Dissemination and Discovery of Diverse Data: Do Libraries Promote Their Unique Research Data Collections?

From “big deal” purchase agreements (Frazier, 2001) between libraries and journal publishers to sizeable increases in e-book holdings and on to larger innovations in scholarly communications, collection management in research libraries has undergone some significant change. As formats change and user expectations change [“if it’s online, I can find it myself”], libraries face new challenges in fulfilling their role in connecting users to the resources acquired and described. Libraries leverage catalogs, digital repositories, wayfinding across websites, and vended discovery layers to meet demand (Parry, 2014).

Among the resources that libraries make available to researchers, there is an increasing emphasis on the library’s role in connecting users to research data (in all of its formats). For example, in the pages of this journal, Witt (2012) made a strong case for the role of libraries in delivering data. Like many commentators, Witt points up the National Science Foundation’s mandate for grant proposals to include data management plans as one impetus for the library community to become more engaged with data services. His argument moves the conversation forward with a description of the process for libraries to work with campus research offices to develop a digital data repository and attendant services. There are more general commentaries sorting out efforts in North American and European libraries to establish themselves as centers for managing research data (Cox & Pinfield, 2014; Borgman, 2012; Corall, 2012; Carlson & Garritano, 2010). Amidst these varied suggestions, what emerges is an arising consensus that libraries must continue to evolve to meet the varied practices of users.

Part of this evolution is an acknowledgment of the expansive definition of research data beyond such established formats as numeric datasets or survey data. In *Sustainable Economics*

for a Digital Planet, the Blue Ribbon Taskforce on Sustainable Digital Preservation and Access (2010) notes that "[Data are] the primary inputs into research, as well as the first order results of that research" (p. 56). Add to that definition the National Science Board's (2005) categorization of data produced through observational, computational, and experimental means, and a larger context begins to emerge for defining data types and broadening expectations for the research infrastructures that contain them.

Within this context, it is easy to find descriptions of libraries' efforts to support the curation and dissemination of the many types of non-numeric digital objects that could meet the expanding definition of "data" in support of scholarly research. For example, within a recent single issue of the *New Review of Academic Librarianship*, which was focused on special collections in a digital age, the range of articles available included case studies of the digitization and preservation of in-house recorded media, especially lectures and special events (White, Bordo & Chen, 2015); a postcard collection (Ladd, 2015); a regional musical heritage collection (Doi, 2015); and the complete image collection of an internationally famous photographer, whose output spanned a number of formats (Harkema & Avery, 2015). While articles such as these focus on the professional decision-making and careful application of technical processes that are crucial for preservation and access of these materials, we are interested in examining how (or if) these unique objects are specifically positioned as inputs to research.

There are many examples, often presented within a discipline-specific context, that describe the ways that non-typical (non-numeric) materials can function as inputs to established or emerging research methods. McCay-Peet and Toms (2009) interviewed historians and journalists to better understand image-use patterns, specifically demarking differences between the use of images as objects or as data. Crucially, their findings suggest that images as data are of

growing importance to these researcher groups. Kononenko (2013) considered the use of sound files in research. Concerned about over-reliance on text-based transcripts of primary audio files, she highlights the "loss of the expressive qualities of speech", noting that "...intonation, inflexion, volume and other important indicators of meaning are lost" (p. 133). In a similar vein, Grimmer and Stewart (2013) provide a guide for political scientists to automated text analysis for large collections of digital texts. While acknowledging the power of automated content-analysis methods, they are quick to point out that solid research that involves text analysis requires careful thought and constant validation. And in a recent addition to this collection of investigations into data descriptions for different disciplines, Jackson Wheeler and Quinn (2015) explore how scholars from the performing arts research community use sound recordings and visual performances as research data.

Just as published books and articles provide foundational support for original research, so too do existing data lend themselves to being repurposed (or reused) in support of original research. But researchers' use of these data may be stymied by a lack of consistent methods for description and discovery. While most researchers understand processes and best practices for discovering books and articles that support their research agenda, the processes that researchers must undertake for discovering relevant research data seem to be haphazard and nonuniform—even with the aid of librarians or other information specialists.

