Scholar-built collections: A study of user requirements for research in large-scale digital libraries

Katrina Fenlon¹, Megan Senseney¹, Harriett Green², Sayan Bhattacharyya¹, Craig Willis¹, J. S. Downnie¹

¹Graduate School of Library and Information Science
²University Library
University of Illinois at Urbana-Champaign
{kfenlon2, mfsense2, green19, sayan, willis8, jdownie}@illinois.edu

ABSTRACT
To realize the great potential value of large-scale digital libraries, we need a fuller understanding of the range of ways in which scholarly communities conduct research, or want to conduct research within them. Scholars build collections in the course of their work. How can we anticipate and support various kinds of collection-building and -use, in order to support the diversity of researchers who work in libraries of digital books? This paper reports selected results of a study of how potential user groups of the HathiTrust Digital Library create and use collections in their research. This study aims to contribute to our broader understanding of scholarly practice, particularly of humanities scholars’ collecting activities. The results of the study inform ongoing work to develop a workset-creation tool for the HathiTrust Research Center.

Keywords
Collections, digital libraries, metadata.

INTRODUCTION
As large-scale libraries of digitized books gain critical mass, they increasingly demonstrate immense potential value as research platforms in numerous disciplines. To realize that value, we need a fuller understanding of the range of ways in which scholars conduct research, or want to conduct research, in the context of massive libraries of digital books.

Researchers using digital libraries must be able to find or discover, select, access or obtain, and organize collections of materials. Whether operating in digital environments or not, scholars tend to rely on collections. As a first step in their analytic work, scholars often gather bodies of evidence, which they perceive as relevant to their questions or topics of inquiry (Brockman, 2001). In each case, these bodies of potential evidence constitute research collections; yet collecting processes are highly idiosyncratic. Research collections take many different forms. Scholars collect a variety of materials, which they organize and analyze using different techniques. For example, one scholar may collect all of the books of a single author for close reading, supplementing them with the archival writings of the author, biographies, and other secondary sources. Another scholar may collect any and all texts in a particular language, from any point in history and regardless of authorship, for computational analysis of the gross. One scholar may work entirely with bibliographies, another only with page images, another with richly encoded TEI representations of poems within books, and a fourth with all of the above.

How can we anticipate various kinds of collection-building and -use, in order to support the diversity of researchers who work in libraries of digital books? This paper reports selected results of a study of how potential users of the HathiTrust Digital Library create and use collections in their research. This study was conducted by the Workset Creation for Scholarly Analysis (WCSA) prototyping project team of the HathiTrust Research Center at the University of Illinois, and the HathiTrust Digital Library.

It aims to further our understanding of scholarly collection-building practice as well as inform the development of a workset-creation tool for the HathiTrust Research Center, which will allow scholars to create sets of materials for analysis.

The study comprised a series of focus groups and interviews with humanities researchers, computational linguists, computer scientists, technologists, and librarians, in order to elicit their collection-building practices, transformation and preprocessing steps, and desired units and methods of analysis. The questions guiding this research include:

R1. What purposes and functions do collections serve for researchers?

R2. How do researchers create collections: What sources, tools, and techniques do they use for finding, identifying, selecting, and obtaining texts and other resources? What are their inclusion and exclusion criteria?

{This is the space reserved for copyright notices.]
R3. How could collections be used or reused for other purposes than researchers’ primary research?

R4. What are the units of analysis for scholarly research, and how do they inform collection-building activity? What kinds of entities (such as logical or physical representations of objects, features, and metadata) do scholars gather into collections as items?

R5. What challenges do researchers face in the course of collecting, from identification and selection to access and analysis?

We anticipate that the results of this study will prove relevant not only to ongoing and future development of bibliographic digital libraries, but also to digital collections and aggregations of other kinds, especially systems that aim to allow users to build their own collections of resources. This study aims to contribute to our broader understanding of scholarly practice, particularly scholars’ collecting activities.

Collections and Worksets in the HathiTrust Research Center

The HathiTrust Digital Library1 is a repository of over 11 million volumes, comprised of some 3.9 billion pages of text. This expansive aggregation of distributed sources serves as a corpus within which researchers can concentrate related sources into densely thematic bodies of evidence. The HathiTrust Digital Library provides bibliographic metadata for the entirety of its holdings as well as full-text access to the approximately 3 million volumes that have been released into the public domain. The HathiTrust Research Center2 (HTRC) is the research branch of the HathiTrust, offering scholars a suite of tools and services, which enable computational access to HathiTrust contents.

