

Ryan Perry
IST 616
Final Project: Information Retrieval System Analysis
December 5, 2014

Archiving the Internet

The Library of Alexandria represents an enduring, if overly mythologized, reference point for collection development and access. All of the world's knowledge bounded by four walls; accessible to knowledge-seekers and enabling them to climb onto the shoulders of giants. The Internet, in its relatively brief existence, has made similar promises in collecting the world's knowledge—before the realization set in that web pages have the potential to be even more ephemeral than printed works. The Internet Archive has as its lofty goal to preserve this ephemera on a truly massive scale, as trumpeted by their motto of “universal access to all knowledge.”¹ However, preservation on its own proves to be insufficient when confronting the massive scale of the Internet Archive's collection; access, more than availability, needs to be taken into consideration. The Internet Archive subsumes all four FRBR User Tasks within a web browser: find, identify, select, obtain. This paper will describe and assess the Internet Archive's functionality across these four tasks.

The Internet Archive comprises a collection of collections. Rather than being merely a library of surrogate records, users can obtain a digital copy of the item they are seeking directly from the archive. The collection has two major facets: material that has entered the public domain (often due to expired copyright), and born-digital materials (especially web content). Owing to the breadth of materials available, users seek a wide range of items from the collection.

¹ As written in the top right of the Internet Archive's website. Internet Archive, accessed December 2, 2014, <http://www.archive.org>.

The Live Music Archive collection , for example, has a passionate user base that uploads, reviews, and comments on concert recordings by bands such as The Grateful Dead, whose collection includes 9,906 items.² The most prominent collection of born-digital materials are searchable via the Wayback Machine, which includes captures of 435 billion web pages and allows users to view websites as they existed at specific points in time.³

Founded in 1996, the Internet Archive is a non-profit organization whose mission includes “offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format.”⁴ Collection development happens through a combination of partnerships with organizations like the Smithsonian and the Library of Congress, as well as user uploads and web crawlers.

Find

Users can find materials on Internet Archive through browsing and searching. In terms of browsing, materials are categorized at the top level by media type: web, video, text, audio, and software. Within these broad categories, curated sub-collections feature materials that were added en bloc. Project Gutenberg represents one prominent example that includes plain text ebooks of works in the public domain. Once on the Project Gutenberg page, users can browse by title, author, recently reviewed items, “this just in,” most downloaded items, and the curators include a “spotlight item” drawn from the collection. Users can also participate in forums, which could assist with the task of finding an item, but would be more likely to be used by those who

² “Welcome to Grateful Dead,” Internet Archive, accessed December 2, 2014, <https://archive.org/details/GratefulDead>.

³ “Wayback Machine,” Internet Archive, accessed December 2, 2014, <https://archive.org/web/>.

⁴ “About the Internet Archive,” Internet Archive, accessed December 2, 2014, <https://archive.org/about/>.

are discussing the finer points of an item they have already found. This general format is used for most collections housed on the archive.org website. With a large number of sub-collections contained in the Internet Archive, users' attempts to find specific items are undermined by artificial divisions. The use of subject-driven sub-categories would alleviate this problem, but would also have the effect of removing items from context of the specific collection.

While finding an item within Project Gutenberg does not pose a significant challenge, browsing to one of the many smaller sub-categories can be very cumbersome. For example, a user attempting to find a collection of Alan Lomax recordings would have to browse through a list of collections alphabetized by the first letter. The "A" list alone contains over 1,235 collections—which do not include any Alan Lomax-related recordings because the collection titles do not begin with his name. As a result, finding through browsing alone can be near impossible depending on the subject.

The search function allows for another, likely more widely used, method for users to find an item. While the search bar can be useful, the advanced search page reveals an impressively extensive set of fields and limiters. These fields utilize the available metadata, which will be discussed in a later section. For users who need assistance with the available boolean modifiers, the bottom of the pages has instructions and examples. Even with instructions, the advanced search functionality is likely to be too complicated for many casual users. The lack of controlled vocabulary limits the usefulness of advanced search parameters, as the results of a search could be spread across multiple different spellings of a given title, creator, or subject. As an aggregation of a materials in various formats, organized in a variety of collections, the Internet Archive inevitably feels disjointed in browsing and searching for materials.

