years of Oklahoma Governors	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
1936 Gainesville tornado : disaster and recovery	Thematic Research Collection	
Abraham Lincoln Historical Digitization Project	Thematic Research Collection	
Ada Lois Sipuel v. Board of Regents University of Oklahoma, 1948-	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
Adjutants General of Arizona	Homogenous Object Collection	sub-collection of Arizona Memory Project
Africa Focus: Sights and Sounds of a Continent	Homogenous Object Collection	
AlabamaMosaic	Aggregation	contains 9 Super-Collections
American Journeys	Thematic Research Collection	
American Missionary Association and the Promise of a Multicultural America: 1839 - 1954	Homogenous Object Collection	sub-collection of LOUISiana Digital Library
American Natural Science in the First Half of the Nineteenth Century	Thematic Research Collection	
Arizona Archives Historic Photographs	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Attorney General Opinions	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Aviation History: The Ruth Reinhold Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Bushmasters	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona County and Local Publications	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Electronic Atlas	Thematic Research Collection	Unavailable
Arizona Executive Orders	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Landscapes Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Latinos in Public Service	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Military Museum Images	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Mines	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona State Agency Publications	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona State Archives - State, County and Local Government Records	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona Territorial Post Offices	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona-related Federal Publications	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona's Saints and Shady Ladies	Homogenous Object Collection	sub-collection of Arizona Memory Project
Arizona-Sonora Documents Online	Super-Collection	15 sub-collections -- part of the University of Arizone Digital Collections (22 total sub-collections)
Arthur, Once Upon a Time - Local History Images of Arthur, Illinois	Homogenous Object Collection	sub-collection of Illinois Digital Archive
ArtsConnectEd II	Collection	
ASU Science Pioneers: 1955 - 1970	Homogenous Object Collection	sub-collection of Arizona Memory Project
Atchison, Topeka & Santa Fe Railway Company (A.T.&S.F.Ry.Co.) Collection Highlights	Homogenous Object Collection	sub-collection of Arizona Memory Project
Atlanta History Center Album	Super-Collection	81 sub-collections
Audio-Video Barn	Thematic Research Collection	
Banana: A Chinese American Experience	Website	

Appendix B (cont.)

Collection	Class	NOTES
Basketry from the Pueblo Grande Museum	Homogenous Object Collection	sub-collection of Arizona Memory Project
Beauty in Stone : the industrial films of the Georgia Marble Company	Homogenous Object Collection	2 film collection -- not searchable
Beyond the Shelf: Serving Historic Kentuckiana Through Virtual Access	Homogenous Object Collection	
Black Swamp Memories	Homogenous Object Collection	
Blues, Black vaudeville, and the silver screen, 1912-1930s : selections from the records of Macon's Douglass Theatre	Thematic Research Collection	
Boston Streets: Mapping Directory Data	Thematic Research Collection	
Brooklyn Daily Eagle Online	Homogenous Object Collection	
Building A Globally Distributed Historical Sheet Map Set	IR Portal	
California Design Collection, 1955-1984	Homogenous Object Collection	sub-collection of the On-line Archive of California
California Ethnographic Field Photographs, 1900-1960	Homogenous Object Collection	sub-collection of the On-line Archive of California
Capturing Their Pasts Veteran Oral Histories	Thematic Research Collection	
Celebration of the Human Spirit: Japanese-American Relocation Camps in Arizona	Homogenous Object Collection	sub-collection of Arizona Memory Project
Centennial Exhibition Digital Collection	Thematic Research Collection	
Central Florida Memory	Super-Collection	15 sub-collections, plus 10 newspapers
Charles Overstreet Collection	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Charles W. Cushman Photograph Collection	Homogenous Object Collection	
Chicago Botanic Garden Plant Evaluation Website	Thematic Research Collection	
Chinese Paintings 12th century - 20th century	Homogenous Object Collection	sub-collection of the On-line Archive of California
City of Glendale Council Minutes of 1910-1914	Homogenous Object Collection	sub-collection of Arizona Memory Project
Civil Rights Digital Library	Aggregation	185 institutions
Civil Rights in Mississippi Digital Archive	Presumably a super-collection	Unavailable
Civil War Newspapers Project	Presumably a homogenous object coll.	Unavailable
Clifford K. Berryman Collection	Homogenous Object Collection	
CLIOH: Cultural Digital Library	Collection	
Coal Mining in Illinois, Machine vs. Man	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Cochise County Clerk of Superior Court - Bisbee Deportation Documents	Homogenous Object Collection	sub-collection of Arizona Memory Project
Cochise County Historical & Archeological Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Cochise County Territorial Court Documents	Homogenous Object Collection	sub-collection of Arizona Memory Project
Code City	Website	
Collection of Photographs by Carleton E. Watkins	Homogenous Object Collection	sub-collection of the On-line Archive of California
Colorado Plateau Digital Archives Selections	Homogenous Object Collection	sub-collection of Arizona Memory Project
Colorado's Historic Newspaper Collection	Homogenous Object Collection	
Columbia River Basin Ethnic History Archive	Thematic Research Collection	
Columbus Public Library Association minutes, 1881-1883	Homogenous Object Collection	
Congressman John Rhodes Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Connecticut History Online	Thematic Research Collection	

Appendix B (cont.)