Scholars use bibliographic records from the HathiTrust Digital Library to select subsets of texts for computational analysis according to their particular research objectives. We refer to these subsets, along with associated, external data sources, as “worksets.” Worksets are a type of machine-actionable, referential research collection. User requirements for workset creation will grow increasingly sophisticated and complex as humanities scholarship becomes more interdisciplinary and more digitally oriented over time. Worksets will allow scholars to formalize their collections of resources from within and outside the HathiTrust, run computational analyses across them, and then reference persistently, and describe the data sets in disseminations of results.

WCSA is a two-year effort aiming to engage scholars in designing tools for exploration, location, and analytic grouping of materials so that they can routinely conduct computational scholarship. The three major goals of the WCSA project are to 1) enrich the metadata in the HT corpus, 2) improve access and discovery through referenceable metadata, and 3) formalize the notion of collections and worksets in the context of the HTRC.

LITERATURE REVIEW

Extensive work on the development, representation, and description of research collections to support scholarship informs our concept of collection in the context of this study (e.g., CLIR, 2010; Hill, et al., 1999; Palmer, et al., 2010; Wickett, et al., 2013; Sinn & Soares, 2014). The collections literature tends to focus on collections created by libraries, archives, and museums — that is, institutional collections. We distinguish institutional collections from those generated by scholars and other users for their own purposes. Rather than “information seeking contexts” (Lee, 2000), which is a characterization more relevant to institutional settings, we take a broader view of scholarly collections in this study, aiming to elicit a wide range of responses and avoid assumptions about what or why scholars collect for their own purposes. Palmer (2004) notes: “All collections, either physical or virtual, are formed through collocation”. Taking this as a starting point, our study sought to understand anything a scholar identified as a collection — any collocation of sources of evidence.

Humanities researchers are a key user group for this study. They are known to gather information from various sources as an essential, often preliminary step in the research process (Palmer & Neumann, 2002; Palmer, 2004; Palmer, et al., 2009; Sukovic, 2008; Sukovic, 2011; Toms & Flora, 2005; Toms & O’Brien, 2008). Indeed, thematic research collections are a “scholarly genre” on the rise: scholar-created research collections are likely to increase in number as the work of producing them becomes more widely accepted as legitimate scholarship” (Palmer, 2004).

Humanities scholars “build their own personal libraries to support not only particular projects but also general reading in their field”, largely out of a need for constant, convenient access to materials for rereading or analysis (Brockman, et al., 2001; Palmer, 2005). Palmer (2009) identifies “gathering” and “organizing” as primitives of the scholarly “collecting” activity. Scholars’ personal research collections include both primary and secondary sources, in numerous media and formats drawn from heterogeneous sources (Brockman, et al., 2001; Palmer & Neumann, 2002; Palmer, 2005). Mueller (2010) employs the metaphor of the library carrel to describe how digital humanities scholars collect texts and subsets of texts that are amenable to computational analysis. Indeed, a survey of scholars working with large-scale text corpora found that they want improved ways of finding and handling relevant subsets of the corpora:

“Researchers do not necessarily need huge sets of data to do interesting work, but the implication is that they do need flexible data delivery services that can deliver different kinds of data in different formats based on different searches for different

1 http://www.hathitrust.org/
2 http://www.hathitrust.org/htrc
kinds of research at different times.” (Varvel & Thomer, 2011).

Palmer (2004) provides a detailed exploration of thematic research collections, which are purpose-built for community use and support traditional humanistic methods of inquiry. The current study aims to build on Palmer (2004) with an empirical investigation of the details of collection-building practice, defined more broadly to include collections built for personal use and for computational analysis.

User-generated collections more generally have been treated in studies of how users retrieve and synthesize materials from digital libraries (Feng, et al., 2004); personal data collections (Beagrie, 2005); preservation of faculty-created digital collections (Beaudoin, 2011); collections of photographs on Flickr (Stvilia, et al., 2009; Rorissa, 2010); and in one study of journalistic research practice (Attfield & Dowell, 2003). Beagrie (2005) does note the high potential value of scholarly collections: “their importance for current scholarship is growing along with the power and reach of software tools and communications available to individuals to create, manage, and disseminate them.”

