User Navigation in Large-Scale Distributed Digital Libraries: The Case of the Digital Public Library of America

Krystyna K. Matusiak
University of Denver, krystyna.matusiak@du.edu

Follow this and additional works at: https://digitalcommons.du.edu/lis_facpub

Part of the Library and Information Science Commons

This work is licensed under a Creative Commons Attribution 4.0 License.

Recommended Citation

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by the Library and Information Science: Faculty Scholarship at Digital Commons @ DU. It has been accepted for inclusion in Library and Information Science: Faculty Publications by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu,dig-commons@du.edu.
User Navigation in Large-Scale Distributed Digital Libraries: The Case of the Digital Public Library of America

Comments
This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of Web Librarianship on October 18 2017, available online: "http://www.tandfonline.com/10.1080/19322909.2017.1356257".

Publication Statement
Copyright held by the author or publisher. User is responsible for all copyright compliance

This article is available at Digital Commons @ DU: https://digitalcommons.du.edu/lis_facpub/7
ABSTRACT:
This article presents the findings of a case study that examined user navigation in a large-scale digital library in the context of academic use. Using the Digital Public Library of America as case, the study explored user navigation and understanding of a distributed model of large-scale digital libraries. The DPLA model involves two- or three-step navigation pathways. Most study participants could navigate the DPLA distributed, multi-layered system effectively. This study found some confusion among the participants when they had to move through a three-step process to locate digital objects provided by a metadata aggregator service hub. The study participants also pointed out the advantages of a distributed model in collocating digital resources and connecting users to a wide range of digital libraries.

INTRODUCTION
Two decades of extensive digital library development and mass digitization have resulted in a critical mass of digital versions of books, journals, and primary sources. Until recently, however, resource discovery in the digital library (DL) environment was difficult, as it required locating and searching individual digital libraries and collections. Despite a great selection of digital resources, the awareness of standalone digital libraries (DLs) and their adoption for educational and personal use has been limited (Liu and Luo 2011; Matusiak 2012; McMartin et al. 2008, Singh, Sharma, and Singh 2015). Large-scale digital libraries (DLs) represent the next step in digital library development by providing a single access point to and the ability to search across a multitude of scientific and cultural heritage collections. Large-scale distributed systems, such as the Digital Public Library of America (DPLA) or Europeana, gather metadata from individual DLs or other aggregators and offer a central portal for searching and linking to digital objects.

Large-scale distributed DLs, however, are a relatively new phenomenon and, like many emergent information systems, face challenges in regards to usability and acceptance (Dobreva and Chowdhury 2010). Very little is known about user interaction with large-scale DLs and the ways users search and locate digital objects in a distributed network. This paper presents the findings of a study on user navigation through a large-scale DL system, using the DPLA as case. The purpose of this study was to examine user understanding of a large-scale distributed DL system and user navigation through a multi-layered structure. The focus of the study was not only on user search behavior, but also on pathways in locating and retrieving digital objects. This study was exploratory in nature and adopted a qualitative research strategy with direct observations and interviews as its primary data collection techniques.

LITERATURE REVIEW
Large-scale digital library systems connect smaller individual DLs and enable global searching and retrieval. The aggregation of content and services can take place on consortial, regional, national, or international levels. These large-scale DL systems are built either as centralized aggregators of content, metadata, and services, or use a distributed network model with repositories of harvested metadata and a service layer to facilitate metadata searching and
connecting to content from originating sites. (Henry 2012; Xie and Matusiak 2016). HathiTust is an example of a centralized repository with millions of full-text digital volumes contributed by partnering academic and research libraries and a single user interface for access and discovery (Christenson 2011). The DPLA, Europeana, and the National Library of Australia’s Trove represent large-scale distributed systems aggregating metadata from cultural heritage collections on a broad national or international scale. The National Science Digital Library (NSDL) is also built on the metadata aggregation model but with a focus on a specific domain, harvesting metadata for educational resources in science, technology, engineering, and mathematics (STEM) (Lagoze et al. 2006).