Collection	Class	NOTES
Cornell University Collection of Political Americana	Homogenous Object Collection	sub-collection of Cornell University Collections
Creating Communities	Super-Collection	12 sub-collections
Cuban Heritage Digital Collections	Super-Collection	28 sub-collections
Cuneiform Digital Library	Super-Collection	17 sub-collections
Cylinder Preservation and Digitization Project	Homogenous Object Collection	
Cyrus F. Jenkins Civil War diary, 1861-1862	Monograph	
Dallas Museum of Art Collections	Super-Collection	12 sub-collections
Day Family Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Day Family Records	Homogenous Object Collection	sub-collection of Arizona Memory Project
Detroit Historical Museums and Society Historic Costume Collection	Homogenous Object Collection	sub-collection of Digital Dress
Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Digital Dress: Historic Costume Collection	Super-Collection	4 sub-collections
Digital Humphrey Winterton Collection of East African Photographs	Thematic Research Collection	
Digital Past	Presumably a super-collection	Unavailable
Digital Video Library Toolkit for Museums and Libraries with Limited Resources	Website	
Discover Babylon	Video Game	
Dorothea June Grossbart Historic Costume Collection	Homogenous Object Collection	sub-collection of Digital Dress
Dorothea Lange Collection 1919-1965	Homogenous Object Collection	sub-collection of the On-line Archive of California
Early Cave Creek, Arizona	Homogenous Object Collection	sub-collection of Arizona Memory Project
Early Life in Taylor, Arizona 1878 - 1940	Homogenous Object Collection	sub-collection of Arizona Memory Project
Early Publications of Yavapai College	Homogenous Object Collection	sub-collection of Arizona Memory Project
Eastern Illinois University Yearbook - Warbler	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Education by Design: Educational Visual Aids from the Bienes Center's WPA Museum Extension Project Collection	Homogenous Object Collection	sub-collection of Broward County Florida Digital Collections
Edward S. Curtis' The North American Indian	Homogenous Object Collection	
Edwin C. Bolles Collection on the History and Topography of London	Presumably a thematic research coll.	Unavailable
eFloras.org	Aggregation	21 collections
Enduring Communities: The Japanese American Experience in Arizona, Colorado, New Mexico, Texas, and Utah	Homogenous Object Collection	
Enoch Pratt Free Library's E-Stories: A Multimedia Celebration of our Multicultural Heritage	Homogenous Object Collection	
Erie Railroad Glass Plate Negative Collection	Homogenous Object Collection	sub-collection of Syracuse University Digital Library
Ethnographic Photographs of California Indian and Sonora Indian Subjects by Alfred L. Kroeber, 1901-1930	Homogenous Object Collection	sub-collection of the On-line Archive of California
Exit Art Digital Archive	Catalog	
Exploratorium Digital Asset Management Collection (EDAM)	Catalog	
Fairchild Tropical Botanic Garden Palm Collection	Thematic Research Collection	
Father Augustine Schwarz Photograph Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Federal Publications about Oklahoma	Homogenous Object Collection	sub-collection of Oklahoma Crossroads

Appendix B (cont.)

