Beyond Size and Search: Building Contextual Mass in Digital Aggregations for Scholarly Use

Carole L. Palmer
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel
Champaign, IL, 61820
1-217-244-0653
clpalmer@illinois.edu

Oksana L. Zavalina
Department of Library and Information Sciences
University of North Texas
1155 Union Circle #311068
Denton, Texas 76203-5017
1-940-565-3736
Oksana.Zavalina@unt.edu

Katrina Fenlon
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel
Champaign, IL, 61820
1-217-265-5406
kfenlon2@illinois.edu

ABSTRACT
At present there are no established collection development methods for building large-scale digital aggregations. However, to realize the potential of the collective base of digital content and advance scholarship, aggregations must do more than provide search of sizable bodies of content. Informed by empirical understanding of scholarly information practices, the IMLS Digital Collections and Content project developed an aggregation strategy for building Opening History, one of the largest digital cultural heritage aggregations in the country. The strategy applied policy-driven collecting, based on the principle of contextual mass, and conspectus-style evaluation of collection-level metadata to identify strong subject areas within the aggregation. Analysis of density, interconnectedness, diversity, and small/large collection complementarity determined subject concentrations and thematic strengths to be prioritized for future collection development and used as organizational structures for browsing and visualization. The approach models how scholars build their own personal research collections, as they follow leads from collection to collection across institutions near and far, and adds value that cannot be achieved through conventional retrieval and browsing at the item-level.

Keywords
Collection evaluation, collection policy, digital aggregations, thematic research collections, scholarly information use, subject access, subject analysis, collection-level metadata

INTRODUCTION
As researchers do more and more of their work online in the Google-centric Web environment, curated digital collections will become increasingly important as anchors for meaningful engagement with digital information. At the same time, large aggregations of these curated collections will be essential as the backbone of our evolving e-research platform, if we are to exploit the full potential and economies of scale made possible by the vast range of distributed digital content.

There are a number of collection development challenges unique to building aggregations. For example, how do we retain the context and identity of individual collections as they are funneled into massive repositories or linked together in extensive networks of items? Just as importantly, how do we uncover the new cohesive areas of content as they build up? And, how do we systematically build aggregate resources in ways that cumulatively produce new cohesive units of value to users?

In the cultural heritage domain, institutions and funding agencies have invested intensively in projects that have produced countless digital collections, but we have not yet leveraged their collective value. Individual projects have been aimed at worthy objectives, such as promoting a unique special collection or providing a scholarly community broader access to particular primary materials. However, overall production has not been guided by more global, long-term goals or even basic collection development principles. Thus, libraries and museums have produced thousands of resources that have made substantial contributions at the local level, but most have not planned their digital programs for participation in large national aggregations (Bishoff & Allen, 2004). Moreover, there are no established collection development principles or practices for building large-scale aggregations.

Digital aggregations can provide essential metastructures for unifying distributed content. However, the act of bringing together and providing access to a large number of collections does not guarantee that the resulting aggregation

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

will be a useful resource for researchers. While it is true that critical mass and search usability are important, they are not sufficient for realizing the full benefits of a large, rich collection of collections. Size and search are baseline requirements, but they cannot be the sole priorities in development, or we risk ending up with aggregations that lose what scholars value most from research collections, and will fail to exploit the potential to generate new, significant collections that can advance scholarship.

The IMLS Digital Collections and Content (IMLS DCC) initiative developed an aggregation strategy drawing on empirical understanding of scholarly information practices and traditional research library conspectus style collection evaluation (cf. Ferguson, Grant, & Rutstein, 1988; Wood, 1992). The approach provided the impetus and the guiding principles for a new IMLS DCC derivative aggregation—Opening History. The resource has grown steadily since its inception in 2008 and is now likely the largest digital cultural heritage aggregation in the country.

Development was guided by our understanding of the information practices of historians and humanities scholars, with a particular focus on how they value and make use of research collections of primary sources. We know, for example, that historians search mostly for unpublished primary sources: text-based objects, such as diaries, wills, letters, manuscripts, as well as images, such as photographs, portraits, architectural drawings, moving image materials and such (Case, 1991). They often use electronic means to locate primary materials for their research, and visiting websites of known repositories has been a more frequent behavior than using search engines (Tibbo, 2003). Historical researchers greatly value digital archives and their finding aids, which often help them locate materials that they have sought for years (Duff & Johnson, 2002), with use peaking during the initial stages of a research project (Buchanan et al., 2005). Browsing has been a long-standing and crucial information seeking practice of historians (Ellis & Oldman, 2005), and contextual information, about relationships among materials and how they are organized is critical for navigating through content and for the ongoing process of interpretation (Duff & Johnson, 2002).

