JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 5, General Issue (2017)

Making Visualization Work for Institutional Repositories: Information Visualization as a Means to Browse Electronic Theses and Dissertations

Leila Belle Sterman, Susan Borda

Sterman, L.B. & Borda, S. (2017). Making Visualization Work for Institutional Repositories: Information Visualization as a Means to Browse Electronic Theses and Dissertations. *Journal of Librarianship and Scholarly Communication, 5*(General Issue), eP2140. https://doi.org/10.7710/2162-3309.2140

ISSN 2162-3309 10.7710/2162-3309.2140

PRACTICE

Making Visualization Work for Institutional Repositories: Information Visualization as a Means to Browse Electronic Theses and Dissertations

Leila Belle Sterman
Scholarly Communication Librarian, Montana State University
Susan Borda
Digital Technologies Development Librarian, Montana State University

INTRODUCTION An attractive repository with clear, well-structured, and accessible content can be a powerful recruitment and publicity tool for administrators, graduate admissions officers, and others trying to bolster support for repositories. Digitizing ETDs is a lengthy process that provides digital access to valuable materials. We demonstrate the benefits of visualizing repository content to increase visibility of these newly digitalized resources. **DESCRIPTION OF PROJECT** The goal of the project was to create an interactive visualization to make our newly digitized theses and dissertations more discoverable. Although this content is described in rich metadata and searchable, we wanted to ensure that it was easy to browse the collection to encourage discovery and use of these historical papers. By employing an interactive visualization of these resources, we provide users an alternative to navigating menus and browsing text. The visualization allows users to see ETDs by college, department, and graduation date, at a glance. The process began with data cleanup involving extracting and normalizing repository metadata, then the data was processed and the Data-Driven Documents (D3) JavaScript library was used to generate the visualization. **NEXT STEPS** Visualizations have vast potential for creating engaging user interfaces for digital library content. We will explore how people are using the visualization as we move forward with this process to visualize multiple collections.

Received: 05/06/2016 Accepted: 02/24/2017

Correspondence: Leila Belle Sterman, Montana State University Libraries, Bozeman, MT, 59717, leila.sterman@montana.edu

INTRODUCTION

As unique library collections items are digitized, libraries should take advantage of the opportunities of having well described, linked, digital items, and provide some means of easily accessing these collections. While there are multiple options for facilitating access to digital materials, this paper outlines the reasoning and process used to develop a three-tiered visual representation of a collection of electronic thesis and dissertations (ETDs) for easy, interactive discovery. Visualizations allow users to understand vast amounts of information quickly based on visual compaction that often presents humans with a more clear understanding of trends, gaps, and clusters. Our visualization relies on uniform cataloging practices and the unique but not unusual organization of our university, specifically, the metadata terms for "College," "Department," and "Year." By clearly arranging the visualization by known concepts, we provide easy routes for those familiar with the university or similarly structured institutions to seek relevant content based on common organizational structures. Digital systems allow multiple pathways to the same digital content, a mechanism that allows items to be retrieved through multiple search and browse tools if the items are adequately cataloged. The increased visibility of ETDs is important locally as they make up the majority of the authors' institutional repository, to the public as they hold great amounts of new knowledge, and to their research fields as they are given less formal publicity than other research products such as journal articles or monographs. This visualization provides an alternate, engaging pathway to our theses and dissertations, papers that provide a timely and broad view into the research produced on campus. We aim to engage scholars, administrators, and the public by facilitating discovery of these digital items.

LITERATURE REVIEW

Although universities began adopting electronic journal formats as soon as the technology was available in the 1990s, the transition from print to electronic theses and dissertations was much slower, especially in the United States (Copeland, 2011; Edminster and Moxley, 2002). The cost (Weisser, 1997; Brown, 2010) and the amount of work entailed (Copeland, 2008) slowed the transition. Even given the time and cost of digitization, many agree that theses and dissertations (TDs) are important stores of information that represent years of research (Copeland, 2011; Lewis, 2002). Specifically, their importance in the literature lies in the fact that "dissertations are longer than journal articles and cover their topics more comprehensively. They are more responsive to past literature than journal articles and are usually researched, refined, and revised over a longer period of time" (Suber, 2008, p. 26). TDs delve deeply into the existing literature—responding to it, filling in gaps, and engaging in the conversation of scholarship that is sometimes less represented in journal articles.