METHODS
This study is concerned with questions of how researchers create and use collections in the course of their research. While it is motivated by the work of the HTRC, with a focus on digital book and serials collections, we are also interested in how researchers incorporate non-digital and non-textual sources into their work. Since this study is largely exploratory, we used a qualitative methodology to answer the research questions stated above. Data were collected through a set of semi-structured focus groups and interviews, allowing moderators to modify the protocol to address topics identified as interesting. Interview and focus group protocols were developed, but researchers were flexible in adapting and probing participants for more detailed explanations. Participants also completed an optional demographic survey.

We conducted the focus groups and interviews at the 2013 Digital Humanities, Joint Conference on Digital Libraries (JCDL), and HTRC UnCamp conferences. Participants were recruited primarily by email. Emails were sent directly to individuals with known expertise or interest in focus-group or interview participation, based on conference attendee and speaker lists. Additionally, emails for mass participation were sent to relevant mailing lists prior to the start of each conference. Thirteen individuals participated in focus groups and five were interviewed individually for a total of eighteen participants.

We recorded, transcribed, and coded the focus groups and interviews using ATLAS.ti 7, and used qualitative content analysis (Zhang & Wildemuth, 2009). Preliminary codes were developed inductively based on themes identified in the raw transcripts. A codebook was developed and the codes were tested on a subset of the data and assessed for consistency. Two separate people coded each transcript to ensure consistency and thematic coverage. For each code group, a separate researcher (i.e., not responsible for coding) was responsible for drawing final conclusions from the coded data. Overall, 28 codes were created in six code groups with five coders. Only results from a subset of these code groups are reported here.

ANALYSIS
In this section we report selected results of the study, elaborating on three themes identified as prominent in our preliminary findings (Green, et al., 2014): (1) roles of collections; (2) units of analysis; and (3) metadata enrichment.

Participant Demographics
The study comprised eighteen participants who either conduct or support research in academic or other research environments across a variety of disciplines and institutions. Specific fields include English, classics/philology, digital humanities, computer science, information science, computational linguistics, and librarianship. Fourteen participants voluntarily reported demographic information, which is summarized in Table 1. Participants were predominantly from the United States (64%), male (57%), and academic faculty (64%). The average participant age is 50 with an average of 14 years of experience conducting related research. We also asked participants to specify their research interests, which are summarized in Figure 1. Note that participants were allowed to identify more than one area of interest.

<table>
<thead>
<tr>
<th>Nationality</th>
<th>US 64%</th>
<th>German (2), French (1), Italian (1), Indian (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male 57%</td>
<td>Female 43%</td>
</tr>
<tr>
<td>Position</td>
<td>Professor 64%</td>
<td>Full (5), Associate (2), Assistant (1), Emeritus (1)</td>
</tr>
<tr>
<td></td>
<td>Other 26%</td>
<td>Engineer (2), Librarian (1), Scientist (1), Student (1)</td>
</tr>
</tbody>
</table>

Table 1. Demographic information

| Avg. age (yrs) | 50 Min=26, Max=74 |
| Avg. experience (yrs) | 14 Min=2, Max=30 |
Roles of Collections
A central theme of the focus groups and interviews was the role of collections in scholarly research. Collections serve a number of purposes for the people creating and using them. Our results suggest partial answers to research questions R1 and R2. The most prominent roles that came to light in our findings include:

- Collections serve as research products in their own right
- Collections serve as reusable resources

Collections as research products
Scholars build collections to serve as platforms for analysis; yet collections themselves are often products of intensive analytic work and constitute, for some, a significant portion of their research labor. Scholars view the collections they create as research products in their own right, meriting credit and citation, in much the same vein as their publications. One respondent acknowledged a trend toward “the valorization of corpus-building...The recognition at the scientific level” (P10). Respondents shared the view that collections as research products warrant broader recognition by the scholarly community at large, in tenure and promotion processes, in grant funding processes, and in other evaluations of humanities and linguistic laboratories or research groups.