While there's been a renewed focus on metadata and documentation to foster findability (and therefore use and reuse) of datasets, we hypothesize that local research data collections are generally not showcased—or even easily discoverable— from the library's website. If academic libraries are showcasing their local holdings—particularly digital texts, image files, audio

archives, and other non-numeric collections—we hypothesize that these are being promoted not as research datasets, but rather as artifacts of limited local and/or historical interest.

Because there is a longer history of government-funded, national-level initiatives to enable data curation and discovery in the United Kingdom and Canada than in the United States, we further hypothesize that to the extent individual libraries *are* successfully showcasing research data collections, it is likely that libraries in the UK and Canada will outperform their US counterparts in enabling such discoverability.

Methodology

We examine how academic libraries -- particularly at large research universities in the US, Canada, and the UK, where data curation and research data services are functions that have become embedded into the research infrastructure -- are enabling cross-institutional and interdisciplinary discovery and use of locally produced research data collections. To enable this examination, we first sought to identify a target universe from which to draw a reasonable sample. Our sample was drawn from member lists of the Association of Research Librarians (ARL) and Research Libraries United Kingdom (RLUK), membership organizations representing, respectively, 125 US and Canadian research libraries and 34 leading libraries in the UK and Ireland. To enable a sample, we transferred an alphabetical listing of the library names to an Excel spreadsheet, where a random number was assigned to each library after employing the random-number generator. This universe was sorted numerically to afford a sample of 20 ARL libraries and 20 RLUK libraries [n=40].

As libraries continue to refine the effectiveness of collection and service delivery via the Internet, these refinements have not yet resulted in uniform best practices to enable the description and discovery of research data. We borrow the concept of Universal Design (Preiser,

2011) to capture the essence that use of library collections ought to be broad as possible and specifically with objects as data not confined to niche audiences. So what then might be the implication for library websites? The function of websites in meeting goal-directed need is well documented (Singh & Dalal, 1999; Kim, 2011). Users see each website as a set of features or attributes with a capacity to meet their needs. Navigability and general wayfinding are key elements. We investigate these features and draw upon information theory to guide our analysis. Kim explains it plainly: "Success can be measured according to the website's purpose: to what extent does the website meet users' needs?" (Kim, 2011, p. 101).

We sought theoretical grounding for our coding schema, onto which we attempted to impose a (semi)structured series of steps to achieve a semblance of uniformity for the disordered chaos that characterizes human search behavior. The *Encyclopedia of Library and Information Science*'s entry on information searching and search models states, "searching can be defined as users' purposive behaviors in finding relevant or useful information..."(Xie, 2010, p. 2592). We especially considered tactics and strategy to inform our coding. Strategy is the multi-dimensional, planned approach to information retrieval while tactics are the individual moves a user might employ. We further draw inspiration from the concept of wayfinding (Arthur & Passini, 1992) used frequently in discussions of the built environment. This concept translates into the information sphere and is characterized by one of three different modes: locate, explore, and meander (Marchionini, 2006).

Once our search and browse techniques were settled, we set out to establish our code schema. As the building blocks for our analysis, a master list of codes based on themes related to the core elements of the study was developed. Codes are best if they are dichotomous, indicating presence or non-presence of a particular feature, yet not all codes are amenable to that best

practice. To combat problems, we wrote brief descriptions for each code and drafted guidelines for when to apply each. We then set out to perform a qualified intercoder reliability test, recognizing that judgements based on complex paths require intuitive decision-making. We were vigilant against the consequential bias, and our test for intercoder reliability was an attempt to mitigate such bias. Our tactic was to have each author code 20% of the sample (without knowing each other's results). Results were reviewed so that inconsistencies could be discussed and clarified. After satisfactory resolution, we established a revised codebook (see Appendix 1), reflecting revisions arising from the intercoder reliability test. The full sample of 40 library websites was then coded, with each author responsible for half of the sample, and each author's half reflecting an equal representation of RLUK and ARL libraries.