Moreover, "TDs represent a substantial contribution to advances in these disciplines. This valuable knowledge is clearly worth disseminating as widely as possible" (McCutcheon, 2011, p. 64). Clear routes to discovery and access are important. "Because of [ETDs'] high quality," Peter Suber asserts, "the access problem is worth solving" (Suber, 2008, p. 26). Not only the research community will benefit from accessible, digitized TDs; universities can use them to increase their visibility and reputation. In terms of institutional repositories, "following published journal articles in pre- or post-print formats, ETDs are the most important document type in open archives" (Schöpfel, 2013, n.p.). Institutions gain prestige from all of their published literature—not just papers and books, but theses and dissertations as well (McCutcheon, 2011).

A majority of graduate papers are now submitted as *electronic* theses and dissertations (ETDs) rather than TDs (Lyon, 2015). Electronic papers benefit from their digital nature, multimedia-embedding capabilities, greater accessibility, and lower cost (due to printing, shelf space, etc.) (Boock and Kunda, 2009), and these digital items can be read by more than one person at a time from anywhere in the world. Some universities now digitize graduate papers from the past in addition to their current practice of accepting "born digital" copies of theses and dissertations (Boock and Kunda, 2009; Park and Richard, 2011). Access to ETDs varies widely, however. Frequently, electronic copies of these graduate papers, especially those that have been digitized from paper copies, are posted on university web sites or in library catalogs. Before digitization, these papers are "some of the least well-disseminated and accessible scholarship generated on academic campuses" (Shreeves and Teper, 2012, p. 532). After digitization, they are more accessible but not necessarily more visible. While digitization is a huge benefit to scholars who cannot travel widely to read papers or who do not have access to InterLibrary Loan, more can be done to surface and highlight this research.

For many catalogers, cataloging ETDs was the first large-scale electronic process (McMillan, 1996), that required them to learn new practices and significantly change their workflow (Boock and Kunda, 2009). The challenge of cleaning and standardizing metadata is amplified by the large numbers of digitized items that arrive with over 100 years' worth of cataloging styles, nuances, and mistakes. McCutcheon states that "[a]nother obstacle to access has to do with the representation of scientific symbols, diacritics, and some punctuation in author-supplied metadata" (McCutcheon, 2011, p. 66), yet another challenge to standardizing metadata. As we experienced, looking at metadata in the aggregate visually reveals the errors, the inconsistencies, and the outliers that may be hidden otherwise. To provide records that may be reused for multiple projects in the future, the task of applying metadata includes ensuring consistent standards are applied to each metadata record (Schöpfel, 2013, Liu, 2004).

Another aspect of information retrieval that has changed since these historic papers were cataloged is that the individual doing the retrieval is often a layperson, whereas prior to the personal computer and digital text searches, specialists largely performed that task (Smeaton, 1996, p. 3). Descriptions and ontologies now need to communicate to a broad audience—one without the specific library or database training that was expected for most of the twentieth century.

Discovery and Visualization

Historically, "access to theses and dissertations is patchy at best, limited by physical lending restrictions" (Andrew, 2004, p. 2). Digitized, theses and dissertations have become accessible to people who cannot travel to an institution or who cannot rely on interlibrary loans. When ETDs are openly accessible online, they also work to bridge the information inequity gap (Edminster and Moxley, 2002). Despite digitization, ETDs will not lose their reputation for being difficult-to-discover resources (Andrew, 2004, p. 2) unless libraries take advantage of the benefits of their digital form. Even as digital objects, ETDs offer are still gray literature and require cataloging to be easily found digital resources.

Institutions, researchers, and graduate student authors benefit from the increased visibility of digital, easily searched materials (Sarkar and Mukhopadhyay, 2010, p. 356). Full text search is one of the benefits that may be extended to electronic items, although it should be noted that "discovery by keyword searching on the internet is successful only when the searcher and the [person] creating the metadata use the exact same terminology" (McCutcheon et al, 2008, p. 42). It remains unlikely that non-expert searchers will use those specific keywords, or that expert searchers will get more than a few items on a highly specialized search in repository software, even as commercial internet search engines attempt to use predictive search to help users find what they are looking for based on a broader, semantic understanding of search terms. Relying on full text search within repositories has limitations as well. Although full text search aids discovery through search in a repository, it does not enable browse functionality for collections of papers. Research in human computer interfaces and information retrieval suggests that "a reasonable way to present complex information is to produce multiple views of the same information" (Lucarella and Zanzi, 1996, p. 122). Librarians are familiar with faceted text searching options in digital catalog software and databases, which allow users to narrow and broaden a search as needed. When conducting visual search or browse, especially in an interactive system, users benefit because "[t]he cognitive overhead required from a user in facing tangled information structures can be alleviated if the system presents only the most relevant pieces of the stored information while hiding the rest." (Lucarella and Zanzi, 1996, p. 122).