The Internet Archive's website redesign, though currently in beta testing, offers an attempt to make it easier for users to find materials on the website. As stated in an Internet Archive blog post on the redesign, the new website will feature "more visual cues to help you find things, facets on collections to quickly get you where you want to go, easy searching within collections, user pages, and many more."⁵ This redesign follows modern trends, especially in library websites, of making the search bar the focal point of the page. Furthermore, browsing in the redesign allows for dynamic filtering and sorting—augmented with a more visual interface. While this redesign remains in need of some improvements, it would certainly empower users to better navigate the "finding" stage of the FRBR user tasks.

Identify


The shortcomings of the Internet Archive's user interface, exacerbated by the size of the collection, become more severe when the user moves into the "identify" stage. In identifying an information resource, Barbara Tillett, the Chief of the Library of Congress's Cataloging Policy and Support Office, the user must "confirm they have found what they looked for, distinguishing among similar resources."⁶ At this stage, the sheer size of the Internet Archive becomes a hindrance due to a lack of unified cataloging standards on one hand, and a one-size-fits-all display of search results.

By way of example, performing a basic search for "Alan Lomax" yields 723 results that can be difficult for a user to parse and identify. Among the top fifty results displayed on the first

⁵ Alex Rossi, "Redesigning Archive.org," *Internet Archive Blogs*, November 5, 2014, accessed December 4, 2014, <https://blog.archive.org/2014/11/05/redesign/>.

⁶ Barbara Tillett, "What is FRBR?: A Conceptual Model for the Bibliographic Universe," Library of Congress Cataloging Distribution Service, accessed December 1, 2014, <http://www.loc.gov/cds/downloads/FRBR.PDF>, 5.

page, most appear to be titles of collections associated with Lomax, with some individual items included as well; a creator record for Lomax does not appear in these initial results.



[Lomax Collection Recording of Blackfoot](#) - [Lomax, Alan](#) (Collector)
This is a recording of Blackfoot. This is a recording from [Alan Lomax's](#) Parlometrics collection. These recordings were made by linguists around the world as well as by [Lomax](#) himself. They have been digitized from the original reel-to-reel tapes. The original notes which accompanied the tapes were often either incomplete, indefinite, illegible or missing. Because of this, the language in this recording may not have been identified or may have been misidentified...
Keywords: [Blackfoot](#); [Alan Lomax Global Language Audio Collection](#); [bla](#)
Downloads: 119

The difficulty in distinguishing between these results stems from the lack of information labels, as the results do not differentiate between collections and items. The title of the resource (or collection) is hyperlinked, with the creator field on the same line, followed on subsequent lines by a brief description, keywords, and number downloaded; a headphones icon indicates the result is an audio recording. While these fields could allow a user to successfully identify certain records, the ambiguity of content hinders identification.

One partial solution would be to divide the results between different categories (such as “collections,” “items,” and “creators”), which would assist the user in quickly parsing the results. These categories could further differentiate by displaying different fields. For example, a creator result could list the lifespan of the name authority, while a collection might indicate the donor. This would save the user from having to click on the result to get the information necessary to identify the resource.

Unfortunately, the proposed website redesign, while more visually appealing, only serves to further obscure the information relevant to the identification process. Results appear as cards, and in the case of audio materials, with the waveform image prominently displayed. While the waveform could be useful in certain contexts, it generally does little to aid in identification and consequently serves mostly as an aesthetic representation. The use of a “by” relationship implies that Lomax is the performer for the



Lomax Collection
Recording of Blackfoot
by [Lomax, Alan](#) (Collector)

 |  117 |  0 |  0

audio, when in fact he recorded the audio. Without an included description field, the appropriate result would be difficult to identify without clicking through into the expanded records for all 723 items. In contrast to the “find” functionality, the redesign actually serves to make it more difficult to identify a resource.

Since browsing by collection simply returns a list of search results to the entire sub-category in alphabetical order, the task of identifying a relevant results become practically impossible due to the sheer length of the list. The results list is in the same format as from searching, so the same shortcomings apply: users are not given the succinct and relevant information they need to identify an item.