For humanities scholars more generally, "collections" are highly important in the production of research. (Brogan, 2006; Palmer, 2005). They are a form of capital that has great pull with scholars, attracting them to visit or even take a position at an institution, to support their need to engage with collections as whole, dense units for exploration and study (Brockman et al., 2001). Library and archival collections, be they digital or physical, are fundamentally resources that provide evidence for inquiry (Buckland, 1999). They are intentionally created wholes (Currall, Moss, & Stuart, 2004), for which "the totality of the records provides information that no individual record can. Historians must comprehend the records in their context rather than as separate disembodied items. Without this context information, the historian could easily misinterpret the meaning or significance of the information in an individual record" (Duff & Johnson, 2002, p.487).

The concept of the scholarly subject collection has long been at the heart of research library collection development and evaluation. The Research Libraries Group (RLG) Conspectus was designed for establishing a "national collection", consisting of multiple, individually strong research-level subject collections. Developed in the late 1970s, the RLG Conspectus is a protocol for evaluating collection strengths and weaknesses, and tracking the depth and intensity of past and current collecting across different institutions. It "supplies a framework that encourages selectors to … direct funds at targeted weaknesses rather than on unneeded duplication" (Ferguson, Grant, & Rutstein, 1988, p. 198), but subject assessments can also be used to identify targets for growth of research-level strengths. Conspectus collection assessment has been informing development of physical research library collections for decades, and, as we show here, the approach can be adapted to guide development of aggregations of digital research collections.

Below, we trace the development of the Opening History aggregation and then report on the principles and methods applied and tested in developing the content, structure, and functionality of the aggregation. In closing, we discuss the implications for building subject-based, nationally-scoped aggregations for scholarly use.

BACKGROUND

Growth of the IMLS DCC

The IMLS DCC began aggregating digital cultural heritage collections in 2003 with the aim of providing a single point of access to nearly all of the digital content funded by IMLS National Leadership Grants (NLG), and selected LSTA-funded material. By September 2007 the collection registry included 202 collections from hundreds of institutions, primarily libraries, museums, and archives. The item-level metadata repository contained more than 300,000 records, representing diverse types of materials, ranging from photographs and manuscripts to maps, sheet music, and multimedia exhibits, harvested using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) from approximately twenty percent of the collections. As the resource grew over the next five years, and we studied the perspectives of resource developers and user interactions with the aggregation, important basic questions arose, many of which represent general concerns of any large-scale resource that depends on metadata harvesting. What's in it, really? What's the rationale for this grouping of content? What are the intended audiences?

The project made important contributions in collection description, interoperability, and sharable metadata (Foulonneau et al., 2005; Shreeves et al., 2005; Cole et al., 2006; Shreeves, Riley, & Milewicz, 2006). At the same

time, the collection itself and its intended user community became increasingly nebulous, partly because, while the collections are uniformly represented through consistent application of a collection-level metadata scheme, item-level content was harvested in varying forms from only a portion of the collections. As the resource increased in size it became increasingly difficult to discern the scope and depth of the subject content, and like most web-based access systems, the interface masked the rich material within, with no means for readily displaying the strengths and contours that were quickly developing.

This set of interrelated concerns became the focus of the second phase of development, which began in the fall of 2007 with the Opening History initiative (see http://imlsdcc.grainger.uiuc.edu/history/). The objectives were more research oriented for this phase of the project, but were directed at the important practical aim of continuing to build the strongest content area in the IMLS DCC—U.S. history, for the primary user group for that content—history researchers.

Toward Greater Cohesion with Opening History

The new Opening History (OH) initiative expanded the base of history-oriented IMLS DCC content to include other cultural heritage collections from across the country, allowing for faster and more focused growth. As of July 2010, OH included more than 850 collections, with more than 1,000,000 harvested items from digital collections of varying sizes from 39 states, the District of Columbia, Puerto Rico, and the Virgin Islands. The rapid growth was fostered primarily through highly productive interactions with the Chief Officers of the State Library Agencies (COSLA), which resulted in state librarians encouraging their institutions to contribute digital content and to help build a more nationally representative aggregation. The expansion was managed by applying two core conventions of traditional collection development: systematic evaluation of the content and a collection development policy (Evans, 2005).

The collection policy was developed to guide selection of relevant history content for history researchers, broadly scoped to include academic and non-academic history scholars; teachers and students, particularly at the undergraduate, graduate, and postgraduate levels; genealogists and "citizen historians"; and others who learn or do research in settings such as museums and public libraries. Two main directives identified in the policy are to (1) build further in the subject areas inherited from the IMLS DCC and (2) to fill gaps in geographic coverage (http://imlsdcc.grainger.uiuc.edu/docs/CollectionDevelopmentPolicy.pdf).

Contextual Mass and Thematic Research Collections

The collection development approach was guided by the principle of "contextual mass", a concept derived from our previous research on the information work of humanities scholars (Brockman et al., 2001) and analysis of scholar-produced thematic research collections (Palmer, 2004). The principle places the emphasis on collecting materials that work together as a system of sources, with meaningful interrelationships between different types of materials and subjects, to support research inquiry. Collections can be of any size, since the idea of contextual mass asserts that striving for critical mass through opportunistic collecting should not drive growth. Instead, selection is based on criteria that produce dense, rich, and cohesive groupings of sources for research and analysis.