Collection	Class	NOTES
Feeding America: The Historic American Cookbook Project	Homogenous Object Collection	
Fenian Brotherhood Collection	Thematic Research Collection	
Field Trip Earth	Homogenous Object Collection	
Find-It! Illinois	IR Portal	
Flora and Fauna of the Great Lakes Region: A Multimedia Digital Collection	Super-Collection	11 sub-collections
Florida Folklife Collection	Collection	sub-collection of Florida Memory & also of Florida Photographic Collection
Folkstreams.net	Homogenous Object Collection	
For our mutual benefit : The Athens Woman's Club and social reform, 1899-1920	Homogenous Object Collection	sub-collection of Digital Library of Georgia
Forman Hanna - Selected Photographs from the Arizona State Museum	Homogenous Object Collection	sub-collection of Arizona Memory Project
Framed Items from the Collection of the Bancroft Library	Homogenous Object Collection	sub-collection of the On-line Archive of California
Franklin Automobile Photograph Collection	Homogenous Object Collection	
From Pi Beta Phi to Arrowmont: Bringing Education and Economic Development to the Great Smoky Mountains, 1910-2004	Thematic Research Collection	
GATT Digital Library: 1947-1994	Homogenous Object Collection	
George Edward Anderson Collection	Homogenous Object Collection	sub-collection of BYU Harold B. Lee Library Digital Collections
George Oiye Album, 1943-1946	Homogenous Object Collection	sub-collection of the On-line Archive of California
Georgia Legislative Documents	Homogenous Object Collection	
Gerrit Smith Broadside and Pamphlet Collection	Homogenous Object Collection	sub-collection of Syracuse University Digital Library
Getting the Message Out: National Political Campaign Materials: 1840-1860	Thematic Research Collection	
Gila County Maps	Homogenous Object Collection	sub-collection of Arizona Memory Project
Glendale Community College Archives	Homogenous Object Collection	sub-collection of Arizona Memory Project
Glendale Public Library History	Homogenous Object Collection	sub-collection of Arizona Memory Project
Global Performing Arts Database (GloPAD)	Homogenous Object Collection	
Goodspeed Manuscript Collection	Homogenous Object Collection	
Hans Hofmann Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
Hard Place	Website	
Harvey Girls of the Winslow Harvey Houses	Homogenous Object Collection	sub-collection of Arizona Memory Project
Hawaii War Records Depository	Homogenous Object Collection	sub-collection of University of Hawai'I at Manoa Library Archives and Manuscripts Collections
Hawaiian Language Newspapers	Homogenous Object Collection	
HEARTH (Home Economics Archive: Research, Tradition, and History)	Collection	
Helen Nestor Free Speech Movement Photographs	Homogenous Object Collection	sub-collection of the On-line Archive of California
Henry Ford Historic Costume Collection	Homogenous Object Collection	sub-collection of Digital Dress
Henry Sugimoto Collection 1928-1990	Homogenous Object Collection	sub-collection of the On-line Archive of California
Heritage West	Aggregation	44 institutions

Appendix B (cont.)

Collection	**Class**	**NOTES**
Highlights of the Catholic Diocese of Tucson	Homogenous Object Collection	sub-collection of Arizona Memory Project
Hisako Hibi Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
Historic Arizona County Road Maps	Homogenous Object Collection	sub-collection of Arizona Memory Project
Historic Downtown Glendale	Homogenous Object Collection	sub-collection of Arizona Memory Project
Historic Pittsburgh Image Collections	Super-Collection	47 sub-collections
Historical Maps Online	Homogenous Object Collection	sub-collection of University Library, University of Illinois at Urbana-Champaign
History at our Hands: The Ponce's Historical Archive & Historical Museum Digitalized (Coleccion Historia de Puerto Rico)	Collection	
History of Sedona	Homogenous Object Collection	sub-collection of Arizona Memory Project
History of Sedona: Farms	Homogenous Object Collection	sub-collection of Arizona Memory Project
History of Sedona: Pioneers	Homogenous Object Collection	sub-collection of Arizona Memory Project
HistoryMakers	Homogenous Object Collection	
Hoagy Carmichael Collection	Thematic Research Collection	
Hohokam of Pueblo Grande	Homogenous Object Collection	sub-collection of Arizona Memory Project
Honore Daumier Lithographs	Presumably a homogenous object coll.	
Illinois Alive! The Heritage and Texture of a Pivotal State During the First Century of Statehood (1818-1918)	Super-Collection	7 sub-collections
Illinois Digital Archives	Super-Collection	56 sub-collections
Illinois Government Information	IR Portal	
Illinois State University History	Homogenous Object Collection	sub-collection of Milner Library, ISU Digital Collections
Images and Ideas: The Collection in Focus at the Berkeley Art Museum, University of California	Homogenous Object Collection	sub-collection of the On-line Archive of California
Images of the Catholic Diocese of Tucson	Homogenous Object Collection	sub-collection of Arizona Memory Project
Indian Miniature Paintings, 1410-1976	Homogenous Object Collection	sub-collection of the On-line Archive of California
Indian Peoples of the Northern Great Plains	Aggregation	5 Institutions
Indigenous Peoples Near Winslow	Homogenous Object Collection	sub-collection of Arizona Memory Project
INFOMINE Scholarly Internet Resource Collection	Catalog	
Integrated Finding Aid to Walt Whitman Manuscripts	Catalog	
Jack Iwata Collection, 1942-1945	Homogenous Object Collection	sub-collection of the On-line Archive of California
Jackson Davis Collection of African American Educational Photographs	Thematic Research Collection	
Japanese Prints	Homogenous Object Collection	sub-collection of the On-line Archive of California
John Bloomfield Jervis Papers	Index	
John Brown / Boyd B. Stutler Collection Database	Thematic Research Collection	
Keystone-Mast Collection, 1870-1963	Homogenous Object Collection	sub-collection of the On-line Archive of California
Kinetic Models for Design Digital Library	Thematic Research Collection	
King County Snapshots: A photographic heritage of Seattle and surrounding communities	Super-Collection	13 sub-collections

Appendix B (cont.)

