Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Assessing Research Data Deposits and Usage Statistics within IDEALS

Christie A. Wiley

University of Illinois at Urbana – Champaign, Champaign, IL, USA

Abstract

Objectives: This study follows up on previous work that began examining data deposited in an institutional repository. The work here extends the earlier study by answering the following lines of research questions: (1) What is the file composition of datasets ingested into the University of Illinois at Urbana-Champaign (UIUC) campus repository? Are datasets more likely to be single-file or multiple-file items? (2) What is the usage data associated with these datasets? Which items are most popular?

Methods: The dataset records collected in this study were identified by filtering item types categorized as "data" or "dataset" using the advanced search function in IDEALS. Returned search results were collected in an Excel spreadsheet to include data such as the Handle identifier, date ingested, file formats, composition code, and the download count from the item's statistics report. The Handle identifier represents the dataset record's persistent identifier. Composition represents codes that categorize items as single or multiple file deposits. Date available represents the date the dataset record was published in the campus repository. Download statistics were collected via a website link for each dataset record and indicates the number of times the dataset record has been downloaded. Once the data was collected, it was used to evaluate datasets deposited into IDEALS.

Results: A total of 522 datasets were identified for analysis covering the period between January 2007 and August 2016. This study revealed two influxes occurring during the period of 2008-2009 and in 2014. During the first timeframe a large number of PDFs were deposited by the Illinois Department of Agriculture. Whereas, Microsoft Excel files were deposited in 2014 by the Rare Books and Manuscript Library. Single-file datasets clearly dominate the deposits in the campus repository. The total download count for all datasets was 139,663 and the average downloads per month per file across all datasets averaged 3.2.

Conclusion: Academic librarians, repository managers, and research data services staff can use the results presented here to anticipate the nature of research data that may be deposited within institutional repositories. With increased awareness, content recruitment, and improvements, IRs can provide a viable cyberinfrastructure for researchers to deposit data, but much can be learned from the data already deposited. Awareness of trends can help librarians facilitate discussions with researchers about research data deposits as well as better tailor their services to address short-term and long-term research needs.

Introduction

Large-sized academic libraries were the early adopters of Institutional Repositories (IRs), and early development of IRs focused on the accumulation, preservation and dissemination of faculty research output in an openly accessible way (Xia & Opperman 2010). Research suggests that only between 15% and 30% of eligible scholars and researchers deposit their work in institutional repositories (Cullen & Chawner 2011). A recent study of scientists and engineers indicated researchers were not aware of the campus repository Wiley & Mischo 2016). Further, a 2015 survey of 327 researchers at UIUC revealed that only 26% of survey respondents were aware of the campus's IR and an even smaller percentage, 12%, utilize the resource (Towns et al. 2015).

Despite challenges in recruiting content and building awareness, IRs have become an established component in the scholarly communication landscape. The IR at UIUC has over 85,000 items. Campus repositories are intended to showcase the research output of an academic or research institution, including research data. Research data as an output may consist of numeric datasets, collections of image files, audio archives, digital texts, and other nonnumeric resources. Many libraries, including UIUC continue to develop research data service programs and infrastructure to support faculty research needs. These needs have increased because of the growth of data-intensive science and federal agency mandates. Assessing the research data deposited within campus repositories allows librarians, research data services staff, and repository staff to evaluate the existing data content and inform future work. For example, the complexity of data is frequently discussed as a challenge for sharing and preservation since multiple files, file types, and interdependence are often expected (Plale et al. 2013). For several years, workshops and consultations at UIUC have emphasized the importance of data documentation, which also may result in additional files. But in practice, do data in the IR most often contain multiple files, especially multiple files of different file types?

Additionally, as academic libraries work to increase awareness and contributions to IRs, being able to demonstrate the utility of data is important. Research done in parallel to this work has revealed that many researchers doubt that anyone is interested in their data (Wiley & Burnette 2017, forthcoming). Can that doubt be addressed through analysis of downloads for datasets, specifically?