The Digital Public Library of America (DPLA)
The DPLA represents the latest effort in the development of large-scale DL systems and has an end goal of evolving into a United States national digital library (Cottrell 2013; Howard 2013; Paulmeno 2015). The site was launched in April 2013, and as of February 2017, provided access to over 15 million digital objects from a wide spectrum of collections in the United States (DPLA 2017). The DPLA aggregates metadata and offers a single user interface for searching and accessing millions of cultural heritage materials held at libraries, archives, museums, and other cultural heritage institutions in the United States (Scardilli 2014).

The DPLA gathers metadata from participating institutions and provides links to digital objects stored and maintained on its partners’ servers. Metadata interoperability is key to building the DPLA repository, where records come from a variety of sources and are often based on different schemas (Sandy and Freeland 2016). Participating institutions are involved in preparing metadata for ingest and mapping to the DPLA’s metadata application profile to ensure interoperability (Gregory and Williams 2014; Sandy and Freeland 2016). This process enhances the quality and consistency of harvested metadata.

The structure of the DPLA is built on a distributed model consisting of content and service hubs (Piper 2013). Content hubs, large institutions such as Harvard University, The New York Public Library, The National Archive, and The Smithsonian, contribute more than 250,000 metadata records directly to the DPLA’s platform. Service hubs are state or regional aggregators that gather metadata from local digital collections and archives, and in turn provide them to the DPLA. At the time of its launch the DPLA partnered with six service hubs, including the Digital Library of Georgia; the South Carolina, Kentucky, Minnesota Digital Libraries, the Mountain West Digital Library, and the Digital Commonwealth. Since then the number of service hubs has grown, ensuring more comprehensive coverage of the United States. The DPLA provides a simple search interface, consistent metadata records, and search refinements through facets and visualization tools. The distributed model, however, requires end users to traverse two or three layers in order to locate actual resources.

Use Studies of Large-Scale DLs
The emergence of large-scale DLs has dramatically changed the information landscape with major aggregators of metadata, such as Europeana and the DPLA, gaining more visibility. Research on the use of large-scale DL systems, however, is still very limited. Use studies have been conducted primarily with the NSDL and Europeana. Both systems represent early large-scale DL developments. The NSDL was designed in the early 2000s to provide students and teachers with a central access point to a wide range of teaching and learning materials in the STEM disciplines (National Science Digital Library 2017). Europeana was launched in 2008 as a
portal to digitized cultural heritage materials from European museums, libraries, archives, and audio-visual collections (Europeana 2017; Purday 2012).

The NSDL, originally funded by the US National Science Foundation, served as a testing environment for a metadata aggregation model (Lagoze et al. 2006). A number of studies were conducted with educators to evaluate the functionality of the NSDL, its support for designing learning activities, and to examine usage patterns (Recker 2006; Recker et al. 2007; Xu and Recker 2012). More recently, Zavalina and Vassilieva (2014) analyzed search queries in the NSDL and Opening History (OH), an aggregator of digital collections with a focus on US history and developed through IMLS National Leadership Grants. The authors found that user searching differed between those two domain-specific DLs in regard to query lengths, frequencies, and search categories. The search queries of the NSDL users varied in length and occurred more frequently, while OH users employed more search categories. The study also reported a high level of advanced searching, especially in the subject categories. Zavalina and Vassilieva (2014) emphasize that large-scale DLs need to consider the patterns in user information-seeking behavior in order to improve user interaction and to facilitate efficient information retrieval.

User studies of Europeana offer insight into user needs and perceptions, search behavior, and navigation using mobile devices. An early usability study of Europeana was conducted with 89 participants in four European countries (Dobreva and Chowdhury 2010; Dobreva et al. 2010). Dobreva et al. (2010) report that a younger generation of participants tends to prefer search engines, like Google, or general websites, like Wikipedia, to curated digital libraries. The authors also observed that participants from different countries varied in their perceptions and expectations of Europeana. Dobreva and Chowdhury (2010) focused on findings related to the usability and functionality of Europeana. They noted that despite initial positive perceptions of Europeana, several participants encountered a number of usability challenges when interacting with the system. The difficulties were related to multilingual content of Europeana as well as limited abilities to prioritize and refine search results. Some users encountered broken links when connecting to digital objects at original sites. Nicholas et al. (2013) compared desktop and laptop users of Europeana to those who were accessing the site using mobile devices. The authors found that mobile users were the fastest-growing Europeana group and demonstrated different information seeking behavior. Mobile users engaged in shorter visits and viewed less content.