Collection	Class	NOTES
Landscape Prints and Drawing Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
League of Nations Statistical and Disarmament Documents	Homogenous Object Collection	
Linking Florida's Natural Heritage	Homogenous Object Collection	sub-collection of State University Libraries of Flor.
Liverpool (N.Y.) Public Library Local History Photographic Collection	Homogenous Object Collection	
Living Museum Online: Preserving and Digitizing the Story of Illinois	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Long Island Memories	Super-Collection	139 sub-collections
Louisiana Gumbo: A Recipe for Empowerment	Homogenous Object Collection	
Louisiana Purchase: A Heritage Explored	Homogenous Object Collection	sub-collection of LOUISiana Digital Library
Louisiana State Museum Jazz Collection	Homogenous Object Collection	sub-collection of LOUISiana Digital Library
Maine Memory Network	Homogenous Object Collection	
Maine Music Box	Super-Collection	5 sub-collections
Making of Modern Michigan: Digitizing Michigan's Hidden Past	Super-Collection	102 sub-collections
Making Sense of Modern Art	Multi-media Website	
Marcel Breuer Architectural Drawings and Sketches	Homogenous Object Collection	sub-collection of Syracuse University Digital Library
Maricopa Pottery (Connell Collection)	Homogenous Object Collection	sub-collection of Arizona Memory Project
Mark Twain's Mississippi	Thematic Research Collection	
Masterworks Online	Super-Collection	7 sub-collections
Meadow Brook Hall Historic Costume Collection	Homogenous Object Collection	sub-collection of Digital Dress
Medallion Papers	Homogenous Object Collection	sub-collection of Arizona Memory Project
Medieval manuscripts in the Digital Age: New Paths to Old Books in the Free Library of Philadelphia	Homogenous Object Collection	aka Medieval Manuscripts
Men, Mines and Money	Homogenous Object Collection	sub-collection of Arizona Memory Project
Mind Models: Artificial Intelligence Discovery at Carnegie Mellon	Website	contains links to 2 separate collections
Mining and Mother Jones in Mount Olive	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Minnesota Historical Society Online Resources	Super-Collection	26 or more sub-collections
Minnesota Immigrant Oral Histories Online	Homogenous Object Collection	sub-collection of Minnesota Historical Soc. Colls.
Missouri Botanical Garden Library Heritage Materials	Thematic Research Collection	
MOAC: California museums working with libraries and archives to increase and enhance access to cultural collections	Super-Collection	26 sub-collections
Mohave Museum - History of Transportation in Mohave County	Homogenous Object Collection	sub-collection of Arizona Memory Project
Montana Memory Project	Super-Collection	42 sub-collections
Montezuma's Castle Historic Photo Archive	Homogenous Object Collection	sub-collection of Arizona Memory Project
Moriyuki Shimada Scrapbook, 1942-1945	Homogenous Object Collection	sub-collection of the On-line Archive of California
Murder & Mayhem: The Strange Saga of Winnie Ruth Judd	Homogenous Object Collection	sub-collection of Arizona Memory Project
National Collection of Endangered Plants	Homogenous Object Collection	
Native American Collection - McLean County Museum of History and ISU	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Nature Museum Online	Website	

Appendix B (cont.)

Collection	Class	NOTES
Navajo County Historical Society Collection Highlights	Homogenous Object Collection	sub-collection of Arizona Memory Project
Nebraska Memories	Super-Collection	30 sub-collections
Nebraska Western Trails	Homogenous Object Collection	
New Jersey Digital Highway	Super-Collection	4 sub-collections
New York Public Library's Picture Collection Online	Homogenous Object Collection	
NJVid: State of the Art Video Access	Super-Collection	18 sub-collections
North Carolina Experience, Beginnings to 1940	Homogenous Object Collection	sub-collection of Documenting the American South (DocSouth)
North Carolinians and the Great War	Homogenous Object Collection	sub-collection of Documenting the American South (DocSouth)
Oak Ridge Cemetery, Illinois Interment Records	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Object of History	Website	
Oklahoma Authors	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
Oklahoma Crossroads	Super-Collection	10 sub-collections
Oklahoma Image	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
Oklahoma Resources	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
Oklahoma State Government Publications	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
Olasee Davis Newspaper Articles	Homogenous Object Collection	
Old Master Prints	Homogenous Object Collection	sub-collection of the On-line Archive of California
Old Trails Museum Collection Highlights	Homogenous Object Collection	sub-collection of Arizona Memory Project
Oliver Family Photograph Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
Olympic Peninsula Virtual Community Museum	Super-Collection	12 sub-collections
Oneida Community Collection in the Syracuse University Library	Collection	
Oral Histories of Gila County	Homogenous Object Collection	sub-collection of Arizona Memory Project
Oral Histories of the American South	Homogenous Object Collection	sub-collection of Documenting the American South (DocSouth)
Oral Histories of the White Mountains	Homogenous Object Collection	sub-collection of Arizona Memory Project
Oral History Collection of the University of Illinois at Springfield	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Our Americas Archive Partnership	Super-Collection	2 sub-collections
Park Forest - An Illinois Planned Community	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Phoenix College - The Early Years	Homogenous Object Collection	sub-collection of Arizona Memory Project
Phoenix Jewish News Photographs Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Photograph Connoisseurship Resource	Thematic Research Collection	
Photohio.org	Super-Collection	10 sub-collections
Photomuse	Website	
Picturing Augusta: historic postcards from the collection of the East Central Georgia Regional Library System	Homogenous Object Collection	sub-collection of Digital Library of Georgia
Plant Images at Missouri Botanical Garden	Index	
PlantCollections(TM)	Collection	
Powwow Photographs by Ann Leonard	Homogenous Object Collection	sub-collection of Arizona Memory Project

Appendix B (cont.)