Applying the principle of contextual mass to the development of OH generated a coherent U.S. history aggregate with latent subject strengths. These strengths are conceptually analogous to thematic research collections—aggregates of primary sources and related materials, created by special collections curators, but also increasingly by scholars whose expertly selected collections are often at the heart of their digital scholarship. Digital scholarly products can be significant "collections" in their own right, designed to support personal research or that of a specialized research community (Smith, 2004; Palmer, 2004).

The thematic research collection has been recognized as an important genre within the field of history (Rogers, 2008) and in digital humanities more generally (Jewell, 2008-2009; Ciula & Lopez, 2009; Price, 2009). As suggested by Palmer (2004), it will play an important role in how research materials are reconfigured in the digital environment, as libraries become more involved in providing access to digital resources collected and organized by scholars, who contribute important expertise in selection, collocation, interpretation, and integration of the sources they study.

METHODS

Development of the IMLS DCC and OH aggregations has been informed by several stages of research applying multiple methods, including surveys, interviews, transaction log analysis, case studies, and metadata analysis (Palmer & Knutson, 2004; Palmer, Zavalina, & Mustafoff, 2007; Zavalina et al., 2008). The aggregation strategy presented here draws on a recent set of systematic collection evaluations of the OH aggregation, using a combination of quantitative and qualitative analyses of collection-level metadata records. Supplementary analysis of transaction logs, and interview and observation sessions with academic historians, provide insights into user interactions with collection-level information in the OH aggregation, as well as content and functionality of value to users (see Zavalina, 2010, for a detailed account of these methods). As illustrated in Figure 1, these analyses provided the input for the collection evaluation, which identified and assessed *subject concentrations*, and more specialized *thematic strengths* within the concentrations. These results were then used to inform updates of the collection development policy and in the design of interface features that exploit collection

metadata for visualization and interaction with the OH aggregate.

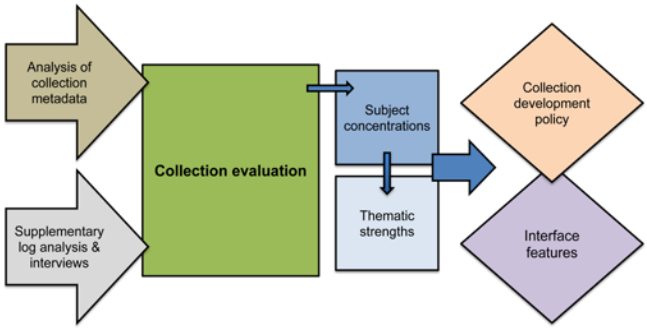

Figure 1. Aggregation strategy

Collection Evaluation

The first collection evaluation of OH was conducted to set a baseline when the new aggregation was made public in 2008. Results were then updated biannually to monitor growth and change in the aggregation as a whole over time. The primary data for the evaluation process were derived from the OH collection-level metadata records, stored in an SQL database and identified by collection ID. Values from the *Subjects* fields (represented in GEM, LCSH, and/or other controlled vocabularies) were extracted, and based on manual analysis, specific terms (e.g., adjutants, veterans, guided missile ranges) were associated with broader terms (e.g., naval history, Hawaii in World War II) to create general subject concentrations (e.g., military history).

After the initial sorting of records into general subject concentrations, the groupings were ranked by frequency, and strength factors were determined for the top subject concentrations. The *Size* field was, of course, essential for determining quantitative strength. When size information was missing from the collection metadata record, item-level metadata was reviewed, when available from harvested collections. Geographic coverage and item types from the respective Dublin Core Collection Application Profile metadata fields were also assessed, as well as information from the free-text *Description* field, which was especially important for identifying subject specific strength factors.

As outlined in Tables 3 and 4 in the Findings section, the four key strength factors were counts of 1) subject-focused collections (collections for which the primary focus is the given subject), and harvested items from those collections 2) subject-inclusive collections (broader collections, which include a subset of materials on the given subject), and harvested items from those collections; 3) item types represented; and 4) U.S. states covered.

Collection dimensions associated with contextual mass—density, interconnectedness, diversity, and small/large collection complementarity—were then assessed based on strength factors. A large number of subject-focused collections indicates high density. Lower density is associated with more subject-inclusive collections. However, interconnectedness may be indicated by subject-inclusive collections in combination with a small number of large subject-focused collections. Diversity is represented by range of genres and geographic coverage, measured by number of item types and U.S. States represented. Additional subject-specific strength factors also emerged, and have proven particularly important for assessing small/large collection complementarity.

Assessment of a subject concentration or a thematic strength is ultimately a holistic account. It combines quantitative measures based on values in metadata fields with qualitative content analysis of subject terms and unique free-text descriptions, necessary for discriminating sub-collections within a collection and identifying specialized themes across collections.