In comparison to Europeana or the NSDL, the DPLA is a new and evolving digital library system; research examining user information seeking and the use of the DPLA is still in the exploratory phase. Recent reports indicate that DPLA participating institutions are beginning to see an increased usage of their digital collections as a result of referrals from the DPLA portal (Biswas and Marchesoni 2016). However, there is no empirical research on user interaction with the DPLA. This study contributes a user perspective on the navigation through the DPLA’s distributed system.

METHODS

Large-scale DLs address the issue of resource discovery in the first phase of information seeking. However, relatively little is known about user navigation through large-scale DLs beyond initial query formulation. Similar to searching with the Google search engine, users may find thousands of hits from a search, but the path to finding its results is different. The structure of distributed large-scale DLs is built on a model of multiple layers that users need to navigate in order to find digital objects at originating institutions. The purpose of this study was to examine user understanding of large-scale DLs and to explore how users navigate this multi-layered
environment to find resources that they can apply in scholarly and educational practices. The following research questions were posed for the study:

RQ1: What is the nature of user experience in interacting with large-scale DLs such as the DPLA?
RQ2: How do users understand and navigate the DPLA?
RQ3: How can the DPLA support teaching and learning in higher education?

This paper reports the findings related to the first and second research question. This study used the DPLA as a case and selected faculty, undergraduate, and graduate students as study participants. The DPLA was selected purposefully as an information-rich case and a primary manifestation of a large-scale, distributed DL. This study was designed as a qualitative case study with an emphasis on users of digital resources and their experience in the process of navigating a large-scale DL. The case study approach allowed for an in-depth investigation of the phenomenon and helped to establish the boundaries for research (Creswell 2013; Yin 2013).

Data Collection Techniques and Analysis
Multiple qualitative data collection techniques were used to provide a thorough examination of the phenomena under study and to ensure standards of credibility and trustworthiness. Triangulation was achieved by using multiple sources of evidence. The data collection techniques included:

• Questionnaire administered at the beginning of the observation session to gather demographic data and information about participants’ needs and prior experience with DLs.

• Direct observations were used to provide data on user interaction with the DPLA and navigation paths through the distributed network. The participants were presented with two predefined scenarios, but were also asked to conduct two additional searches on the topics of their choice. Appendix A provides a copy of the observation script and scenarios.

• Interviews were conducted after observation sessions to record participants’ reactions about the nature of their experience with the DPLA and to explore in more depth their understanding of large-scale DLs and their preferences for information searching. A semi-structured interview guide was used for interviews. Interviews were recorded and transcribed.

Data were collected over a period of three months in the fall of 2015. During observation sessions, participants were asked to complete the scenarios from the point of performing a search, finding a resource, to downloading it, if possible.

At the beginning of observation sessions, participants were prompted to conduct searches based on predefined scenarios. These task-oriented scenarios provided examples of realistic activities that users would undertake in their teaching or learning practices, such as locating an oral history recording to be presented in class or finding images to include in a paper or PowerPoint presentation. Appendix A outlines the predefined scenarios used in this study.

Next, the participants were encouraged to search on two topics of their own interest. The topics selected by the participants varied widely. Some examples include: Frank Lloyd Wright, Native Americans, World War II posters, photographs from southern China, anti-war activism, and African American clergy. A laptop computer was used during observations. The sessions were recorded with Camtasia, a screen video capture software produced by TechSmith that records on-screen video activities plus accompanying audio. This software allowed the
researcher to capture participants’ comments and navigation pathways, the goal of which was to gather data on users’ navigation through the layers of the DPLA.