Several reasons emerged from responses, for understanding collections as a scholarly genre, in the terms of Palmer (2004):

1. Collections are investments of substantial scholarly work;
2. Collections should be shareable and citable for the sake of reproducibility of analytic results; and
3. Collections have reuse value for scholarly communities at large, to be discussed in the next section.

On (1) Collections as investments: It is well known that humanities researchers invest significant time, effort, and money into the creation of collections for their own use (Palmer, 2004; Brockman, et al., 2001). Our findings shed light on how scholars’ sense of investment contributes to their increasing desire to view collections as research products.

Respondents reported that the work of finding, evaluating, selecting, and acquiring items to gather into collections often takes years. One respondent described the process as one of purposive, long-term accretion upon a core of primary sources:

“Here’s the core historical text, and then the papers and the backstory just build that core text, and then building out from there to support all of the other literature coming up to it...The other writing: all of these authors—what did they write? Pulling those into the collection. Pulling into the collection the works they’re responding to. So it’s kind of starting with a core of texts and networking out. And then networking forward to look at some of the follow-on publications” (P15).

The investment described here lies in the time spent seeking and acquiring relevant items. Some scholars even employ teams to help with this aspect of collection-building. Respondents described gathering or purchasing different kinds of materials from a great variety of sources, including digital libraries (such as HathiTrust, Google Books, library collections as cataloged in online public access catalogs; known archives; content providers that publish well-known digitized collections like Early English Books Online); other scholars’ personal collections; and eBay and other retailers.

Some respondents even described collecting practices that entail personal risk, as they collect in a grey area of legality. One respondent acknowledged illicit digitization work: “I’ve seen and been affiliated with places that literally took bulk published books within the last year and scanned all the images out of them and put them on the web. Those projects aren’t currently active” (P15). Another described secretive, peer-to-peer exchanges of texts and other materials: “It’s important that you talk to people and they say, ‘Oh I’ve got ten, fifteen, I’ve worked on them and I can give you them. Please tell no one’...[T]hey are an important source” (P13).

The gathering alone is labor-intensive, but the investment also lies in the interpretive work of determining the relevance of an item to a growing collection. Highlighting the extent to which this gathering and assessment work is considered a scholarly contribution, one respondent noted that a comprehensive collection of electronic editions was an intended outcome of her dissertation work: “That’s another part of my PhD, is creating a catalog of what’s already been done – electronic editions, it’s just gathering” (P12; emphasis added).

There is an acknowledged tradeoff between cost and quality of a collection, which alone is evidence that collections are valuable investments to the scholars creating them. Some scholars are actively involved in digitization and consider it a phase of their collecting work: “we paid quite a bit of money for a vendor to double-keyboard and produce good TEI text” (P8). Others are more opportunistic in their gathering: “whatever I find I include and I actually make just a distinction between what is what...The aim is to get as
much data in there as possible. I’m not really picky about what I include.” (P17).

Yet the act of gathering is not the end of the collecting process. Many scholars, especially those employing computational methods of analysis, must preprocess the texts and other information they gather before they can begin any analytical work. One respondent acknowledged that, “fixing the data, preprocessing it in all kinds of ways, is always like 85 percent of what you’re doing” (P18). Another noted that much of the labor of collection-building for computational analysis is now devoted to resolving technical problems of access and data processing, yet voiced optimism that technical barriers will erode over time, and that the curatorial work of selection will become more central to the process: “Today it is viewed as something very technical to prepare a corpus…one day, it will be unrelated to technical stuff, and it will get closer to something of value” (P10).

On (2) Collections as support reproducibility: A second factor in how collections serve as scholarly research products is the growth of humanities data. As scholars increasingly employ computational methods to texts and other sources on a large scale, there is a growing sense that their bodies of evidence must be treated on a par with scientific data sets. One respondent argued that, “we must be able to relate publication to corpus” (P10). Respondents suggested that ensuring the citability of research collections and data sets was particularly important at this moment in the evolution of digital humanities research, as scholars try to justify the use of novel humanities methods, or demonstrate their accountability:

“[I]f you just say, I have a corpus and nobody is allowed to see it but wonderful things come out of it… That’s not really research. That is a problem here, I think, for us, because we are just starting with this kind of work. We are trying to get accountability for the kind of work we are doing. And it’s important for us to show the basis our work” (P13).

Additional motivation for viewing collections as research products stems from their reuse value to other scholars.