Collection	Class	NOTES
Project Introspection	Homogenous Object Collection	
Public Art in the Bronx	Homogenous Object Collection	
Publishers' Bindings Online, 1815-1930: The Art of Books (PBO)	Homogenous Object Collection	
Pullman Company Car Drawings, 1870-1969 (bulk 1919-1969)	Homogenous Object Collection	sub-collection of CARLI Digital Collections
Quilt Index	Super-Collection	29 sub-collections
Raid on Deerfield: The Many Stories of 1704	Thematic Research Collection	
Remembering the Houses of Western Springs	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Revolutionary City: Developing a Virtual Reality Model of Williamsburg in 1776	Website	
Richard Vogler Cruikshank Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
Robert B. Honeyman, Jr. Collection of Early Californian and Western American Pictorial Material	Homogenous Object Collection	sub-collection of the On-line Archive of California
Rochester Images	Homogenous Object Collection	
Rock Art of Cochise County	Homogenous Object Collection	sub-collection of Arizona Memory Project
Rome-Turney Radiator Company Records Collection	Thematic Research Collection	
Ronald G. Becker Collection of Charles Eisenmann Photographs	Homogenous Object Collection	sub-collection of Syracuse University Digital Library
Sahuaro Ranch History	Homogenous Object Collection	sub-collection of Arizona Memory Project
Samuel Hugh Hawkins diary, January-July 1877	Monograph	sub-collection of Digital Library of Georgia
Sanborn fire insurance maps for Georgia towns and cities, 1884-1922	Homogenous Object Collection	sub-collection of Digital Library of Georgia
Scottsdale Remembers: Recollections of Our Past	Homogenous Object Collection	sub-collection of Arizona Memory Project
Scottsdale's History in Images	Homogenous Object Collection	sub-collection of Arizona Memory Project
Senator Barry M. Goldwater: An Arizona Legend	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum American Indian Image Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Audio Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Buildings Image Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Map Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Military Image Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Mining Image Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot Hall Museum Transportation Image Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Sharlot M. Hall: Arizona's Curator	Homogenous Object Collection	sub-collection of Arizona Memory Project
Shipler Photograph Collection	Homogenous Object Collection	sub-collection of Utah State History Online Photos
Ships for victory: J. A. Jones Construction Company and Liberty ships in Brunswick, Georgia	Homogenous Object Collection	sub-collection of Digital Library of Georgia
Show Low Collection Highlights	Homogenous Object Collection	sub-collection of Arizona Memory Project
Shuffle Along: The Eubie Blake Collection	Thematic Research Collection	
Sound Model: Collaborative Infrastructure for Digital Audio	Homogenous Object Collection	
Southeast Asia Visions	Thematic Research Collection	

Appendix B (cont.)