Data analysis was conducted primarily as qualitative content analysis that involved open coding and discovering categories, patterns, and themes (Corbin and Strauss 2008). A codebook was developed based on the initial open coding and was used systematically to code data gathered through observations and interviews. NVivo software was used in coding and further data analysis. Quantitative data was recorded on the completion of scenarios.

Participants

The study employed a purposeful sampling strategy. It was conducted at a private, mid-size university in the United States with a large graduate student population. The participants were recruited from the social sciences and humanities programs by posting flyers in the library and campus public spaces. Social sciences and humanities educators and students were selected because of the high concentration of primary sources and cultural materials in the DPLA. The sample size consisted of 21 participants, including two faculty members, six undergraduates, and 13 graduate students. Among the graduate students there were five PhD and eight Master level students. Fifteen participants identified themselves as female and six as male. Table 1 provides a summary of the participants’ backgrounds.

Table 1. Participant status and programs

<table>
<thead>
<tr>
<th>Participant Status</th>
<th>No.</th>
<th>Academic Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>2</td>
<td>History; Art History</td>
</tr>
<tr>
<td>Graduate students</td>
<td>13</td>
<td>Art History; Communication; Education; English; Emergent Digital Practices; International Studies; Religious Studies; Spanish</td>
</tr>
<tr>
<td>Undergraduate students</td>
<td>6</td>
<td>Art History; Biology; Computer Science; Emergent Digital Practices; History; International Studies</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

FINDINGS AND ANALYSIS

Nature of User Experience: Resource Discovery and Usability

Most study participants were not familiar with the DPLA and used it for the first time for this project. Four participants heard about the site, but only two had actually used it prior to the study. One student reported that he was introduced to the DPLA in class and used it for a class project. Although many participants were first-time users, they understood quickly that the DPLA represented an effort to bring resources from multiple institutions. A phrase of “one-stop
shop” appeared as a reoccurring theme in many comments, as exemplified by Participant M, “I actually appreciate the one-stop shop where you don’t have to go to twenty different sites to find something. So if there’s a collaboration or partnership where it’s feeding into that one portal, I think that’s helpful.”

The participants started their resource discovery on the DPLA home page with a basic keyword search by typing terms in the search box, such as ‘railroads maps’ or ‘civil rights activist interview’. Next, they used the “Refine Search” feature on the results page to narrow down their search (see Figure 1). Almost all participants, except for one international undergraduate student who did not notice it, used this faceted “Refine Search” feature. The participants liked having the ability to narrow down their searches, as noted by Participant D: “Once you start searching, it was nice to put those like refining filters on it, I mean, it made it really easy.” The following facets were used most frequently:

- Format (Type)
- Subject
- Location.

A few students used a map to further narrow down their search by location (see Figure 2). The DPLA map is a visualization tool that demonstrates the location and the number of digital objects coming from particular participating institutions. However, when used as an approach to refining a search, the map feature elicited mixed results. Some participants used it quite effectively, while others struggled with understanding that the results retrieved when using a single point on the map were not necessarily about this particular location.

The participants also commented about the lack of advanced search. The question, “Is there an advanced search?” came up several times during observations and was posed by graduate students and faculty who were familiar with an advanced search function from their interactions with library databases. Three graduate students attempted to construct complex Boolean queries in the DPLA simple search box and had to revise their searches because they retrieved no results.
Most participants had a very positive experience searching the DPLA and found the site relevant to their academic needs. Several participants commented on the usability attributes, such as ease of access, ease of search, clear layout, and aesthetic appeal. Participant A said that he was surprised to find the site relatively easy to search: “Yeah, it was surprisingly easy to search. I expected such a large repository to be difficult to use.” Participant U stated: “I like it, I think it’s very user friendly and very streamlined and what I like about it is it seems accessible to a wide range of viewers at all different expertise levels [...] It is aesthetically clean, and it’s appealing and it looks professional but it also looks inviting and it’s not too complicated.”