Results

Website Browsing

In order to determine if the growing emphasis on data services—including the collection and dissemination of datasets—among research libraries was manifested by the prominent positioning of data discovery tools, we first examined library homepages in search of an answer to these very simple questions: Are research datasets discoverable from the library's homepage and, if so, are local (institutional) datasets especially showcased?

From our sample, the answer to both of these questions is mostly "No" (see Figure 1). Among research libraries in the UK, only one institution's library homepage hinted at the presence of a digital data collection (in this case, image files). Three research libraries in the US pointed to research datasets from their homepage; among these (including the two institutions that featured two types of data), one institution pointed to numeric datasets (which were clearly local), two pointed to digital image files (one local), and two noted the availability of digital text

(linguistic corpora—both of which were local). Two Canadian libraries (a small number, but a large percentage of the Canadian libraries in our sample) pointed to a total of three types of data collections—two numeric datasets and one digital image collection.

[Place Figure 1 here]

This finding was not entirely unexpected and may be aligned with a trend toward uncluttered homepage design, so we proceeded to dig deeper in pursuit of data discovery. This involved a careful browse of each institution's library website, with a deliberate goal of finding pointers to specific types of research data. If any data were discovered, we expected that the most common of these would be numeric datasets. But we also searched for image files, voice recordings or other sound files, digital text (linguistic corpora), and other non-numeric datasets, including data that supports research in disciplines as diverse as chemistry, music, biology, design, physics, theater, and so on.

Not unsurprisingly, the website browse for numeric datasets yielded very fruitful results (see Figure 2)—although we did not expect that our search would be less successful within UK libraries (where we were unable to discover pointers to numeric datasets at six institutions) than within US and Canadian libraries (where only one library within our sample yielded a fruitless search). Among the 19 American and Canadian libraries that pointed to datasets from their websites, six of these specifically highlighted local (institutionally sourced) datasets, while the remaining 13 either pointed only to external datasets [N=9], or did not provide enough detail to determine the origin of the datasets [N=4]. Image data files proved to be even more prevalent than numeric data files (see Figure 3), as these were discovered after browsing their libraries' websites in fully 37 of the 40 institutions in our sample. Of these, 34 institutions (17 in the UK and 17 in America and Canada) pointed to unique local holdings.

[Place Figure 2 here] [Place Figure 3 here]

The presence of voice (sound) data files was more prevalent than anticipated (see Figure 4), as we were able to browse to these in 23 of the 40 institutions in our sample. Of this total, 22 featured local (institutionally-sourced) files, with more than twice as many from North American institutions (N=15) than from institutions in the UK (N=7). Of the 17 institutions where no voice (sound) files were found, 13 of these were in the UK and four in America and Canada.

[Place Figure 4 here]

When browsing for digital text files (e.g. linguistic corpora), only one UK library was found to have a data file of local (institutional) origin. Externally sourced digital text files were discovered at two more UK libraries and at three libraries in the US. A total of 34 libraries in our sample (17 in the UK, 14 in the US and three in Canada) revealed no pointers to digital text files as the result of a website browse. Similarly, the results of a browse for other non-numeric digital data yielded viable results at just 10 of the institutions in our sample (two in the UK, seven in the US, and one in Canada). Seven of these 10 pointed to local (institutionally sourced) data files (six in the US and one in the UK). The types of non-numeric data found from this browsing exercise were quite diverse, ranging from mouse and fly genomes to digital images of stage and costume designs from theater performances.

Results from Searching

After exhausting our exploration for data files through website browsing, we then turned to website searching in an attempt to discover data files that didn't reveal themselves through a methodical browse. Again, as described in the methodology, we drew on the core concepts of tactics and strategy and Xie's idea of "purposive behavior" to guide our efforts. It is also important to note that we deliberately chose to omit searching of OPACs, as wayfinding through

the website was our chief aim. An unexpected result from this phase was the absence of a library-specific, site-search tool at 10 of the institutions in our sample (nine in the UK and one in the US). Among the remaining 30 institutions in our sample, a search of the library website at 15 of these institutions (four in the UK, 10 in the US, and one in Canada) for numeric datasets turned up no new data files that had not been found by browsing (see Figure 5). In 11 of the institutions for which an initial browse had yielded fruitful results, a search of the library website turned up additional viable numeric data files (four in the UK, five in the US, and two in Canada). At three more of these institutions (two in the UK and one in the US), a search revealed numeric datasets after none had been found by browsing.