Card, Mackinlay, and Shneiderman define "Visualization" as "the use of computer-supported, interactive visual representations of data to amplify cognition" (1999). Information Visualization has many benefits to users and information practitioners. These techniques and tools have the ability to both give users an overview and insight into data or a library collection and can be used to see patterns in data or metadata that are more difficult to see otherwise (Shneiderman, 2002). Visualization is beneficial based on the visual compaction of vast amounts of information that may present humans with a more clear understanding of trends, gaps, clusters, and potentially suggest topics of future study (Shneiderman, 2002). Some collections will benefit from information visualization more than others. These include: collections with meaningful metadata structures, collections that have unknown content or the organizations of that content is unknown (as in the case of laypersons browsing), for people who have a hard time verbalizing or understanding their own information need, or when a desired item is easily seen but not easily described (Fekete, et al. 2008, p. 3). Specifically, collections with meaningful metadata structures will be easily transformed into visual depictions of information: content in collections that was previously hidden may be surfaced without a user knowing the specific topic, vocabulary or organization of a collection.

Visualizations make a concrete connection between the diverse objects within a collection and the groupings that an organization believes will help communicate the narrative of that collection. If a user interface allows the user to retrieve information easily or more clearly understand that item's relationship to other items (Rose, 1991 p. 13), then the user interface enhances the collection for that user. Visualization allows users to see the contents of a collection based on groupings and provides "an almost physical depiction of data" (Jisc, 2013). It can effectively showcase entire collections in easily digested ways and help guide the development of user-focused features. Information visualization is powerful because it exploits an innate cognitive aptitude: the human brain processes large amounts of visual information much faster than data on spreadsheets. For this reason, visualizations are an ideal way for users to access large amounts of information that might otherwise remain hidden.

An attractive repository with clear, well-structured, and accessible ETD content can be a powerful recruitment and publicity tool for administrators, graduate admissions officers, or those trying to bolster support for repositories. Further, in a 2009 study, researchers found that one repository hoped to have tools to "support publication analysis, visualization tools, and social networking" (Palmer et al, 2009, p. 152) to help with easy and attractive content recruitment. These added tools would help integrate a repository into the social- and metrics-based aspects of scholarly communication, helping to assimilate repository content into the landscape of other academic and web-based workflows or habits. Looking at the services that users expect and shaping the user experience to meet those expectations helps increase

the accessibility and ease of use of repositories. Collins et al. explore the digital experience of users finding items after a museum visit (2005). When describing their collection, they are careful to present the collection according to concepts that are familiar to their visitor audience (Collins et al., 2005). Lessons from these studies shape our understanding of user's understanding of "clear well-structured content" and help drive future projects to increase use and support of repositories.

DESCRIPTION OF PROJECT

Overview

Multiple libraries document the often lengthy ETD digitization process (Shreeves & Teper, 2012). Although digitization and ongoing digital capture of these graduate papers is critical to their visibility in repositories and impact on current scholarship, further action such as proper metadata application is often needed to provide access to the information. Without proper metadata, users will not be able to easily find the items they are looking for (Lopatin, 2006). This became evident at our home institution: although ETDs had been digitized and placed in the library's digital repository, they were not optimized for discovery. To facilitate browsing of the ETD collection, we created a sunburst visualization to provide a visual navigation of this specific collection of well-cataloged digital objects. We propose that this interactive visualization will help elevate the design, metadata, availability, and accessibility of ETDs in our DSpace-based institutional repository (see figure 1). In this paper, we detail the processes and considerations involved in creating this visualization of our digital collection and include code to enable others to replicate this process.

The Montana State University library digitized more than 5,000 theses and dissertations in the fall of 2015 dating 1901 to 2003 (which are now *electronic* theses and dissertations or ETDs) adding to our collection of digitally held papers in our DSpace repository. The graduate school has accepted digital theses and dissertations since 2003, and following the success of that program, digital deposit of ETDs into DSpace became mandatory in the 2004-2005 school year. We now have a rich scholarly resource in our repository with more than 100 years of content. Although this content is described in rich metadata and searchable, we wanted to ensure that it was easy to browse the collection to encourage discovery and use of these historical papers. By employing an interactive visualization of these resources, we provide users an alternative to navigating menus and browsing text. The visualization allows users to see ETDs by college, department, and graduation date, at a glance.