Select

Many of the same elements affecting the identification of resources in the Internet Archive also apply to their selection. According to Tillett, selecting “involves meeting a user’s requirements with respect to content, physical format, etc. or to reject an entity that doesn’t meet the user’s needs.”⁷ The prevalence of duplicate copies of the same information resource impedes the user’s ability to select the most appropriate result after searching for and identifying the item. Without enough relevant information on the results page, distinguishing features can be ascertained using the metadata on the page for an individual item.

I used the Internet Archive as a reference for Professor Ken Lavender’s Management & Organization of Special Collections course, while performing research for a descriptive bibliography study. Seeking additional digitized copies of our assigned book, *Dictionarium Polygraphicum*, I had to select copies of the correct edition and volume. Given few

⁷ Tillett, 5.

differentiating details on the search results page, my first instinct was to select the copy that had logged the greatest number of downloads—under the logic that if others had found the copy useful, I would as well. Of the available copies, this item also had the most detailed metadata.

Regarding the metadata driving the Internet Archive's backend, the collection of text resources appears to be the only media collection on the website to use library catalog metadata standards. Surrogate records for text resources are encoded in MARCXML, while other resources appear to be encoded in a proprietary form of XML. The completeness of the descriptive metadata varies from item to item, but most of the text resources include information on the book contributor as well as a technical metadata relating to the digitization process. Other media formats lack this level of detail, and tend to vary widely in terms of quality, due to the prevalence of user-uploaded content in many collections. Each item in the Internet Archive is given a unique alphanumeric identifier, which acts as an internal URI and serves as the basis for the item's public URL. With such varying levels of quality and quantity, metadata often cannot be relied on in the process of selecting a resource.

While the available metadata can be useful in selecting text resources, the quantity and uneven quality of user uploaded media in other categories make it difficult to select an appropriate resource. The Grateful Dead collection, being one of the more active areas of the Internet Archive, offers an example of the impact of user involvement. More so than in less active collections, a critical mass of ratings and reviews can help to differentiate the 9,906 recordings in this collection. In addition to numerous detailed reviews, the collection also has a forum for additional discussions. User-generated content can help to add value to records in

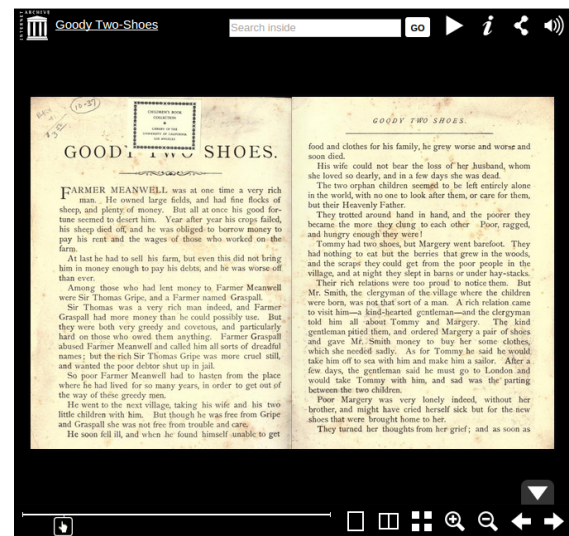
these active areas, but the majority of the Internet Archive lacks a dedicated enough user base to factor into the selection process.

Obtain

Once an information resource has been found, identified, and selected, the user moves to the final stage of the FRBR user tasks: obtaining the item. While the Internet Archive's information retrieval system remains stymied by limitations resulting from the user interface and lack of consistent metadata—even within the proposed redesign—the true strength of the collection is in the ease through which users can obtain information resources. If users navigate the website effectively, they can download, view, stream, or play items from the detail pages of the Internet Archive website.