Collections as reusable resources

Responses suggested that collections could and do facilitate scholarly communication in a number of ways, and thus have real and potential value for reuse. Our results suggest a partial answer to research question R3.

Respondents saw user-generated collections as potentially valuable for helping users navigate massive, overwhelming information spaces. One respondent, for example, suggested that “[p]eople are building subcollections”, and that these curated collections on specific themes have the potential to remedy this problem: “you come to a very broad, very general site like [the HathiTrust], and it’s hard for somebody who’s interested in a particular domain to go in and find a subfield within that broad space” (P15).

This is true not only for the collections people create for their own use, but also for collections that others have created. The same respondent suggested that his research team hoped to build upon the value of the collecting work done by another scholar, who

“…spent the last seven years combing through Google Books, finding books, building up a catalog, and downloading the PDFs. There’s a lot of value in that, but we can’t really access and represent that value using the standard APIs that Google provides very easily and build value-added on top of that” (P15).

As scholars – especially those working online – seek and discover items relevant to a specific theme or inquiry from different corners of the Web and from within digital libraries, they create navigational resources of potential value to other researchers pursuing related lines of inquiry, effectively by gathering a set of pointers to relevant items. One respondent suggested that, while gathering and organizing were imperative to the analytical process, the resulting collection was intended as a contribution to the scholarly community: “I’m learning a lot through this organizing of my material and it’s informing what will be the main argument of my research … At the same time I think, I believe, it would be something useful for the whole digital humanities, electronic editing community” (P12).

One respondent noted that a collection garnered interest from other scholars, not as much for its content as for the technical methods by which it has been developed and maintained: “I hear people who are interested in the history of women are interested in using my bibliography. I get queries and people use it. Not a lot. Frankly. No, I get more interest the other way, which is [about] our approach” (P8).

The most common reuse of research collections reported in our responses was for teaching. Scholars reuse their own research collections to teach undergraduate and graduate students: “the corpus I built, I’m giving it to students in classrooms so we can work together” (P13). In addition, certain collections are repurposed for teaching in primary and secondary educational contexts, though this seems to be more often the case with library or institutional collections rather than scholars’ personal research collections. Respondents also suggested that some collections may have reuse value for the general public, or at least communities of niche interest outside of academia, citing the example of the reuse of an archival or historical collection for genealogical research: “[genealogists, people who want to study family histories and such, … [this] is probably one of the most active sources of interest among the general public.” (P2).

Units of Analysis

Units of analysis are the actual targets of scholars’ analytic work: what kinds of things they aim to study, which
Table 2. Reported units of analysis

correspond directly to the kinds of things they aim to collect. Our results suggest a partial answer to research question R4. Units of analysis described by respondents varied widely in both abstraction and granularity. For example, one respondent noted:

“It is very essential to work at the level of a particular chapter, with the actual text...We cannot talk so meaningfully about the work of a writer as a whole, in the abstract. The interpretation is based on actual text, at smaller units of analysis” (P7).

Table 2 lists in alphabetical order the various units of analysis enumerated by respondents. The list is not intended to be comprehensive or representative of the field, but rather to convey the wide range of analytical approaches that researchers are taking. The units of analysis researchers employ differ in granularity and level of abstraction.

Units of analysis may be amorphous. One example is the narrative theme as a unit of analysis. One respondent studied recurring narrative threads across a corpus, specifically the narrative theme of “secret marriage,” which recurs in the novels of some women writers of the period of study. The respondent hoped to comprehensively collect novels manifesting that theme for close reading.

Objects of analysis may also constitute well-defined units within the hierarchy of a text, such as paragraphs or individual words. Scholars may analyze a more concrete unit as a proxy for another, more amorphous unit. One respondent, interested in comparing the “basic plots” of women’s lives as depicted in different biographies, broke texts for study down to the level of individual paragraphs, describing the process as such:

“Paragraph 1, we tag whether such and such occurs in it. So, is this kind of event in it? You know, what do we think of as the basic plot of it? Where do we see the use of such-and-such a kind of metaphor?...So it’s a kind of paragraph-level, to the point where [for] fourteen versions of the same woman’s life we can kind of say, Oh do they [the biographers] use that event?” (P8).