This project builds on work presented in a poster at the Open Repositories conference in June 2015 (Borda and Sterman, 2015). Although we continue to use the Data Driven

Documents (D3)[1] javascript library to visualize our ETDs, as previously described, we have now also automated the extraction of the original metadata that describes our ETDs, reformatting the metadata for D3 to reorganize and create the final visual presentation. D3 is an open source javascript library created specifically for visualizing data. It was chosen because it has built-in data nesting capabilities that can turn flat data (i.e. comma separated values (CSV)) into a hierarchical format. It is also flexible and easy to embed into our repository's landing page. Flexibility was additionally important for the ability to accommodate semi regular updates to the underlying data from our DSpace repository. We are running a locally hosted instance of DSpace 5.x XMLUI, which can provide data in CSV format when using the "Export Metadata" function via the user interface or in XML format when accessing it via OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting). We use both CSV and XML data to produce our visualization.

To encourage the replication of this process, we used a Python script to extract metadata from our DSpace repository and the D3 javascript library to create the visualization. We chose the sunburst design to facilitate the exploration of this collection based on our existing metadata, for reasons elaborated in the next section.

[1] http://d3js.org/

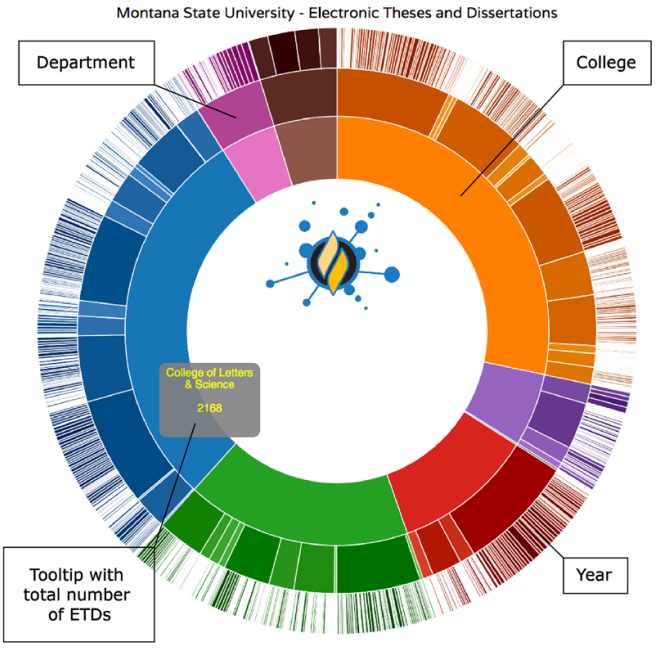

Figure 1. Sunburst visualization http://www.lib.montana.edu/~susan/sunburst/etd_sunburst_labelv2.html

Considerations and Process

While knowledge discovery in general, and of grey literature in particular, can be a time-consuming challenge (Richardson, Srinivasan & Fox, 2008), our project aimed to enable the discovery of ETDs and surface useful information for users. Users of this tool and the repository are a potentially diverse group. While many are campus-based and include students, researchers, administrators, and admissions officers, our hope is that there will be many users who discover these papers through search and browse in our repository regardless of their location or university affiliation so we wanted a format that made sense to both internal and external users. To facilitate browsing our ETD collection, we used a process that could easily be extended to promote awareness and facilitate use of other digital objects so that this might be useful to other projects. DSpace is visually bland, yet "these types of visualization techniques improve the user's ability to understand the content in the repository, while also aiding in serendipitous discovery – traditionally one of the weakest features of a digital repository" (Phillips et al. 2007). Our visualization allows users to interact with it and narrow or broaden their search in a visual manner that is navigated by clicking on each wedge of the "donut." Each ring of the "donut" represents a metadata element. It moves hierarchically from *College* to *Department* and finally to *Year* of Degree conferment. At any stage the user can move up or down in the visualization or click through to the records that fit into that specific category in the repository itself. As users scroll over each section, "tool tips" or textual aides appear to facilitate clear navigation of the content.