Rather than having to travel to a library to obtain a copy of text resources, the collection can be viewed through an embedded ebook viewer on the website. Users can page through the resource and immediately reflect on whether or not it meets their needs. With this instant feedback, users can perform the cycle of FRBR user tasks in rapid

succession and can re-select much quicker than in collections that do not have embedded resources. Additionally, the resource can be downloaded in a number of formats, including: PDF (with an option for a black and white copy), EPUB, Kindle, Daisy, Full Text and DjVu; through several download protocols, including HTTPS and Bittorrent. The Internet Archive does an admirable job in making content easily accessible, and in formats that can be used with little



effort on the part of the user. If the user has any difficulties in using the files, the record links to a page that has instructions and explanations for how the files can be used in various formats.

The audio and video collections have similar tools in place that allow users to obtain resources directly within a browser window or via downloading the files. The software collection, however, has posed a greater challenge since many programs in the collection require obsolete hardware or emulation software in order to function. In addressing this limitation for a portion of the software collection, the Internet Archive launched the Internet Arcade in early November 2014. The Internet Arcade allows users to play 723 games published from the 1970s to 1990s in a browser using JSMAME emulation.⁸ With a large number of websites spreading the news of the new service, the Internet Archive experienced a sudden burst of over 100,000 visitors.⁹ This overwhelming interest in having an ability to obtain resources immediately speaks to how users expect to use information retrieval systems.

Conclusion

The Internet Archive has been at the forefront of online media preservation for its entire existence. However seamless the Internet Archive has made the FRBR user tasks, it must be taken in the context of how resources used to be found, identified, selected, and obtained. In contextualizing the website redesign (which will be the first major overhaul since 2002), Alexis Rossi reminds us:

Having thousands of movies available on the Internet in 2002 was actually pretty rare (remember, Youtube didn't exist until 2005). Those 5,000 media items couldn't be played on our site – you had to download them to your own computer to watch or listen.

⁸ “Welcome to Internet Arcade,” Internet Archive, accessed December 5, 2014, <https://archive.org/details/internetarcade>.

⁹ Jason Scott, “Inviting the Internet Over to Play,” *Internet Archives Blogs*, November 5, 2014, accessed December 4, 2014, <https://blog.archive.org/2014/11/05/inviting-the-internet-over-to-play/>.

It was very difficult to add your own files to the Internet Archive – and who would have had the bandwidth to do it anyway? In 2002 only 21% of U.S. homes had “high speed” internet connections. High speed back then meant 200 kb per second.¹⁰

Now that we live in the age of YouTube (and Spotify and Kindles), users expect to be able to obtain and view media instantly from any browser, anywhere. Somehow, the ability of the Internet Archive to offer instant access to resources feels less remarkable than it should.

The Internet Archive would greatly benefit from reflecting on FRBR user tasks in developing a more fundamental redesign to make the process of finding, identifying, and selecting more fluid. The fact that users are able to obtain a resource within a browser is only useful if they are able to get there first. While the Internet Archive does indeed need the new coat of paint the redesign promises, it would benefit much more from a new engine.

¹⁰ Rossi.

References

“About the Internet Archive.” Internet Archive. accessed December 2, 2014.

<https://archive.org/about/>.

Internet Archive. Accessed December 2, 2014. <http://www.archive.org>.

Rossi, Alex. “Redesigning Archive.org.” *Internet Archive Blogs*. November 5, 2014. Accessed December 4, 2014. <https://blog.archive.org/2014/11/05/redesign/>.

Scott, Jason. “Inviting the Internet Over to Play.” *Internet Archives Blogs*. November 5, 2014. Accessed December 4, 2014. <https://blog.archive.org/2014/11/05/inviting-the-internet-over-to-play/>.

Tillett, Barbara. “What is FRBR?: A Conceptual Model for the Bibliographic Universe.” Library of Congress Cataloging Distribution Service. Accessed December 1, 2014. <http://www.loc.gov/cds/downloads/FRBR.PDF>.

“Wayback Machine.” Internet Archive. Accessed December 2, 2014. <https://archive.org/web/>.

“Welcome to Grateful Dead.” Internet Archive. Accessed December 2, 2014.

<https://archive.org/details/GratefulDead>.

“Welcome to Internet Arcade.” Internet Archive. Accessed December 5, 2014.

<https://archive.org/details/internetarcade>.