Units ranged widely in levels of abstraction, from parts-of-speech trigrams to poems within books, to encoded TEI elements, and whole lexicons. One respondent described his research work as traversing several levels of abstraction, and entailing analysis at multiple levels of granularity:

“We have words, text units, and intermediate structure. Those three levels hold different types of properties. For words, it’s lemmas, categories, things like that. For intermediate structure, it depends on the structure, the type of text — and the texts have metadata, which are usual metadata for text units. Those three units of the target we search in TEI stuff. And it can have a lot of different ways of being expressed, those three levels. Those are the textual levels, and semantically we try to develop textual planes, where a scholar can search in different planes” (P10).

An object of analysis may not always consist of contiguous chunks of text, but may contain text culled from different parts of a collection and reassembled. Researchers want collections, worksets, texts, and other objects of analysis to be highly divisible, and prefer the resultant pieces to be identifiable, movable, and readily associable with highly granular metadata. A respondent spoke of needing different “ways to slice this book”:

“So we need to slice it by page. That’s pretty natural. We need to slice it by poem, which doesn’t conveniently overlap or match the page boundaries. We potentially need to slice it by sections within a poem, so if there’s a section that’s all sonnets, followed by a section that’s all free verse... ...We actually need to slice these works in fairly fine-grained ways, depending on what particular analytical approach we’re looking for” (P15).

Another respondent observed that this sort of recombination occurs throughout the scholarly research process:

“[W]ords are the main entry point for those scholars, and they use a lot of corpus configurations, like subcorpora...And partitions-building. Partition[ing] is to slice the corpus in parts, the sum of which is the whole. So this is for contrastive analysis” (P10).

Some objects that scholars work with are themselves composites of other potential analytic units; in these cases, the complexity of analysis may be compounded. Page images, which are often made available in libraries of digitized texts, are themselves complex, compound objects, each piece of which is some scholar’s preferred unit of analysis. A respondent described “a project to develop tools for the computational analysis of the visual elements of digitized page images” (P18).

The extraction of different units for analysis is a known problem for scholars. Analyzing words from raw text requires high-quality OCR, the lack of which was a
frequently reported challenge among respondents. Analyzing paragraphs or other elements of the hierarchical structure of a text may require high-quality markup (which itself may depend on high-quality OCR). As one respondent noted:

“[T]he preexisting collections where [consistent encodings] exist are fantastic…People… who work with XML-encoded texts [say] that everyone aims for the same TEI standard and everyone hits it differently… [If] you could get volumes from a single source that are consistently encoded, it makes things like metadata extraction much easier” (P9).

Metadata Enrichment
Scholars use item-level metadata to identify and assess items for inclusion in their collections. As such, the quality and extent of metadata available to scholars is vital to the collection-building process. In digital scholarship, metadata may extend beyond the volume to deeper levels of enrichment, including scholarly annotations, markup, and encoding. Data enrichment activities span the entire spectrum of scholarly units of analysis described above. Within the context of the HathiTrust Digital Library, volumes are currently described exclusively by bibliographic records imported from library catalogs. Participants emphasized the limitations of relying on volume-level cataloging records. As one scholar states:

“[T]he book is not a unit of great interest in many cases. In many cases you want to get all the poems that are quoted in the book that are not listed in the metadata. Or whatever of interest to you. So the metadata from the library is very coarse” (P17).

Indeed, one of the most prevalent themes arising from this study is the researchers’ need for more extensive and more accurate metadata at varying levels of granularity. This result suggests a partial answer to research question R5. Strict adherence to bibliographic cataloging standards does not afford ways to record the many attributes and properties necessary for identifying and selecting items for thematic scholarly research collections and computational research. Despite perceived deficiencies in available metadata, researchers have also made use of bibliographic metadata as a rich resource in its own right for scholarly analysis. One scholar acknowledges that access to metadata will always be less restricted than access to full text, and further emphasizes:

“[T]here’s a lot you can get just from the metadata. Franco Moretti has done interesting work just on titles. But when you have titles and publication places and publication dates and author information --- there’s a whole bunch of … potential network analysis stuff to be done there, even without having any access to the underlying texts” (P9).