In the initial assessment of our ETD metadata, we looked at the existing fields to determine which would create the most engaging visualization for discovery and browsing. We wanted to visualize metadata terms that helped users understand the subject matter of the ETDs, yet was broad enough to be quickly understood. In digital visualizations, there is also the opportunity to provide a layered or interactive experience, allowing users to refine the search using visual cues and provide multiple points to access the items themselves in the repository. Our visualization has three layers of information (i.e. college, department, and year) that are easily navigated with the click of a mouse to narrow or broaden the visualization by the visualized metadata elements. While this could be done with any normalized metadata element, we attempted to use metadata terms that were descriptive, comprehensible, and meaningful to a lay audience. With over 11,000 subject headings for 7,000 files, we ruled out the "topic" field [dc.subject.lcsh]. While this rich source of metadata is useful for search, a visualization of 11,000 non-hierarchical terms could add more confusion than clarity to our smaller number of files. Additionally, the metadata was generated by years of cataloging, so terms vary in the description of similar items. To facilitate understanding of the collection in relation to the university's structure

we used the institutional organization of *College*, *Department,* and *Year of Graduation* to help users understand the content as a whole and find specific items more easily. We exported this information from DSpace using the "Metadata Export" function. Then we stripped unnecessary columns from the resulting CSV file to access the pertinent data, i.e., "dc.date.issued," "dc.publisher," and "mus.relation.department" (local term) more easily.[2] Although we used only three fields, we found significant differences between current ETD records and those generated over the past 100 years. For the visualization to work, these needed to be normalized.

By looking at the data in aggregate, we readily identified inconsistencies that required significant initial metadata cleanup. Some were easy to correct and required little strategic thought. For example, typos, differences in special characters such as "&" vs. "and," varying punctuation, and inconsistent capitalizations were identified and cleaned using the "OpenRefine" tool.[3] Beyond those issues, we found inconsistencies resulting from the nature of historical records. These were more difficult to normalize, as we needed to make decisions about the metadata standards we would comply with. For example, the date term [dc.date.issued] could be "1932" or if the defense and graduation occurred in different years, the term could be "1932/1933." These were standardized to include only the year the item was created. Additionally, over time, names of departments changed. For example, a department that used to be known as "Horticulture" is now "Plant Sciences & Plant Pathology." Department names in the metadata were fixed to align with current university structure under the assumption that users looking for historical content would understand item groupings in the frame of our current structure while also more easily browsing a consolidated list of departments. Without historically consistent controlled vocabularies and standard metadata models, it is difficult to create easily browsed or searched records. This cleanup process is formulated based on philosophical as well as logistical considerations: it can be time consuming and data may be lost from the records as fields are edited to fit retroactive standards, yet it provides a possibility for enriched use of a collection. This initial cleanup also included filling in missing terms and fixing typos. This cleaned up metadata allows users to browse departments more easily due to factors such as decreasing the number of departments and research groups from over 140 to 69. We loaded this information back into the repository so both the repository and the visualization could take advantage of more consistent, normalized metadata. This process should only be necessary once: after the initial cleanup future records will follow current practices and will not require manual normalization.

[2] See https://scholarworks.montana.edu/xmlui/handle/1/2726?show=full for an example of full ETD metadata

[3] http://openrefine.org/

Building the Visualization

Based on our available data and our goal to create an interactive tool, we decided that the sunburst (figure 1) would be most appealing to users. Its circular shape is pleasing and clean, it loads in relatively short time, and it is easy to navigate without multiple instructions, even on mobile devices.

Next, we fit our data into the necessary format for the D3 Sunburst. This included using tab delimited files (TSVs) instead of comma separated (CSV) files. We found that the TSV format accommodated specific terms like college names "Education, Health & Human Development" better than the CSV format: although DSpace automatically generates CSV files, the comma in the name separated that term into two entries. We opted to use XOAI[4] instead of the basic OAI-PMH format to harvest all of our relevant metadata fields because the "mus.relation.department" field is not part of the DSpace default Dublin Core.

The organization of our repository was a key consideration in preparing the visualization. All of our theses and dissertations are in the same collection (col_1_733) which meant we did not have to concatenate collections to get the complete set. The higher-level community, a DSpace collection structure, also contains ETDs in process that we did not want to include in our visualization. A brief description of the process is in the Appendix, and the full code is posted on Github. We used the following basic steps to build the visualization:

1. We used the pyoaiharvester.py script to extract data from DSpace via XOAI and saved the extracted data in XML format.
2. Then we used the xml_parser.py to parse the xml file (created in step 1) into *college*, *department* and *date issued* fields and created the etds.tsv file with these new fields.
3. Using the D3 javascript library to complete the work of converting the flat TSV file into a hierarchical format (required by the sunburst), we generated the graphic itself.
4. To increase user friendly features and link the graphic to our ETD content we used additional javascript to create the tooltips and add in the links to ScholarWorks.
5. A "cron" job continuously updates the Sunburst visualization with new ETD content as it is added to the repository. The paths in the previous scripts had to be modified to ensure the transfer of data between systems (see appendix). The

[4] https://github.com/DSpace/xoai

file path indicating the location of etds.tsv in the xml_parser.py script had to be modified to match the location of the sunburst.js as the "cron" job could not run with the Python scripts in the production location of the sunburst.js file.