In 2015, a study was conducted on the UIUC campus repository to begin assessing item records categorized as "data" or "dataset." The results of the earlier study revealed that text files were the most frequently deposited file type, followed by Excel spreadsheets and PDFs. A variety of research disciplines and communities were represented, but deposits were dominated by just a few areas like the Illinois Department of Agriculture and Rare Books and Manuscripts. The goal of this follow-up study is to look at deposited data more closely to answer the following lines of research questions: 1) what is the file composition of datasets ingested into the UIUC campus repository, IDEALS? Are datasets more likely to be single-file or multiple-file items? and 2) what is the usage data associated with these datasets? Which items are most heavily downloaded? Can we start to determine if popularity is steady over time or does it fluctuate?

Literature Review

A university-based IR is a mechanism for capturing, archiving, and managing digital research outputs of the institution (Marsh 2015). In recent years, this content has expanded to include institutional records, digitized materials, and research data. The value of good research data management and practices become more apparent as research funders place ever greater importance on data as an output of research (Ball et al. 2012).

Research data is not solely described as material underlying conference papers, journal articles, and books. Data is defined as any information that can be stored in digital form, including text, numbers, images, video or movies, audio software, algorithms, equations, animations, models, simulations (National Science Board 2005). Yet this definition is open and subject to a lot of interpretation. Researchers can share data through deposit in a data center, archive, or institutional repository (IR), through submission to a journal as supplements to articles, through discrete publication, websites, and peer exchange (Akers & Doty 2013; Van den Eyden et al. 2010, Wallis et al. 2013).

Academic libraries are increasingly sources of infrastructure and research support in the area of data stewardship (Akers & Doty 2013). IRs may support project conception, proposal development, scheduling, documenting, embargoing, and communicating within and among research groups, data exchange, and storage (Kunda & Anderson-Wilk 2011; Ray 2014). Institutional repositories have the ability to manage scholarship, data, software tools, and code (Cragin et al. 2010; Walters 2014).

Although IRs have become a more integrated library service at large academic institutions, the literature indicates repository managers experience issues obtaining faculty cooperation in content acquisition (Xia & Opperman 2010). An initial investigation conducted on repository users and repositories in New Zealand found users were more interested in externally developed, discipline-specific repositories than in repositories housed at their own institutions (Cullen & Chawner 2010). A follow-up study in 2011 revealed ongoing barriers to depositing in IRs to include faculty and institutional repository staff workload, challenges of IR use, lack of awareness, and concern of data confidentiality (Chawner & Cullen 2011) and a lack of awareness of the institutional repository and the deposit process among Texas A&M University faculty (Yang & Li 2015).

Research suggests that metrics can be used to understand how repositories are used, and this information informs policy decisions on future investment (Kelley et al. 2012). Counts of item downloads are among a number of metrics that should be assembled based on institutional mission and on audience (Bruns & Inefuku 2016). The ability of a system to make available the number downloads and views of full-text files is listed as one of the top critical success factors for IR (Lagzian, Abrizah & Wee 2015). Libraries determine the most appropriate benchmark for success within their respective IR (Fralinger & Bull, 2013). Furthermore, the work of Kratz and Strasser found that researchers value download counts second only to citation (Kratz & Strasser 2015).

Recent IR literature indicates there are distinct perspectives on content recruitment, use/non-use of IR, awareness of campus repositories, and researchers' willingness to contribute to campus repositories and assessment. Although this literature is insightful, it does

not indicate the importance of assessment in the context of research data within repositories. Examining the file composition of dataset and associated usage of datasets within the UIUC. IR provides an important opportunity to examine trends in data deposits and adjust expectations, services, and/or recruitment strategies as appropriate.

Methods

The dataset records collected in this study were identified by filtering item types categorized as "data" or "dataset" using the advanced search function in IDEALS. Returned search results were collected in an Excel spreadsheet to include data such as the Handle identifier, date ingested, file formats, composition code, and the download count from the item's statistics report. The Handle identifier represents the dataset record's persistent identifier. Composition represents codes that categorize items as single or multiple file deposits; for example "Multiple-Mixed" indicates the item contains multiple files of mixed file types. Date available represents the date the dataset record was published in the campus repository.

Download statistics were collected via a website link for each dataset record and indicate the number of times the dataset record has been downloaded. Specifically, download counts for an item in IDEALS go up by 1 for each individual file downloaded from the item. Known bots and crawlers are blocked, and only one download is counted per IP address per calendar day to avoid over-estimation.