Collection	Class	NOTES
Southeastern Native American Documents, 1730-1842	Super-Collection	37 sub-collections
Southern Homefront, 1861-1865	Homogenous Object Collection	sub-collection of Documenting the American South (DocSouth)
Southern Oregon History Collection	Thematic Research Collection	
Springfield Aviation Company Collection	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Springfield College YMCA Historical Image Collection	Super-Collection	6 sub-collections
Street and Smith Publishers' Archive and Dime Novel Cover Art	Homogenous Object Collection	sub-collection of Syracuse University Digital Library
Summons to Comradeship: World War I and II Posters	Homogenous Object Collection	
Teaching with Digital Content	Homogenous Object Collection	sub-collection of University Library, University of Illinois at Urbana-Champaign
Ten O Clock News	Homogenous Object Collection	sub-collection of WGBH Open Vault
Tennessee Documentary History, 1796-1850	Super-Collection	6 sub-collections
Terence Vincent Powderly Photographic Prints	Homogenous Object Collection	
Territorial Kansas Online	Thematic Research Collection	
Texas ETD Repository	Super-Collection	18 sub-collections
Texas Heritage Online	Presumably a super-collection	Unavailable
Thar's gold in them thar hills: Gold and gold mining in Georgia, 1830s-1940s	Super-Collection	3 sub-collections
Theresa Hak Kyung Cha Collection 1971-1991	Homogenous Object Collection	sub-collection of the On-line Archive of California
Thunderbird School of Global Management - Historical Collections	Homogenous Object Collection	sub-collection of Arizona Memory Project
TIDES: Teaching, Images & Digital Experiences	Super-Collection	15 sub-collections
TIMEA (Travelers in the Middle East Archive)	Thematic Research Collection	
Timepieces	Website	
Timothy Vedder Letters	Homogenous Object Collection	
Tohono O'odham Collection, 1970 - 1980, Helga Teiwes Photographer	Homogenous Object Collection	sub-collection of Arizona Memory Project
T-RACES: a Testbed for the Redlining Archives of California's Exclusionary Spaces	Thematic Research Collection	
Trading Post Families of Winslow, AZ	Homogenous Object Collection	sub-collection of Arizona Memory Project
True North: Mapping Minnesota's History	Thematic Research Collection	
Trust Territory of the Pacific Islands	Homogenous Object Collection	
Tucson Territorial Pioneer Project	Homogenous Object Collection	sub-collection of Arizona Memory Project
Tulsa Race Riot Documents	Homogenous Object Collection	sub-collection of Oklahoma Crossroads
UCLA Fowler Museum of Cultural History Collection	Homogenous Object Collection	sub-collection of the On-line Archive of California
University Goes to War, World War I Women	Homogenous Object Collection	sub-collection of Illinois Digital Archive
University of the Virgin Islands Research Reports and Occasional Papers	Super-Collection	7 sub-collections
Upper Mississippi Valley Digital Image Archive	Super-Collection	11 sub-collections
USS Arizona Silver Service Collection	Homogenous Object Collection	sub-collection of Arizona Memory Project
Utah Digital Newspapers	Super-Collection	175 sub-collections
Vachel Lindsay Collection	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Vanishing Georgia	Homogenous Object Collection	sub-collection of Digital Library of Georgia

Appendix B (cont.)

Collection	Class	NOTES
Views of Old Morenci and Metcalf	Homogenous Object Collection	sub-collection of Arizona Memory Project
Virgin Islands Funeral Memorial Booklets	Homogenous Object Collection	
Virgin Islands Historical Photographs	Homogenous Object Collection	
Virtual Motor City: Images from the Detroit News	Homogenous Object Collection	
Visual Index to the Virtual Archive of the Skyscraper Museum	Super-Collection	3 sub-collections
Voices of the Colorado Plateau	Homogenous Object Collection	
vPlants: A Virtual Herbarium of the Chicago Region	Thematic Research Collection	
Walt Whitman Archive	Thematic Research Collection	
Walter Muramoto Collection, 1942-1945	Homogenous Object Collection	sub-collection of the On-line Archive of California
WebERA	Website	
Western New York Suffragists: Winning the Vote	Thematic Research Collection	
Western Soundscape Archive	Homogenous Object Collection	sub-collection of the J. Willard Marriot Library at the University of Utah
Western Trails: An Online Journey	Homogenous Object Collection	10 exhibits
Western Waters Digital Library	Aggregation	15 institutions
Western Ways Features Company Photographs	Homogenous Object Collection	sub-collection of Arizona Memory Project
WGBH Open Vault	Super-Collection	7 sub-collections
William Gedney: Photographs and Writings	Thematic Research Collection	sub-collection of Duke University Libraries Digital Collections
William Hayes Collection, 1820-1860	Homogenous Object Collection	sub-collection of Illinois Digital Archive
Worklore: Brooklyn Workers Speak	Super-Collection	4 sub-collections
World's Columbian Exposition of 1893	Thematic Research Collection	
World's Columbian Exposition of 1893, and the Founding and Early History of The Field Museum	Homogenous Object Collection	sub-collection of Illinois Digital Archive
WPA TVA Archaeological Photograph Archive	Homogenous Object Collection	
Writers of the Purple Sage: Origins of a National Myth	Homogenous Object Collection	sub-collection of Arizona Memory Project
YMCA College Image Collection	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection
YMCA Historical Image Collection (subset)	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection
YMCA Portrait Collection	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection
YMCA Poster Collection	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection
YMCA Training School Image Collection	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection
YMCA World War I Image Collection	Homogenous Object Collection	sub-collection of the Springfield College YMCA Historical Image Collection

Appendix C: Collections selected for experimentation.