NEXT STEPS

By creating an interactive visualization, we hope to draw users in and entice them to engage with a collection of research that is uniquely tied to our institution. One of the motivations for this process was our desire to create a way to compare the graduate student output from each college and department at our university. This comparison gives potential students, new hires, and grant reviewers a quick reference for the composition of the university and encourages transparency and potentially collaborations between colleges and departments. Another benefit of this option for resource discovery is that it allows for the sort of serendipitous discovery that users celebrate in shelf-browsing. For instance in just two clicks a you can see that the earliest publication in "Home Economics" is 1924. One need not know what one is looking for before starting to look, like items can be grouped in multiple ways, and unfocused browsing is rewarded with full text documents.

We hope that with the understanding that metadata may be repurposed to construct visualizations, repository managers will have added incentive to apply consistent, meaningful metadata to each new item and collection and review current collections' metadata for flaws, inconsistencies, and missing data. It could also promote the use of tools that standardize the input of metadata. For example, interface developers could create environments that draw metadata from controlled vocabularies and drop-down menus instead of relying solely on freeform, manual entry (Surratt & Hill, 2004, p. 208). Even though we cannot plan for every future possibility, controlling the input clean metadata decreases the time spent on future recataloging projects (Boock and Kunda, 2009) or cleaning up human errors.

We hope that this paper will encourage other repositories to build visualizations and provide direction for developing their repositories further. Repositories should consider visualizing metadata elements to promote the use and awareness of scholarly items, guide infrastructure development, and invigorate the culture of sharing. Users can quickly and easily interpret visualizations of elements such as size, topic, interconnectedness, and growth. Although this project visualized ETDs, this process could be applied to any digital collection.

At the Montana State University library we plan to make future visualizations for the research data files and the journal articles in our repository. We also plan to make visual navigations for our abstract-only undergraduate content that has rich metadata about students,

their advisors, and community based projects. These projects will help users navigate and understand collections and allow the library to highlight collections that are deemed important, overlooked, or currently difficult to find.

Visualizations of digital repositories increase awareness of items within a collection, allow users to explore collections, and surface trends and patterns within collections that might otherwise remain unseen. Displays of this nature contribute to open source tools remaining competitive with other web-based platforms in general, and commercial repository/digital collection software, specifically. Of equal importance, their visual structure helps define and check the consistency of metadata application to large numbers of items.

REFERENCES

Andrew, T. (2004). *Intellectual property and electronic theses*. JISC Legal Information Services. https://www.era.lib.ed.ac.uk/handle/1842/612

Boock, M., & Kunda, S. (2009). Electronic thesis and dissertation metadata workflow at Oregon State University Libraries. *Cataloging & Classification Quarterly*, *47*(3–4), 297–308. https://doi.org/10.1080/01639370902737323

Borda, S. & Sterman, L. (2015) Visualizing electronic theses and dissertations. Presented at Open Repositories, June 8-11, 2015. Retrieved from: https://www.conftool.com/or2015/index.php?page=browseSessions&form_session=63

Brown, J. (2010). Literature review of research into attitudes towards electronic theses and dissertations (ETDs) [Working / discussion paper]. UCL Library Services/ SHERPA LEAP: London, UK. http://discovery.ucl.ac.uk/20424/

Card, S. K., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: Using vision to think* (1 edition). San Francisco, Calif: Morgan Kaufmann.