Supplementary Log Analysis

Transaction log analysis was performed periodically in our ongoing studies of the use of OH, and these results contributed to our interpretation of the collection evaluation work. This paper draws on the two most recent log analyses: one performed in the fall of 2009 on a sample of 12 weeks of OH data from 2008-2009, and another performed in spring 2010 on 3 months of OH data from January to March, 2010. The log results provided an important perspective on the degree of collection-level user interaction compared to item-level interaction, focusing on collection search, collection browse, and viewing of collection metadata records.

Supplementary Interviews and Observations

Interview and observation sessions were conducted with academic historians to study how collection-level metadata facilitates scholarly access and use. A semi-structured interview protocol was used, and participants were also observed searching and browsing on their current research topics in OH. Purposeful sampling was applied to identify three participants, two PhD candidates and a professor, with two working in the area of Native American history, and one in the area of Japanese American history during World War II. The data collection sessions lasted between 40-60 minutes and were documented through audio recording and field notes, and analyzed through iterative coding of salient themes. The results represent scholarly perspectives on collection-level information and show how scholars engage with collection-level metadata and use collection search and browse functions.

Brief results from the log analysis and the interview sessions are presented first to provide context and perspectives on collection level interaction with OH and the use of collection-level information by scholars. This is followed by a detailed report on the systematic OH collection evaluation.

FINDINGS

Use and Value of Collection Records

Transaction log data presented in Table 1 show a high level of engagement with collection metadata records, with the total page views for collections more than 4 times greater than item page views. Although collection browse is used considerably less than item browse, subject browsing is clearly the most important of the various browse options.

User interactions	Page views
Viewing collection metadata records	1,760
Viewing item metadata records	368
Collection browse:	2,939
Subject browse	953
Geographic browse	533
Project browse	502
Object type browse	487
Institution browse	311
Collection title browse	153
Item browse	4,388
Collection search	880
Item search	1,860

Table 1. User interactions with metadata records in OH

Results from the sessions with historians provide additional insights on the value of collection-level representation and how scholars think about collections as they explore sources in OH. For example, while looking at 244 results retrieved by an item-level search, one historian stated their preference for information organized at the collection level: "If I am searching for something initially, this is too much information. I'd rather see it grouped by collection and have good metadata about the collection as a whole." Another scholar found collection record information sufficient for understanding the content of interest, noting: "I don't necessarily need to see it [item metadata record]. Generally by that time [of looking at an individual item], I know how it's been described as a collection. I know what I am looking for in general." For this researcher, the availability of item records was secondary to the contextual role of the collection information.

The historians also followed links from collection records in OH back to collection homepages at the institution. In this activity, OH was navigated as a central hub that successfully connected the researcher with the originating context of a collection and allowed them to explore related digital collections or information about physical collections at the hosting institution. We also observed historians seeking out provenance information in collection metadata records, to better assess the unique nature of a collection and its historical relevance or importance. Narrow, specialized collections were of particular interest, with one historian considering them much more useful for his needs and of higher quality than more general collections.

Collection information in OH is clearly being used, as seen in the transaction log results, and it is useful to practicing scholars for navigation, interpretation, and assessment of content, as seen in the interviews and observations. The more detailed collection evaluation results presented below provide in-depth analysis of OH collections as the basis for advancing content, structure, and functionality for scholarly use. Note that the frequencies reported in the text and tables refer to total collections (subject-focused plus subject-inclusive) unless specified as one or the other.

Identifying Emerging Subject Concentrations

The most recent phase of collection evaluation was performed on the aggregate of 803 collections in April 2010. The subject concentrations identified are ranked in Table 2, based on total number of collections with materials on the subject. The largest, Military History, grew by 57% (from 53 to 83) over the recent one-year period, with approximately 10.3% of collections now designating Military History as a subject. The smallest concentration, Exploration and Travel History, was represented in 23 collections. The most prominent growth was in the second largest concentration, Native American History, which increased from 32 to 78 collections in the past year.

Below we profile the top two subject concentrations, Military History and Native American History, to illustrate the strength factors documented through the evaluation process. In keeping with our contextual mass approach, both concentrations are high in density, interconnectedness, and diversity of content. They also provide excellent examples of complementary contributions made by small collections, and the emergence of a thematic strength within the concentration.

Subject concentration	Collections with subject, April 2009	Collections with subject, April 2010	Percent of OH with subject, April 2010
Military history	53	83	10.3%
Native American history	32	78	9.7%
Transportation history	33	48	6.0%
Asian American history	41	44	5.5%
African American history	29	33	4.1%
Mining history	17	26	3.2%
Exploration and travel history	19	23	2.9%

Table 2. Largest subject concentrations in OH

Profile of Two Subject Concentrations

(1) Military History

As seen in Table 3, approximately two-thirds (56) of the 83 collections covering Military History are focused on the

subject. The subject-inclusive collections, however, account for 71% of the overall 67,473 harvested items. The concentration as a whole includes 39 different item types. Photographs/slides/negatives are the most common (represented in 47 collections), followed by Books and pamphlets (32), Prints and drawings (15), Posters and broadsides (10), and Letters (9). Seven new item types emerged in the past six months, adding databases, glass slides, interviews, paintings, sketches, and sermons.