[Place Figure 5 here]

Our discovery of additional image data files was also quite successful when we supplemented our initial browse with a search of the libraries' websites (see Figure 6). Among the 30 institutions in our sample that offered a library-specific search function, a search yielded image data files in addition to the ones that were found by browsing in 15 of those institutions (seven in the UK, six in the US, and two in Canada), and viable results in three institutions where browsing had been unsuccessful.

[Place Figure 6 here]

At 14 of the institutions in our sample where browsing revealed the presence of voice and other audio data files, searching yielded additional viable results (see Figure 7). Ten of these instances were at US institutions, with two each at institutions in the US and Canada. Additionally, library website searches at five more institutions (three UK and two US) revealed sound-related data files where none had been found by browsing.

[Place Figure 7 here]

The small number of text files (linguistic corpora) that we found from browsing library websites was supplemented by viable additional results at nine institutions after searching library websites at the 30 institutions in our sample that offered a library-specific search function. Of these nine, five (two UK and three US) were at institutions where browsing yielded no results, and four (one UK and three US) at institutions where browsing was fruitful, but searching revealed even more results. It was interesting to note that these types of data files were not found at any of the Canadian institutions in our sample—either by browsing or by searching. Our search for other types of non-numeric data files was successful at nine of the institutions in our sample (seven in the US and one each in the UK and Canada), where searching turned up additional results to supplement results from browsing.

Conclusion and Discussion

The accessibility of research data holds great potential for advances in research. It allows the verification of study results and the reuse of data in new contexts. The role of academic libraries in this endeavor continues to evolve and may be influenced, and sometimes restricted, by conversations around open access that are beyond the scope of this study; considerations of existing discipline-specific data sharing practices; and the need for confidentiality by data creators and data users. The collection and dissemination of the subset of research data whose consumption could be regarded as non-rivalrous might prove to be the means by which libraries establish a stronger presence in this realm. This could be achieved by removing barriers to access and leveraging the longstanding infrastructure protocols and expertise of librarians to enable unfettered discovery, use, and analysis of all types of research data, as broadly defined by the examples in this study.

Creating rich collections and exposing them for discovery and use is the overarching goal of academic libraries. With the increasing digitization of library collections, it is reasonable for researchers to expect that these collections will be situated, and therefore easily findable, on libraries' websites. The impetus for this study was to examine how libraries present components of their collections as research data—and how users might subsequently find these collections. Fundamentally, as a descriptive analysis, this study points up some current patterns and provides a snapshot of how well academic libraries in the US, Canada, and the UK are presenting non-numeric collections as research data on their websites. We find that large UK and Canadian libraries do no better than their US counterparts at these tasks. While we found a substantial amount of non-numeric files to complement the expected presence of numeric datasets, these were rarely promoted as potential research datasets and were more likely to be a way of showcasing collections of institutional or historical interest.

While the results of our analysis suggest that libraries have been slow to make the leap from promoting most types of specialized digital collections as generic inputs to research (data) rather than as objects of local or institutional interest, this is not necessarily indicative of deficiencies or substandard professional practices. These results are more likely reflective of an ongoing transitional time, during which uniform expectations and standards have not yet been established. There are a number of reasons (including researchers' expectations) why it might not make sense to showcase locally created collections as research datasets (or potential datasets) from a library's website.

Admittedly, our attempt to replicate "typical search behavior" imposes a uniform process onto individuals (researchers) who, in reality, approach information-seeking in non-uniform ways. Moreover, this approach applies an assumption of standardization (i.e., all websites are

created equal) which, of course, is not true. This was apparent enough during the browsing exercise but was manifested even more strongly when browsing was supplemented by searching per our methodology. At some institutions, the website search function was powered by an inadequate search tool; at other institutions, the library website search offered up results from a library-hosted institutional repository, a resource that was not consistently present at all of the institutions in our sample.