Candidate Collections	Clearly an OAI Data Provider?	Collection "Size"
Beyond the Shelf: Serving Historic Kentuckiana Through Virtual Access		410000
Brooklyn Daily Eagle Online		147000
Colorado's Historic Newspaper Collection		150000
Cuneiform Digital Library		95000
Dallas Museum of Art Collections		13000
Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois		15921
Exploratorium Digital Asset Management Collection (EDAM)		10022
Florida Folklife Collection		10000
GATT Digital Library: 1947-1994		47130
George Edward Anderson Collection	Yes	14020
HEARTH (Home Economics Archive: Research, Tradition, and History)	Yes + MODS	385308[1]
History at our Hands: The Ponce's Historical Archive & Historical Museum Digitalized (Coleccion Historia de Puerto Rico)		10000
John Brown / Boyd B. Stutler Collection Database		20000
Maine Memory Network		10000
Mind Models: Artificial Intelligence Discovery at Carnegie Mellon		323781
Montana Memory Project	Yes	13275
New York Public Library's Picture Collection Online		30000
Plant Images at Missouri Botanical Garden		56000
PlantCollections(TM)		94600
Quilt Index		18674
TIDES: Teaching, Images & Digital Experiences		23000
Upper Mississippi Valley Digital Image Archive		175000
Utah Digital Newspapers		500000
Vanishing Georgia		18000
Virtual Motor City: Images from the Detroit News		15251
	Total Items	2,604,982

Appendix 3 Footnotes:

[1]: Total numbers reflect pages from 934 books and 8 journals.

Appendix D: Target Collection Classifications

Collection	Portal Class	Information Object Class	OAI-PMH Implementation?	Harvesting Candidate?	Broadcast Metasearch Candidate?
Beyond the Shelf	Open Search, Browsable, Indexable	Compound Surrogates	No	No	Yes
Brooklyn Daily Eagle Online	Open Search, Browsable, Indexable	Compound Surrogates	No	No	Yes
Colorado's Historic Newspaper Collection	Open Search, Browsable, Indexable	Compound Surrogates	No	No	Yes
Cuneiform Digital Library	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Dallas Museum of Art Collections	Open Search, Browsable, Indexable	Complex Surrogates	No	No	Yes
Digital Archive of 1936-1941 Historical Aerial Photography of the State of Illinois	Open Search, Browsable, Indexable	Compound Surrogates	Yes	No	No
Exploratorium Digital Asset Management Collection	Open Search, Not Browsable, Not Indexable	Complex Surrogates	No	No	Yes
Florida Folklife Collection	Open Search, Browsable, Not Indexable	Complex Surrogates	No	No	Yes
GATT Digital Library	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
George Edward Anderson Collection	Open Search, Browsable, Indexable	Complex Surrogates	Yes	No	No
HEARTH	Open Search, Browsable, Indexable	Compound Surrogates	Yes	No	No
History at our Hands	Open Search, Not Browsable, Not Indexable	Complex or Traditional Surrogates	No	No	Yes
John Brown / Boyd B. Stutler Collection Database	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Maine Memory Network	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Mind Models	Open Search, Browsable, Indexable	Simple Surrogates	No	No	No
Montana Memory Project	Open Search, Browsable, Indexable	Compound or Complex Surrogates	Yes	No	No
New York Public Library's Picture Collection Online	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Plant Images at Missouri Botanical Garden	No Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
PlantCollections	Open Search, Not Browsable, Not Indexable	Complex or Traditional Surrogates	No	No	Yes
Quilt Index	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
TIDES	Open Search, Browsable, Indexable	Compound Surrogates	No	Yes	No

Appendix D (cont.)

Collection	Portal Class	Information Object Class	OAI-PMH Implementation?	Harvesting Candidate?	Broadcast Metasearch Candidate?
Upper Mississippi Valley Digital Image Archive	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Utah Digital Newspapers	Open Search, Browsable, Indexable	Compound Surrogates	No	Yes	No
Vanishing Georgia	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No
Virtual Motor City	Open Search, Browsable, Indexable	Complex Surrogates	No	Yes	No

Appendix E: Upper Mississippi Valley Digital Image Archive to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Title	dc:title
Photographer	dc:creator
Studio Name/Location	dc:creator
Date Original	dcterms:created
Date Range	dcterms:created
Description	dc:description
Location Depicted	dcterms:spatial
Time Period	dcterms:temporal
Subject	dc:subject
Notes	dc:description>Notes:
Ordering Information	dc:rights
Image Number	dc:identifier>Image No.:
Hardcoding URI	dc:identifier
Repository	dc:source
Repository Collection	dcterms:isPartOf
Physical Location	NOT BEING MAPPED
Object Description	dcterms:extent
Digital Reproduction Information	NOT BEING MAPPED
Date Digital	NOT BEING MAPPED
Acquisition	NOT BEING MAPPED
Restrictions	dc:rights
Publisher	dc:publisher
File Name	dc:identifier>File Name:
Date Record	NOT BEING MAPPED
Record Created By	NOT BEING MAPPED
Date Record Modified	NOT BEING MAPPED
Record Modified By	NOT BEING MAPPED
CD Volume Name	NOT BEING MAPPED
Original Database	dc:publisher
Hardcoding “image/jpeg”	dc:format
**Hardcoding “Photograph” OR “Text”	dc:type