Collins, T., Mulholland, P., & Zdrahal, Z. (2005). Semantic browsing of digital collections. In Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen (Eds.), *The Semantic Web – ISWC 2005* (pp. 127–141). Springer Berlin Heidelberg. https://doi.org/10.1007/11574620_12

Copeland, S. (2008). Electronic theses and dissertations: Promoting "hidden" research. *Policy Futures in Education*, *6*(1), 87–96. https://doi.org/10.2304/pfie.2008.6.1.87

Copeland, S. (2011). Electronic theses and dissertations. In R. Rikowski (Ed.), *Digitisation Perspectives* (pp. 103–113). Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6091-299-3_6

Edminster, J., & Moxley, J. (2002). Graduate education and the evolving genre of electronic theses and dissertations. *Computers and Composition*, *19*(1), 89–104. https://doi.org/10.1016/S8755-4615(02)00082-8

Fekete, J.-D., Wijk, J. J. van, Stasko, J. T., & North, C. (2008). The value of information visualization. In A. Kerren, J. T. Stasko, J.-D. Fekete, & C. North (Eds.), *Information Visualization* (pp. 1–18). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_1

Jisc. (2013). Why is data visualization important. *Data Visualization Guide*. Retrieved from https://www.jisc.ac.uk/guides/data-visualisation/why-is-data-visualisation-important

Lewis, D. S. (2002). Electronic theses and dissertations. *Proceedings of Fifth International Symposium on Electronic Theses and Dissertations*. http://docs.ndltd.org/dspace/handle/2340/194

Lopatin, L. (2006). Library digitization projects, issues and guidelines: A survey of the literature. *Library Hi Tech*, *24*(2), 273 – 289. https://doi.org/10.1108/07378830610669637

Lucarella, D. & Zanzi, A. (1996). Information modelling and retrieval in hyper media systems. In M. Agosti & A. F. Smeaton (Eds.), *Information Retrieval and Hypertext* (pp. 121-135). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-1-4613-1373-1_1

Liu, Y.Q., (2004). Best practices, standards and techniques for digitizing library materials: A snapshot of library digitization practices in the USA. *Online Information Review*, *28*(5), 338 – 345. https://doi.org/10.1108/14684520410564262

Lyon, C. (2015). The state of ETDs in the Lone Star State: An update for 2014/2015. Retrieved from https://tdl-ir.tdl.org/tdl-ir/handle/2249.1/76360

McCutcheon, S. (2011). Basic, fuller, fullest. *Library Collections, Acquisitions, & Technical Services*, *35*(2–3), 64–68. https://doi.org/10.1016/j.lcats.2011.03.019

McCutcheon, S. & Kreyche, M., Maurer, M.B. & Nickerson, J.. (2008). Morphing metadata: Maximizing access to electronic theses and dissertations. *Library Hi Tech*, *26*(1), 41–57. https://doi.org/10.1108/07378830810857799

McMillan, G. (1996). Electronic theses and dissertations. *Cataloging & Classification Quarterly*, *22*(3–4), 105–125. https://doi.org/10.1300/J104v22n03_09

Park, E. G. & Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *The Electronic Library*, *29*(3), 394–407. https://doi.org/10.1108/02640471111141124

Palmer, C. L., Teffeau, L. C., & Newton, M. P., (2009). Strategies for institutional repository development: A case study of three evolving initiatives. *Library Trends* 57(2), 142–67. https://doi.org/10.1353/lib.0.0033

Phillips, S., Green, C., Maslov, A., Mikeal, A., & Leggett, J. (2007). Manakin: A new face for DSpace. *D-Lib Magazine*, *13*(11/12). https://doi.org/10.1045/november2007-phillips

Richardson, W. R., Srinivasan, V., & Fox, E. A. (2008). Knowledge discovery in digital libraries of electronic theses and dissertations: An NDLTD case study. *International Journal on Digital Libraries*, *9*(2), 163–171. https://doi.org/10.1007/s00799-008-0046-9

Rose, D. E. (1991). *A symbolic and connectionist approach to legal information retrieval* (Order No. 9130763). Available from ProQuest Dissertations & Theses Global. (303923027). http://search.proquest.com/docview/303923027?accountid=28148

Sarkar, P., & Mukhopadhyay, P. (2010). Designing single-window search service for electronic theses and dissertations through harvesting. *Annals of Library and Information Studies* 57(4), 354-364 http://hdl.handle.net/10760/17539

Schöpfel, J. (2013). Adding value to electronic theses and dissertations in institutional repositories. *D-Lib Magazine*, *19*(3/4). https://doi.org/10.1045/march2013-schopfel

Shreeves, S. L., & Teper, T. H. (2012). Looking backwards: Asserting control over historic dissertations. *College & Research Libraries News*, *73*(9), 532–535. http://crln.acrl.org/content/73/9/532.full

Shneiderman, B. (2002). Inventing discovery tools: Combining information visualization with data mining1. *Information Visualization*, *1*(1), 5–12. https://doi.org/10.1057/palgrave.ivs.9500006