Types of enrichment
Throughout the focus groups and interviews, researchers identified 15 different types of data enrichments required for conducting research with digital collections (see Table 3). The greatest emphasis by far was on the importance of genre identification (e.g., fiction, poetry, and biography). A scholar interested in poetry acknowledged that “genre is not something that is always explicit. It’s usually not explicit in bibliographic records, so we have to find it in different ways” (P18).

After genre, the most commonly listed attributes included authorial gender, important dates related to the author’s life span, and a book’s initial date and location of publication. Other examples of useful attribute additions cited by participants include circulation data, changes in authors’ names and institutional affiliations over time; lists of frequently named entities within the text; and related editions, such as reprints, previous or subsequent editions, or even unauthorized pirated editions. One respondent succinctly expressed the combined importance of these additional attributes for the process of scholarly collection building as follows:

“You really do want to be able to say ‘give me all the poetry written by women, written between 1850 and 1880 published in Britain.’ And at the moment, that’s not something you can do” (P9).

While bibliographic records indicate the date that a particular copy of a book was published, several scholars expressed a preference for the date the text was originally written or the work was first distributed. Relying on current standards can hinder both the initial identification and selection of texts based on scholarly parameters and the textual analysis that follows. For example, one participant laments that according to bibliographic records “Cicero writes in the 19th century” (P17).

<table>
<thead>
<tr>
<th>Author gender</th>
<th>Integrated tools and methods of exploration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author names and affiliations</td>
<td>Languages</td>
</tr>
<tr>
<td>Data quality</td>
<td>Links to other sources</td>
</tr>
<tr>
<td>Dates</td>
<td>Metadata correction</td>
</tr>
<tr>
<td>Document format</td>
<td>Publication and circulation data</td>
</tr>
<tr>
<td>Encoding and typing and different levels of analysis</td>
<td>Related editions</td>
</tr>
<tr>
<td>Entities within text</td>
<td>Scholarly annotations</td>
</tr>
<tr>
<td>Genre</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Desired kinds of metadata enrichment
Even in cases where cataloging standards provide for representing information like dates and languages, the recorded information may not be correct. One focus group member cites an evaluation conducted comparing languages identified through full-text analysis with how language is represented in library catalogues:

“There were like 24,000 books listed as Latin. About 20-30% of them were not Latin. They were in some other language. For various reasons they were improperly cataloged. And then there were about 3,000 books that were Latin that were not listed” (P17).

A basic, yet necessary, form of enrichment is simple correction of inaccuracies in current bibliographic records. Other forms of enrichment extend beyond the correction and addition of simple attributes. Metadata enrichment is frequently the product of scholarly interpretation and other forms of intensive intellectual labor, such as literary and linguistic annotations or hand-encoded editions.

Scholars also express interest in making use of emerging technologies to embed links to other authoritative sources into cataloging records, integrate alternative tools and methods of exploration with existing bibliographic records, and address issues specific to digital texts such as determining data quality (e.g., confidence rating to identify the best OCR version of a given work or page image quality of a digitized book).

Scholars are already implementing the forms of metadata enrichment that we outlined above, to serve different research goals. One participant noted that it is “important...to be able to access the metadata of a library in a bulk version so that I can download it and process it myself” (P13). This sentiment was also reflected by several other study participants. Scholars also shared and discussed their own tools and techniques for metadata enrichment, which included information retrieval methods, subject visualization, relationship clustering, MySQL schemas for supporting new metadata, linked data approaches to external authority files, and tools like MorphAdorner3 for linguistic annotation.

**Enrichment challenges**

Over the course of conversations with participants in focus groups and interviews, five core metadata enrichment challenges emerged:

1. Contributing scholar-enriched metadata back to cataloging systems.
2. Tracking provenance of enriched metadata.
3. Ensuring metadata interoperability across standards and domains.
4. Accounting for variability across standards implementations.
5. Representing more granular forms of data enrichment in traditionally formatted bibliographic records.

Scholars interviewed as part of this study expressed a keen desire to share their metadata enrichments with the broader community. Yet establishing a metadata feedback framework that provides a context for scholarly contributions remains an unprecedented innovation within large-scale digital library systems. In response to one participant’s preference for bulk download of bibliographic metadata, another participant argued:

“So one thing is getting the data out. But then the next step is, you’ve done all this work, and you then have the authoritative metadata. You have the best metadata in the world, and no one will take that from you” (P17).