Appendix F: King County Snapshots – Wing Luke Asian Museum [sub-collection] to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Title	dc:title
Photographer	dc:creator
Date	dcterms:created
Caption	dc:description>Caption:
Notes	dc:description>Notes:
Subjects	dc:subject
Personal Names	dc:subject
Places	dcterms:spatial
Digital Collection	dcterms:isPartOf
Image Number	dc:identifier>WLAM No.:
Ordering Information	dc:rights
Credit Line	NOT BEING MAPPED (redundant with Digital Collection)
Repository	dc:source
Physical Description	dcterms:extent
Type	NOT BEING MAPPED
Hardcoding “Photograph”	dc:type
Hardcoding “image/jpeg”	dc:format
Digital Reproduction Information	NOT BEING MAPPED

Appendix G: Illinois Digital Archive – Arthur, Once Upon a Time – Local History Images of Arthur [sub-collection] to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Title	dc:title
Subject [LCSH]	dc:subject
Subject [Local]	dc:subject
Description	dc:description
Date.Original	dcterms:created
Relation.IsPartOf	dcterms:isPartOf
Coverage.Geographic	dcterms:spatial
Collection Publisher	dc:publisher
Rights	dc:rights
Hardcoding URI	dc:identifier
Identifier	dc:identifier>Local Identifier:
Type	NOT BEING MAPPED
Hardcoding “Photograph”	dc:type
Hardcoding “image/jpeg”	dc:format

Appendix H: TIDES – Cason Monk-Metcalf Funeral Directors [sub-collection] to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Title	dc:title
Description	dc:description
Creator	dc:creator
Subject	dc:subject
Owner	dc:source
Owner's Website	dc:rights
Associated Dates	dcterms:temporal
Type	dc:type
Format	dc:format
Rights	dc:rights
Hardcoding URI	dc:identifier

Appendix I: Center for Sacramento History to Qualified Dublin Core Crosswalk

Source Display Element	Qualified Dublin Core Element
Collection	dcterms:isPartOf
Catalog Number	dc:identifier>Catalog No.:
Title	dc:title
Subject	dc:subject
Creator	dc:creator
Artist	dc:creator
Photographer	dc:creator
Author	dc:creator
Search Terms	NOT BEING MAPPED
People	dc:subject
Place	dcterms:spatial
Object Name	dc:type
Other Name	dc:type
Medium	dc:description>Material:
Material	dc:description>Material:
Composition	dc:description>Material:
Call Number	dc:identifier>Call No.:
ISBN	dc:identifier>ISBN:
ISSN	dc:identifier>ISSN:
Description	dc:description
Lexicon Category	NOT BEING MAPPED
Lexicon Sub-Category	NOT BEING MAPPED
Date	dcterms:created
Classification	NOT BEING MAPPED (redundant with search terms)
Print Size	dcterms:extent
Year Range from/to	dcterms:temporal
Imagefile	NOT BEING MAPPED
Image	NOT BEING MAPPED
Title added entry	NOT BEING MAPPED
Publisher AND Published Place	dc:publisher
Hardcoding URI	dc:identifier
Published Date	dcterms:created
LCNO	dc:identifier>LC No.:
Scope & Content	dc:description
Dates of Creation	dcterms:created

Appendix J: Lewis and Clark Trail Heritage Foundation to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Collection	dcterms:isPartOf
Catalog Number	dc:identifier>Catalog No.:
Title	dc:title
Subject	dc:subject
Creator	dc:creator
Artist	dc:creator
Photographer	dc:creator
Author	dc:creator
Search Terms	NOT BEING MAPPED
People	dc:subject
Place	dcterms:spatial
Object Name	dc:type
Other Name	dc:type
Medium	dc:description>Material:
Material	dc:description>Material:
Composition	dc:description>Material:
Call Number	dc:identifier>Call No.:
ISBN	dc:identifier>ISBN:
ISSN	dc:identifier>ISSN:
Description	dc:description
Lexicon Category	NOT BEING MAPPED
Lexicon Sub-Category	NOT BEING MAPPED
Summary	dc:description>Summary:
Date	dcterms:created
Imagefile	NOT BEING MAPPED
Image	NOT BEING MAPPED
Publisher AND Published Place	dc:publisher
Hardcoding URI	dc:identifier
Published Date	dcterms:created
Accession Number	dc:identifier>Accession No.:
Scope & Content	dc:description

Appendix K: Longmont Museum Online Photo Collection to Qualified Dublin Core Crosswalk

Source Display Element	**Qualified Dublin Core Element**
Title	dc:title
Subject Heading	dc:subject
Creator	dc:creator
Artist	dc:creator
Photographer	dc:creator
Author	dc:creator
Date	dcterms:created
People	dc:subject
Place	dcterms:spatial
Description	dc:description
Catalog Number	dc:identifier>Catalog No.:
Hardcoding URI	dc:identifier
Object Name	dc:type
Notes	dc:description>Notes:
Exact Date	dcterms:created
Print Size	dcterms:extent
Copyright Information	dc:rights
Image	NOT BEING MAPPED
Business and Organization Keywords	NOT BEING MAPPED
Hardcoding “image/jpeg”	dc:format