Smeaton, A. F. (1996). An overview of information retrieval. In M. Agosti & A. F. Smeaton (Eds.), *Information Retrieval and Hypertext* (pp. 3–25). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-1373-1_1

Suber, P. (2008). Open access to electronic theses and dissertations. *DESIDOC Journal of Library & Information Technology*, *28*(1), 25–34. https://doi.org/10.14429/djlit.28.1.149

Surratt, B. E., & Hill, D. (2004). Etd2Marc. *Library Collections, Acquisitions, & Technical Services*, *28*(2), 205–223. https://doi.org/10.1016/j.lcats.2004.02.014

Weisser, C. R. ; W. (1997). Excerpted: Electronic theses and dissertations: Digitizing scholarship for its own sake. *Journal of Electronic Publishing*, *3*(2). https://doi.org/10.3998/3336451.0003.209

APPENDIX

The following steps outline our process. We have posted the complete code on Github.[5]

1. Used the pyoaiharvester.py script to extract data from DSpace via XOAI and saved the extracted data as xml:

```
python pyoaiharvester.py -l http://scholarworks.montana.edu/oai/
request -o etd.xml -s col_1_733 -m xoai
```

Code Sample 1. Executing pyoaiharvester.py

2. Used the xml_parser.py to parse the xml file (created in step 1) into college, department and date issued fields and created the etds.tsv file with these new fields.

```
for elem in record.findall(".//{http://www.lyncode.com/xoai}ele-
ment[@name='publisher']"):
                    pub = elem.find('*{http://www.lyncode.com/xoai}
      field')
                    coll = pub.text # slice this at the "," to get
      the college
                    college = coll.split(", ", 1)[1]
```

Code Sample 2. Extracting college

3. Used the D3 javascript library to complete the work of converting the flat tsv file into a hierarchical format (required by the sunburst) and to generate the graphic itself.

```
d3.tsv("etds.tsv", function(error, dataset) {
      var hierarchy = {
          key: "ETD",
          values: d3.nest()
            .key(function(d) { return d.college; }).sortKeys(d3.
ascending)
              .key(function(d) { return d.dept; }).sortKeys(d3.
ascending)
              .key(function(d) { return d.year; }).sortKeys(d3.
ascending)
```

[5] https://github.com/mutanthumb/ETD_sunburst

```
          .rollup(function(leaves) {
                        return leaves.length;
                    })
                    .entries(dataset)
              };
```

Code Sample 3. Using d3.nest to convert from flat file to hierarchical format

4. Used additional javascript to create the tooltips and add in the links to ScholarWorks.

```
      else if (d.depth == 2) { // Department
                var urlDept1 = "http://scholarworks.montana.edu/xmlui/handle/1/733/browse?value=";
                    var urlDept2 = "&type=department";
                      if (d.key.match(/&/g)) {
                        var dept = d.key.replace(/ /g,"+");
                        var dept = d.key.replace(/&/g,"%26");
                        var url =  urlDept1 + dept + urlDept2;
                  } else if (d.key.match(/\s/g)) {
                          var dept = d.key.replace(/ /g,"+");
                          var url =  urlDept1 + dept + urlDept2;
                        } else {
                        var url =  urlDept1 + d.key + urlDept2;
                        }
                        //Update the tooltip position and value
                        d3.select("#tooltip")
                       .html("<a href=\"" + url + "\">"  + d.key + "</a><br/><br/>" + d.value + " ETDs")
                        .style("left", (d3.event.pageX) + "px")
                        .style("top", (d3.event.pageY) + "px")
                        .transition()
                            .duration(500)
                            .style("opacity", 0)
                        .transition()
                            .duration(200)
                            .style("opacity", .9)
                        //.attr("fill-opacity", 1);
                      //Show the tooltip
                      d3.select("#tooltip").classed("hidden",false);
                      }
```

Code Sample 4. Creating links and tooltips for the "Department" ring

5. A "cron" job to continuously updates the Sunburst visualization. The paths in the previous scripts had to

be modified accordingly. The file path indicating the location of etds.tsv in the xml_parser.py script had to be modified to match the location of the sunburst.js as the "cron" job could not run with the Python scripts in the production location of the sunburst.js file.

```
20 6 * * 4 python /home/susan/pyoaiharvester.py -l http://scholarworks.montana.edu/oai/request -o etd.xml -s col_1_733 -m xoai
25 7 * * 4 python /home/susan/xml_parser.py
```

Code Sample 5. Cron job for automatically updating the Sunburst