Navigating the DPLA Distributed Structure
Analysis of the observation data demonstrates that the study participants were able to complete most of the predefined scenarios and successfully locate and download relevant objects. In this study, 85% (36 out of 42) of the predefined scenarios were completed successfully. However, the study participants experienced some confusion in searching and navigating the DPLA’s distributed system. Seven participants encountered difficulties in locating images or sound recordings. They were eventually able to find the objects they were looking for, but sometimes had to repeat their steps or try a different approach. Two participants found broken links on the sites of participating institutions. In the set of 42 predefined scenarios, four scenarios were not completed. In those four instances, the participants could not locate relevant resources or gave up during the navigation process. Table 2 summarizes the problems that were observed or reported. The findings of this study indicate that the distributed structure of the DPLA caused some confusion in users’ navigation and played a role in completing tasks and finding objects.

Table 2. Issues in navigating the DPLA distributed system
<table>
<thead>
<tr>
<th>Observed or Reported Issues</th>
<th>No. of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced difficulties in locating images or sound recordings in the predefined scenarios</td>
<td>7</td>
</tr>
<tr>
<td>Unable to complete a predefined scenario</td>
<td>4</td>
</tr>
<tr>
<td>Encountered broken links on the site of the participating institution</td>
<td>2</td>
</tr>
<tr>
<td>Expressed some confusion about navigating the DPLA structure</td>
<td>9</td>
</tr>
</tbody>
</table>

The DPLA’s distributed model puts users on two- or three-step navigation pathways depending whether objects are found through a content or service hub, and whether a service hub hosts content and metadata or is just a metadata aggregator. The two-step search process involves 1.) searching the DPLA site and 2.) linking directly from a metadata record to an object in a content hub (for example, HathiTrust, the National Archive, or Smithsonian) or service hub hosting content such as Digital Commonwealth or The Portal to Texas History. The three-step process involves an intermediate step of connecting through a regional or state service hub that is only a metadata aggregator, like the Mountain West Digital Library. In the three-step path users 1.) search the DPLA site, 2.) connect from a DPLA metadata record to another metadata record in a service hub, and finally, 3.) link to an object at an originating institution.

The two-step search process that involves searching the DPLA platform and following a link to the object did not cause any difficulties. All participants moved through that process smoothly and efficiently. However, the three-step process wherein a user goes from the DPLA site to another aggregator was sometimes confusing and time-consuming. Some participants became disoriented at the middle layer when, after following a link from the DPLA, they encountered another metadata record at the service hub level, but could not see an actual object. Figure 3 illustrates the three-step process where on the service hub site (in this case on the Mountain West Digital Library) users had to click on “View Resource” in order to link to an object at the contributing institution’s website. The lack of thumbnail images on the Mountain West Digital Library was particularly disorienting, as described by Participant P, “At one point I totally lost the image that I was, you know, looking for and that was kind of, that was kind of awkward and a little confusing.” Several participants struggled at the middle step. Some started searching the aggregator site, while a few gave up at this point. Participant F, a graduate student, who was not able to locate a sound recording for the predefined scenario, left the site at that point and went to YouTube instead. Difficulties in locating objects as well as inability to complete scenarios all occurred during the three-step discovery process through a metadata aggregation service hub.
The multi-step process also took more time and caused disorientation because of different graphical interfaces the participants encountered on their pathways. Participant L stated: “It would take a little bit more time and it kind of maybe threw me off a little bit in terms of like what page am I on? Where am I? And then if, I don’t know, sometimes you could get caught like searching on that page instead of going back to the initial one.” As illustrated in Table 2, nine participants expressed some confusion about the DPLA’s structure in the follow-up interviews. Participant R explained, “It is a little confusing to me that it goes to all of these other institutions and then you have to sort of navigate all the buttons and the interface for the institution you’ve been taken to.” Five participants in the sample experienced difficulties in tracing back their steps.

On the other hand, a number of participants navigated the three-layered structure without any difficulties. Even those who had initially experienced some confusion learned how to move through the layers relatively quickly. They adapted to the two- or three-step search process by opening multiple tabs in a browser and navigating between the tabs. Participant E explained this behavior, “I’m a child of the Internet, I’m used to it. So I, that’s why I always open things on tabs”. One participant mentioned that despite some difficulties in locating objects, the structure of the site was not a problem to her. She said that navigating from one site to another is something to be expected, “you have to jump ship multiple times but that’s OK” (Participant H).