Other scholars acknowledged the importance of having a workflow to correct metadata inaccuracies and to contribute scholarly metadata enhancements back to the broad, shared infrastructure of digital libraries. If digital libraries incorporate frameworks for metadata enrichment from a variety of sources, tracking provenance of newly enriched metadata will become an essential aspect of data enrichment activities. This may extend as far as identifying the source of the enrichment, the warrant upon which any given enrichment is based, and a confidence rating for the reliability of the proposed change.

Within the context of multi-disciplinary, tera-scale digital resources, the challenge of interoperability across various standards and domains is inevitable. While such challenges can be extrapolated across a number of cases, one participant reflected on the implications for using museum metadata in other modes of scholarly inquiry:

“[T]he metadata is rather extensive but it’s very museum-centric. It’s mostly accession number and, you know, the standard thing you would expect walking into a museum. Artist, title, everything. But we are looking at using these in other disciplines so what happens if you want to use this in architecture. Then you need to express a completely different kind of structure” (P3).

Even within accepted disciplinary or cross-disciplinary standards, the level of metadata implementation often varies significantly. One participant specifically referred to the multitude of customized implementations within the context of the Text Encoding Initiative (TEI), saying “we can't read TEI because nobody can read TEI” (P10).

As long as digital library systems like the HathiTrust rely on bibliographic descriptions contributed from library catalogs, a core challenge will be mapping more granular metadata

---

3 [http://morphadorner.northwestern.edu/](http://morphadorner.northwestern.edu/)
enrichments (based on scholarly units of analysis) to traditional catalog records.

CONCLUSIONS AND NEXT STEPS
Data from the focus groups and interviews indicate that scholars perceive of collection-building as a vital scholarly activity. Their collections are research products in their own right and hold high reuse potential for the community. Their collections are highly heterogeneous: the units that scholars aim to collect and analyze are wildly divergent in granularity and level of abstraction. It is clear that it is not enough for a bibliographic digital library to simply allow its users to collect books; they want to collect different parts and pieces of books, and gather them along with items from outside of the digital library. Respondents express an ardent need for more and better metadata that transcend the conventions of the bibliographic record. They also expressed willingness to help create and share that metadata. Future work will target metadata enrichment to help accommodate researchers’ demands for varying units of analysis and collection. Particularly for research done in a non-consumptive context, in which metadata is the main entry point to the corpus, high quality and flexible metadata is imperative for scholarly research. Our findings suggest that scholars regularly enrich data in the course of their analytic work; harnessing these vital contributions to metadata enrichment as a component of scholarly discourse is key to maintaining relevant digital resources that serve researchers’ needs.

These findings inform ongoing development work as the Workset Creation for Scholarly Analysis project seeks to implement a workset-creation tool that will allow HTRC users to create a very specific kind of collection, one that is immutable and machine-actionable, to wield in their computational analyses. There is a tension between the mutability of scholarly collections, which grow and change over time, and the growing need for reproducible collections to undergird computational work. Scholars working with computational methods increasingly pursue the scientific gold standard of reproducibility, not only to justify computational work in the humanities, but also to ensure that collections are citable for credit in evaluative processes.

The next phase of the WCSA project will draw upon findings from focus groups and interviews to develop a functional data model for worksets in the context of the HathiTrust Research Center. As a compromise between the empirical mutability of the scholarly collection and the need for persistent datasets, we propose the workset as a snapshot of a scholar’s thematic research collection. The workset would capture the content of a collection at the point in time when that content was subject to computational analysis. A single scholarly collection may be the basis for multiple worksets throughout a career. By its persistence, the workset would become a shareable research product that can be cited and offered to colleagues interested in reproducing the results of a study. We will continue to investigate how to leverage the informative potential of scholarly collections as valuable, reusable research resources.

ACKNOWLEDGEMENTS
The authors gratefully acknowledge the Andrew W. Mellon Foundation for generously funding the Workset Creation for Scholarly Analysis Project; the HathiTrust Research Center, the Center for Informatics Research in Science and Scholarship, and the University of Illinois Library and Graduate School of Library and Information Science for institutional support; and WCSA co-PIs Tim Cole and Beth Plale for project leadership. The authors would also like to thank all who participated in the focus groups and interviews.

REFERENCES


The columns on the last page should be of approximately equal length.