Geographic coverage currently spans twenty-four states, with Arizona represented in (11), Illinois (8), and Louisiana (7). State level geographic coverage, represents, for example, military correspondence from officers and soldiers of a state regiment in the Civil War, or documents related to Japanese American internment in specific states during World War II.

A number of the Military History collections do not have specific region or state geographic coverage, but rather focus on military events, that may have happened outside the United States. For historians, events such as wars, battles, and tribunals are a significant dimension of Military History, and OH has solid representation of World War II (27), the Civil War of 1861-1865 (18), and World War I (12). As is typical of the heterogeneity within OH, item counts for the collections vary significantly, from one subject-focused collection consisting of only two "minute books" on World War I to an extensive collection of 7,140 photographs and posters. The minute books make a significant contribution to the concentration, adding a unique perspective from the Athens Woman's Club, a group that fundraised, lobbied and organized relief efforts during the war. As discussed below, Japanese American internment (15 collections) emerged as a substantial thematic strength within the Military History subject concentration.

Military History Strength Factors	Frequencies
Subject-focused collections	56
Items in subject-focused collections	19,388
Subject-inclusive collections	27
Items in subject-inclusive collections	48,085
Item types represented	39
U.S. states covered	24
Events covered	12

Table 3. Military History strength profile

(2) Native American History

A somewhat lower percentage of Native American History collections are subject-focused, 58% (45) of the 78, again with the subject-inclusive collections providing the majority (78%) of items. Forty-eight item types are represented, with a single collection providing 19 types. Photographs/slides/negatives (44) are the most common, followed by Books and pamphlets (23), Physical artifacts (20), Prints and drawings (18), and Letters (9). Interestingly, fifteen new item types emerged in the past six months, greatly diversifying the content, adding maps, stereoviews, book covers, clothing, correspondence, lantern slides, leaflets, lithographs, music (audio files), notebooks, photographic postcards, sculpture, tables, account books and aerial views. Collectively, the collections cover 31 states, with Arizona represented in 23, and California, Oklahoma, and New Mexico each represented in 10 or more collections.

Native American History Strength Factors	Frequencies
Subject-focused collections	45
Items in subject-focused collections	19,246
Subject-inclusive collections	33
Items in subject-inclusive collections	66,431
Item types represented	48
U.S. states covered	31
Tribes/tribal groups covered	65

Table 4. Native American History strength profile

Although not as extreme as with Military History, the size of the subject-focused collections ranges widely, from 7 to 3,786 items. Tribes was determined to be an important subject-specific factor within Native American History, with OH covering 65 tribal groups, of which 15 are represented by two or more collections. Navajo emerged as an important thematic strength, with 8 collections. Unfortunately, some collection records do not specify tribes, instead using broad statements such as "80 tribes," or "every tribe resident in Oklahoma."

Exploiting Subject Concentrations for Users

Having developed a process for identifying and profiling subject concentrations, the next aim was to develop ways to make them more explicit to OH users. The Transportation History concentration was selected for experimentation, since it is considerably smaller than Military History or Native American History, but no less interesting, with a high degree of contextual mass and a solid growth rate. Twenty-two collections with item-level metadata from the Transportation History concentration were isolated for a testbed, and collection level metadata was exploited to develop features to assist users in grasping and interacting with the range and richness of content. The design provides representation of time periods, a dimension of fundamental interest to historians. It also highlights two key subject concentration strength factors—states covered, visualized as geographic coverage on a map (Figure 1) and item types represented, supported by faceted guided search (Figure 2).

The Transportation History portal experiment demonstrated that enhanced access to a subject concentration could be implemented with relatively low investment. In the next phase of development, the structure will be extended to the Native American and Military History subject concentrations, and points of access to all three subject concentrations will be displayed on the main OH webpage,

making the strengths within the aggregate explicit to users and increasing accessibility by web crawlers.