Moreover, discussions around definitions of "research data" are ongoing, and there are multiple legitimate viewpoints concerning accurate and cogent descriptions of non-numeric digital collections as research datasets. The library web design community could be enlisted to provide advice on best practices to create interfaces that are rich and navigable for large heterogeneous digital collections, and that feature non-jargon-laden nomenclature that is understandable to researchers across disciplines who might consider library web pages as a starting point when seeking research data. Further explorations of the feasibility of implementing visual browsing while engendering contextual representations for unique local datasets also hold great promise.

References

Arthur, P., & Passini, R. (1992). *Wayfinding: People, signs, and architecture.* New York: McGraw-Hill.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable economics for a digital planet: Ensuring long-term access to digital information.* Retrieved from: http://brtf.sdsc.edu/publications.html

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63,* 1059-1078.

Carlson, J. R., & Garritano, J.R. (2010). E-science, cyberinfrastructure and the changing face of scholarship: Organizing for new models of research support at the Purdue University Libraries. Purdue Libraries Research Publications, #137. Retrieved from http://docs.lib.purdue.edu/lib_research/137/

Corrall, S. (2012). Roles and responsibilities: Libraries, librarians and data. In Pryor, G. (Ed.), *Managing research data* (pp. 105-133). London: Facet.

Cox, A. M., & Pinfield, S. (2014). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science, 46,* 299-316. doi:10.1177/0961000613492542

Doi, C. (2015). Local music collections: Strategies for digital access, presentation, and preservation—A case study. *New Review of Academic Librarianship, 21,* 256-263. doi:10.1080/13614533.2015.1022663

Frazier, K. (2001). The librarians' dilemma: Contemplating the costs of the "big deal". *D-Lib Magazine, 7*(3). Retrieved from http://www.dlib.org/dlib/march01/frazier/03frazier.html

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*, 267-297. doi:10.1093/pan/mps028

Harkema, C., & Avery, C. (2015). Milne *en masse*: A case study in digitizing large image collections. *New Review of Academic Librarianship, 21*, 249-255. doi:10.1080/13614533.2015.1034806

Jackson, A. S., Wheeler, J., & Quinn, T. (2015). Data services and the performing arts. *Music Reference Services Quarterly, 18*(1), 13-25. doi:10.1080/10588167.2015.997072

Kim, Y. (2011). Factors affecting university library website design. *Information Technology and Libraries, 30*(3), 99-107. doi:10.6017/ital.v30i3.1768

Kononenko, N. (2013). Groupsourcing folklore sound files: Involving the community in research. *Canadian Slavonic Papers, 55*(1-2), 131-151.

Ladd, M. (2015). Access and use in the digital age: A case study of a digital postcard collection. *New Review of Academic Librarianship, 21*, 225-231. doi:10.1080/13614533.2015.1031258

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM, 49*(4), 41-46. doi:10.1145/1121949.1121979

McCay-Peet, L., & Toms, E. (2009). Image use within the work task model: Images as information and illustration. *Journal of the American Society for Information Science and Technology, 60*, 2416-2429. doi:10.1002/asi.21202

National Science Board, (2005). *Long-lived digital data collections: Enabling research and education in the 21st century.* Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

Parry, M. (2014, April 21). As researchers turn to Google, libraries navigate the messy world of discovery tools. *The Chronicle of Higher Education*. Retrieved from https://chronicle.com/article/As-Researchers-Turn-to-Google/146081

Preiser, W. F. E., & Smith, K. H. (2011). *Universal design handbook* (2nd ed.). New York: McGraw-Hill.