In addition, several participants viewed the distributed structure of the DPLA as an advantage rather than a hindrance and stated that they would prefer it to a centralized model. They were impressed with the amount of aggregated resources and ability to search from one portal, as noticed by Participant D, “It’s nice to have all of the resources in one place.” Others found linking to originating libraries and archives beneficial “because it gives credit to the institutions that provide the resources” (Participant N). Participant Q also felt that the DPLA
increased the visibility of participating institutions and exposed her to a wide range of collections that she was not aware and would not have explored otherwise.

**DISCUSSION**

This study focuses on user navigation through a large scale distributed DL system. It confirms several findings from the previous studies of large-scale DLs, but also contributes new insights on user navigation though a multi-layered system. Researchers exploring use of large-scale digital libraries point out that these new discovery systems serve a very broad, heterogeneous population of users who use different access devices and vary in their perceptions, search strategies, and navigation patterns (Dobreva and Chowdhury 2010; Nicholas et al. 2013; Zavalina and Vassilieva 2014). Studies of Nicholas et al. (2013) and Zavalina and Vassilieva (2014) analyzed logs of large populations, but even in the small sample of users selected for this study, differences in search behavior were evident. Faculty and graduate students demonstrated the characteristics of expert searchers and were interested in using advanced search, while undergraduate students tended to stick to a basic keyword search. Different levels of basic and advanced searches were also found by Zavalina and Vassilieva (2014) in their study of the NSDL. While the NSDL provides an advanced search page, the current versions of the DPLA does not. The DPLA offers a simple search box and means of refining queries through facets but less control in constructing advanced searches. The lack of advanced search in a large-scale environment with millions of resources is rather limiting, especially for academic users who are used to building advanced queries.

This study also echoes some findings of the usability study of Europeana where users experienced difficulties navigating the system and were unable to locate objects due to broken links (Dobreva and Chowdhury 2010). Likewise, the DPLA users in this study came across some broken links when trying to retrieve digital objects from contributing institutions. These findings point to a major challenge of distributed large-scale DLs, especially in comparison to libraries built on a centralized model. Ultimately, access to content in distributed systems depends on the commitment and stable infrastructure of the participating institutions. Broken links, even in small numbers, erode user’s trust in the DL system.

As discussed in the literature review section, use studies of large-scale DLs are still limited and tend to analyze user search queries. This study looks into user information behavior beyond query formulation and examines navigation pathways through the distributed network of DPLA participating institutions. The findings indicate that some users were confused by the multi-layer structure and especially by the lack of thumbnails at the middle layer. A distributed DL system, like the DPLA, would serve users better if user interfaces of service hubs were consistent in providing visual cues in addition to metadata records.

A unique and interesting finding of this study relates to user adaptive behavior. Although most participants were new to the DPLA, they learned quickly how to navigate through the two- or three-step process. Several participants stated in the interviews that they expect moving from one site to another as part of online searching. They mentioned a similar experience with a library discovery tool, Summon, where upon searching the Summon portal they had to link to resources in different library databases. It’s important to point out here that all study participants were academic library users who had prior experience with library federated systems, which may not be the case of other DPLA user groups. This study was conducted with a limited number of participants recruited from university students and faculty. It would be worthwhile to expand it to other user populations, such as K-8 students and teachers and general public.
CONCLUSION

Large-scale digital libraries address the challenge of resource discovery in the DL environment, but pose new challenges for user navigation through the aggregated platforms. The centralized models provide a more reliable and predictable navigation environment but offer limited visibility to contributing institutions. Large-scale distributed DLs, like the DPLA examined in this study, rely on the systems of local content and service providers and are more challenging for user navigation because of their multi-layered structure. This study contributes a user perspective on the navigation through a distributed system where some users find it confusing but others point out the advantages of a distributed model, including serendipitous discovery and an increased visibility of contributing digital libraries. Although this study focuses exclusively on the DPLA, its findings on user navigation though a multi-layered distributed system may have implications for the design of other aggregated digital library systems.

REFERENCES