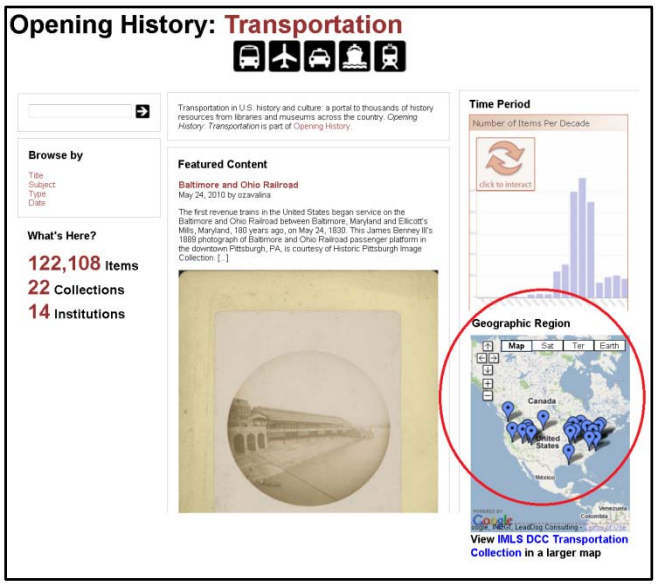

Figure 2. Geographic coverage for Transportation History

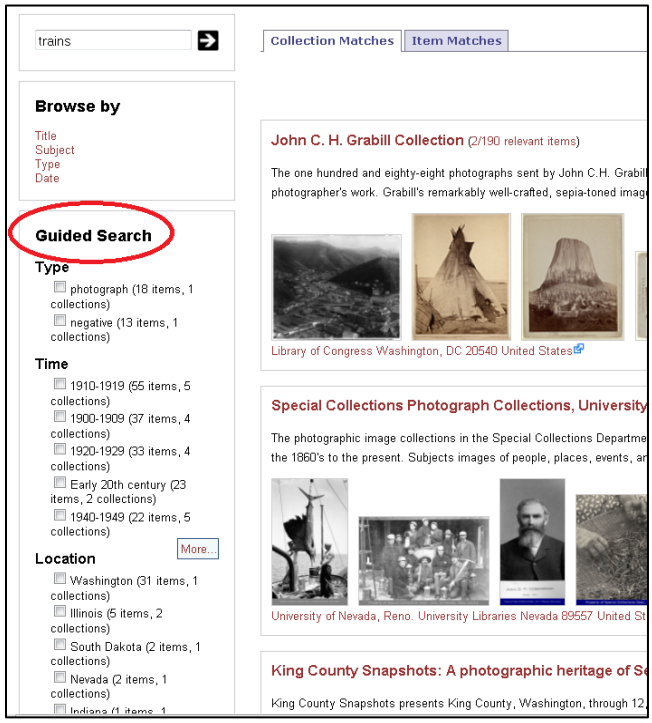

Figure 3. Guided search for Transportation History

Developing Thematic Research Collections

Bringing strong subject concentrations to the foreground is an important part of helping historians interpret the large body of content in the OH aggregate. However, the process of scholarly inquiry is only loosely associated with general subject categories, such as Military History or Native American History. Thematic research collections, discussed in the Background section, are much more closely aligned with how practicing scholars gather materials around their object of study. Scholars work deeply in more constrained subject areas to develop their unique expertise and make original contributions to their field. Their orientation to specialized subjects is evidenced in the way they gather and develop their own personal collections, seeking out and retaining items that would seem highly esoteric to the non-specialist (Palmer, Teffeau, & Pirmann, 2009).

As the body of OH resources grew over time and the collection evaluation evolved, it became apparent that specialized subjects, akin to thematic research collections, were emerging. Like subject concentrations, they could be assessed for density, diversity, interconnectedness, and complementarity. One prominent example is the subset within Military History on Japanese American Internment during World War II, based in a robust set of fifteen subject-focused collections. It is also a minor subject in several other subject-inclusive collections, which contribute valuable items, including photographs, letters, and scrapbooks of relocated Japanese Americans in Oregon and Washington state.

The Navajo thematic strength within Native American History exemplifies the significant, complementary contributions small collections can make to a concentration. One 38-item collection has a unique set of texts, including letters, account books, and notebooks on Navajo history in Arizona, that complements a larger collection of 1,124 similar items in a New Mexico Native American collection. Another small multi-media collection adds 135 oral history audio files and 51 digitized Navajo pottery items. The pottery images have high complementarity with a separate collection of 25 photographs of Indian traders and trading posts in Navajo reservations, where pottery was one of the items of trade.

The utility of including item-level assessment in collection evaluation is also demonstrated by the Navajo case. Eight collections were identified through *Subjects* or *Description* fields of the collection records, but 20 additional collections were identified through terms in item records. For one collection, 247 Navajo items were cloaked among the full set of 29,000 items. For others the numbers were quite small but the content still significant (e.g., 14 high-quality 19th century stereographs from the 7,318-item Library of Congress special collection; 3 historical popular songs about Navajo from the 6,762-item Indiana University Sam DeVincent American Sheet Music Collection; 3 theses on the history of relations between Navajo and Mormons in Utah in the 962-item Brigham Young University collection of Mormonism theses). Collectively these 20 digital collections contain 919 unique and valuable interconnected items within the Navajo theme.

A third thematic strength coalesced in the Transportation History concentration around Pullman railroad cars manufactured in the late 19th and early 20th centuries. This theme is of potential interest to academic historians, and

also to citizen historians who are railroad enthusiasts. Contributing sources include:

- A solid base of 1,500 Pullman train car blueprints from the Newberry Library's Pullman Car Company collection;
- Scattered photographs from three large Connecticut, Washington, and Pittsburgh photograph collections, depicting Pullman car exteriors, Pullman porters, and Pullman cars on various regional railway lines;
- Two glass negatives (one depicting George Pullman's wife, another showing Pullman railroad car damaged in an accident on the New York, New Haven & Hartford Railroad in 1913) from the extensive Library of Congress, George Grantham Bain collection, of more than 40,000 items;
- One obscure 19th century book, "A Story from Pullmantown," satirizing the Pullman company, from the very small 17-item Illinois Art and Literature digitized books collection at the University of Illinois.