Singh, S. N. & Dalal, N. P. (1999). Web home pages as advertisements. *Communications of the ACM, 42*(8), 91-98. doi:10.1145/310930.310978

White, H., Bordo, M., & Chen, S. (2015). Digitizing and preserving law school recordings: A Duke Law case study. *New Review of Academic Librarianship, 21,* 232-240. doi:10.1080/13614533.2015.1024871

Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration, 52*, 172-188. doi:10.1080/01930826.2012.655607

Xie, I. (2010) Information searching and search models. *Encyclopedia of Library and Information Sciences (3rd ed.),* Taylor & Francis, pp. 2592-2604. doi: 10.1081/E-ELIS3-120043745

Appendix 1

Code Schema

Code	Label/Description
1	No mention of datasets, images, etc. on library homepage
1.1	Library homepage mentions numeric data or datasets
1.11	Numeric data or datasets mentioned on library homepage are external to the institution
1.12	Numeric data or datasets mentioned on library homepage include local (institutional) collections
1.2	Library homepage mentions image collections
1.21	Image collections mentioned on library homepage are external to the institution
1.22	Image collections mentioned on library homepage include local (institutional) collections
1.3	Library homepage mentions voice (or sound) collections
1.31	Voice (or sound) collections mentioned on library homepage are external to the institution
1.32	Voice (or sound) collections mentioned on library homepage include local (institutional) collections
1.4	Library homepage mentions digital texts
1.41	Digital text collections mentioned on library homepage are external to the institution
1.42	Digital text collections mentioned on library homepage include local (institutional) collections
1.5	Library homepage mentions other digital non-numeric collection(s)
1.51	Other digital non-numeric collections mentioned on library homepage are external to the institution
1.52	Other digital non-numeric collections mentioned on library homepage include local (institutional) collections
2	Unable to readily browse to numeric data terms on library website
2.1	Numeric data terms found by browsing library website include datasets
2.2	Numeric data terms found by browsing library website point to external datasets only (ICPSR, government resources, etc.)
2.3	Data terms point to numeric data or datasets of local (institutional) origin
3	Unable to readily browse to image collections on library website
3.1	Image collections found by browsing library website; non-local collections only
3.2	Some image collections are of local (institutional) origin
4	Unable to readily browse to voice (sound) collections on library website
4.1	Voice (sound) collections found by browsing library website; non-local collections only
4.2	Some voice (sound) collections are of local (institutional) origin

5	Unable to readily browse to digital text collections on library website
5.1	Digital text collections found by browsing library website; non-local collections only
5.2	Some digital text collections are of local (institutional) origin
6	Unable to readily browse to other digital non-numeric collections on library website
6.1	Other digital non-numeric collections found by browsing library website; non-local collections only
6.2	Some other digital non-numeric collections are of local (institutional) origin
7	Could not readily browse to numeric data terms, and a search of the library website also turned up nothing
7.1	Could not readily browse to numeric data terms, but a search of the library website turned up viable results
7.2	Numeric data terms were found by browsing the library website, and a search turned up additional viable results
7.3	Numeric data terms were found by browsing the library website, but a search turned up nothing new
7.4	No library website search is available
8	Could not readily browse to image collections, and a search of the library website turned up nothing
8.1	Could not readily browse to image collections, but a search of the library website turned up viable results
8.2	Image collections were found by browsing the library website, and a search turned up additional viable results
8.3	Image collections were found by browsing the library website, but a search turned up nothing new
8.4	No library website search is available
9	Could not readily browse to sound recordings (voice or other), and a search of the library website turned up nothing
9.1	Could not readily browse to sound recordings (voice or other), but a search of the library website turned up viable results
9.2	Sound recordings were found by browsing the library website, and a search turned up additional viable results
9.3	Sound recordings were found by browsing the library website, but a search turned up nothing new
9.4	No library website search is available
10	Could not readily browse to digital text collections, and a search of the library website turned up nothing
10.1	Could not readily browse to digital text collections, but a search of the library website turned up viable results
10.2	Digital text collections were found by browsing the library website, and a search turned up additional viable results
10.3	Digital text collections were found by browsing the library website, but a search turned up nothing new
10.4	No library website search is available

11	Could not readily browse to other non-numeric collections, and a search of the library website turned up nothing
11.1	Could not readily browse to other non-numeric collections, but a search of the library website turned up viable results
11.2	Other non-numeric collections were found by browsing the library website, and a search turned up additional viable results
11.3	Other non-numeric collections were found by browsing the library website, but a search turned up nothing new
11.4	No library website search is available