For all three thematic strengths, the combination of content from the subject-focused collections and subject-inclusive collections provide a rich, refined information-seeking environment (Lee, 2000). The units stand in stark contrast to what would be provided by a more standard set of search results, in terms of value to the process of scholarly exploration and inquiry. In the open web or a large aggregation not optimized for collections and contextual mass, very small library collections, such as the Illinois Art and Literature digitized books, would likely be obscured. If the one satirical book on Pullman was retrieved, it would be isolated from the context of the original collection and the other complementary collections made readily accessible by thematic collocation.

DISCUSSION AND CONCLUSION

Although more multi-faceted than the standard RLG Conspectus approach, the OH collection evaluation method can be applied in a similar way to assess and compare digital research collections. More importantly for our purposes, it provides a systematic process for evaluating and enhancing aggregations of research collections. Applying the process to the OH aggregation uncovered organically growing subsets in subject areas of value for historical inquiry. Once identified, subject concentrations and thematic strengths can be further developed into premier, nationally scoped research collections through targeted collection recruitment. As new strengths are identified in the growing base of collections and harvested items, they can be built up and integrated into the fabric of historical subjects and themes. This process will generate aggregate contextual mass—a tightly knit system of collections, rather than individual sources, with meaningful interrelationships among different subject areas and types of materials, as an alternative to the expansive and scattered mass that lies behind a web browser or the typical large-scale portal.

Moreover, by making the subject strengths explicit in the interface through specialized portal views, visualization, browsing, and filtering features, the user is provided with an intrinsically derived organizational structure that assists with interpreting and interacting with the extensive body of content. Unlike many access systems that leave users to trial and error searching at the item level, the OH strategy uses collection description information to disclose content in a way that supports the scholarly practices of exploring and engaging with specialized materials within and across collections.

Collection description, by its nature, provides useful assertions about topics in a collection, which support discovery and identification of patterns within a large aggregation. In OH, the idiosyncrasies of the metadata and subject vocabularies, among the more than 1,000,000 harvested items, would make it impossible to discern trends or comprehensively capture subjects from item-level metadata. Collection-level analysis of subject concentrations and thematic strengths extends results that would be generated through basic item-level retrieval on subject terms, providing more complete recall that includes the contributions of small collections, as described above in the Pullman example. Broad recall, with links back to the context of the originating collections for every item, facilitates serendipitous browsing in ways that are essential to the conduct of historical research (Dalton & Charnigo, 2004).

The collection evaluation process has also clarified criteria for identifying strong emergent subjects within a large-scale aggregation. Subject strengths will have some consistent factors, such as both subject-focused and subject-inclusive collections, and for many subjects in an aggregation like OH, item types and geographic coverage are also important factors for scholarly users. Other factors are salient for particular subject areas, as seen with "events" within Military History. Events will also be a key factor in the regional and local history concentrations that are quickly becoming significant in OH, due to the growing contributions from public libraries, historical societies, and history museums. We also expect additional factors, such as landmarks and people, will prove to be important for local history enthusiasts and genealogical researchers. As we continue to analyze OH metadata as the aggregation grows, we will develop a fuller framework for operationalizing subject strengths within cultural heritage aggregations.

OH provides access to subject materials and collection contexts dispersed across the country, while coalescing new emergent collections in general and specialized subject areas. This process of gathering and reconfiguring sources models how scholars build their own personal research collections, as they follow leads that take them from collection to collections across institutions near and far. The

work of building high quality digital aggregations requires the same kind of purposeful selection of targeted and complementary materials from many institutions to create contextual mass within and among subjects and themes. And, in the digital age, it will be an essential part of professional collection development, if we wish to turn our vast, distributed digital investment into collective resources that can foster and advance scholarship.

ACKNOWLEDGMENTS

This research was supported by a 2007 IMLS NLG Research & Demonstration grant. Project documentation is available at http://imlsdcc.grainger.uiuc.edu/about.asp#documentation. The authors wish to acknowledge Mike Twidale, Sarah Shreeves, Richard Urban, Sunah Suh, Hong Zhang, and Jennifer Parga for their important work on the interface team developing the experimental Transportation History portal.

REFERENCES

Bishoff, L., & Allen, N. (2004). *Business Planning for Cultural Heritage Institutions.* Washington, DC: Council on Library and Information Resources. Retrieved August 2, 2010 from http://www.clir.org/pubs/reports/pub124/contents.html.

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Washington, DC: Digital Library Federation/Council on Library and Information Resources. Retrieved August 2, 2010 from http://www.clir.org/pubs/reports/pub104/pub104.pdf.

Brogan, M. (2006). *Contexts and Contributions: Building the Distributed Library*. Washington, DC: Digital Library Federation/Council on Library and Information Resources. Retrieved August 2, 2010 from http://www.diglib.org/pubs/dlf106.

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J., & Warwick, C. (2005). Information seeking by humanities scholars. In *Proceedings of the European Conference on Digital Libraries* (Vienna, Austria, Sept. 18-23, 2005), 218-229.

Buckland, M. K. (1999). *Library Services in Theory and Context*. (2nd ed.). New York: Pergamon Press. Retrieved August 2, 2010 from http://sunsite.berkeley.edu/Literature/Library/Services/index.html

Case, D. O. (1991). The collection and use of information by some American historians: A study of motives and methods. *Library Quarterly*, *61*, 61-82.

Ciula, A., & Lopez T. (2009). Reflecting on a dual publication: Henry III Fine Rolls print and web. *Literary and Linguistic Computing*, *24*(2), 129-141. Retrieved August 2, 2010 from http://llc.oxfordjournals.org/cgi/content/abstract/24/2/129.

Cole, T., Jackson, A. S., Palmer, C. L., Shreeves, S. L., Twidale, M. B., & Zavalina, O. L. (2006). *Findings Pertaining to the Framework of Guidance for Building Good Digital Collections*. White paper. Retrieved August 2, 2010 from http://hdl.handle.net/2142/722.

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria*, *58*, 131-146.

Dalton, M. S., & Charnigo, L. (2004). Historians and their information sources. *College and Research Libraries*, *65*(5), 400-425.

Duff, W. M., & Johnson, C. A. (2002). Accidentally found on purpose: Information-seeking behaviors of historians in archives. *Library Quarterly*, *72*(4), 472-496.

Ellis, D., & Oldman, H. (2005). The English literature researcher in the age of the Internet. *Journal of Information Science*, *31*(1), 29-36.

Evans, G. E. (2005). *Developing Library and Information Center Collections.* (5th ed.). Englewood, CO: Libraries Unlimited.

Ferguson, A. W., Grant, J., & Rutstein, J. S. (1988). The RLG Conspectus: Its benefits and uses. *College and Research Libraries*, *49*(3), 197-206.

Foulonneau, M., Cole, T. W., Habing, T. G., & Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (Denver, CO, June 7-11, 2005), 32-41.

Jewell, A. (2008-2009). Digital editions: Scholarly tradition in an avant-garde medium. *Documentary Editing*, *30*(3&4), 28-35. Retrieved August 2, 2010 from http://digitalcommons.unl.edu/libraryscience/183.

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

Palmer, C. L. (2004). Thematic research collections. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 348-365). Oxford: Blackwell. Retrieved August 2, 2010 from http://www.digitalhumanities.org/companion/index.html.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*, *56*(11), 1140-1153.

Palmer, C.L., & Knutson, E. (2004). Metadata practices and implications for federated collections. In *Proceedings of the 67th ASIS&T Annual Meeting* (Providence, RI, Nov. 12-17, 2004).

Palmer, C. L., Teffeau, L C., & Pirmann, C. M. (2009). *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Dublin, OH: OCLC Research and Programs.

Palmer, C.L., Zavalina, O.L., & Mustafoff, M. (2007). Trends in metadata practices: a longitudinal study of collection federation metadata. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, Canada, June 19-23, 2007), 386-395.

Price, K. (2009). Edition, project, database, archive, thematic research collection: What's in a name? *Digital Humanities Quarterly, 3*(3). Retrieved August 2, 2010 from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1068&context=englishfacpubs.

Rogers, B. M. (2008). The historical community and the digital future. In *3rd Annual James A. Rawley Graduate Conference in the Humanities, Imagining Communities: People, Places, Meanings.* (Lincoln, NE, April 12, 2008). Retrieved August 2, 2010 from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1025&context=historyrawleyconference.

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is 'quality' metadata 'shareable' metadata? The implications of local metadata practice on federated collections. In *Proceedings of the 12th National Conference of the Association of College and Research Libraries,* (Minneapolis, MN, April 7-10, 2005), 223-237.

Shreeves, S. L., Riley, J., & Milewicz, L. (2006). Moving towards sharable metadata. *First Monday*, *11*(8). Retrieved August 2, 2010 from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1386/1304.

Smith, M. N. (2004). Electronic scholarly editing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* (pp. 306–322). Oxford: Blackwell. Retrieved August 2, 2010 from http://www.Digitalhumanities.org/companion/index.html.

Tibbo, H. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. *The American Archivist, 66*(1), 9-50.

Wood, R. (1992). A conspectus of the conspectus. *Acquisitions Librarian*, *4*(7), 5-23.

Zavalina, O. L. (2010). *Collection-level Subject Access in Aggregations of Digital Collections: Metadata Application and Use*. (Doctoral dissertation). Urbana-Champaign: University of Illinois, 192 p. Retrieved August 2, 2010 from http://hdl.handle.net/2142/16620.

Zavalina, O.L., Palmer, C.L., Jackson, A.S., & Han, M.-J. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata*, *8*(4), 263-292.