With this data collected, each dataset record was coded for "composition" as either a single-file or multiple-file dataset, along with the associated file types. Further, to compare dataset downloads in IDEALS, average downloads per month was calculated by subtracting the month/year of ingest from the month/year of date downloads were recorded for this study. This gives the duration of time available for downloads to accumulate and was used to divide into the downloads recorded to calculate the average download per month. While this number represents the entire dataset, it potentially skews the results towards datasets that have a larger number of files, since those add cumulatively towards an entire dataset's total download counts. To account for this, the Average Downloads/Month/File was also calculated by dividing Average Downloads/Month by the total number of files in the dataset.

Results

A total of 522 datasets were identified for analysis, covering the period between January 2007 and August 2016. Two major influxes occurred during this time frame. The first occurred in the years 2008-2009 when a large number of PDF reports were deposited by the Illinois Department of Agriculture (IDOA). The second influx occurred in 2014 when the Rare Books and Manuscripts Library (RBML) deposited a large number of publication lists as Excel spreadsheets. The total download count for all 522 datasets is 139,663.

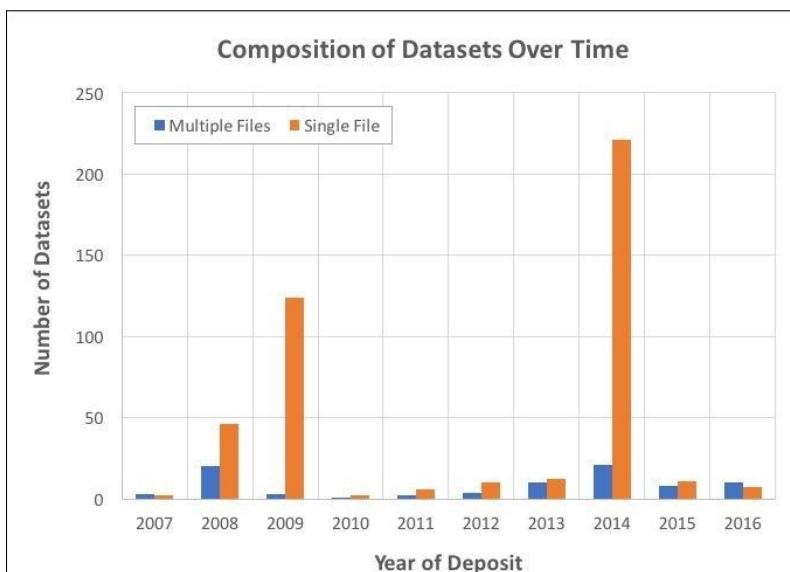*Composition as Single-File or Multiple-File Datasets*

The general guidelines for depositing content IDEALS requires that the work be wholly or in part produced or sponsored by UIUC faculty, researchers, staff, or students. Undergraduate students may submit work under the sponsorship of a faculty member. IDEALS accepts research no matter the file format, and currently does not have any requirements or guidance

for data documentation such as readme files or codebooks. Table 1 shows the breakdown of datasets by composition of single and multiple files deposited.

Table 1: Composition of IDEALS Datasets

Multiple-File Datasets	Count	Single-File Datasets	Count
Multiple Mixed	73	Single Excel	208
Multiple CSVs	3	Single PDF	186
Multiple PDFs	3	Single QuickTime Video	11
Multiple TIFFs	1	Single Zip	10
Multiple XMLs	1	Single AVI Video	9
Multiple Zips	1	Single AIFF Audio	4
Total	**82**	Single CSV	3
		Single HTML	2
		Single JPEG	2
		Single Unknown	2
		Single Text	2
		Single PowerPoint	1
		Total	**440**

Single-file datasets clearly dominate deposits. To account for the influxes mentioned above by IDOA and RBML, as well as the potential for the changing nature of data deposits, the composition was also examined over time in Figure 1. These results indicate that over a nine-year period, the proportions of single-file versus multiple-file datasets into the institutional repository remain variable.

Figure 1: The number of single-file or multiple-file datasets deposited over time in IDEALS.

Usage Statistics

The second line of research questions concerns usage data associated with these dataset records. The usage statistics report lists a cumulative count of all downloads for an item and a line graph that displays accumulation of download counts over time as well as a bar graph of download counts by month for the item.

The average downloads/month/file across all 522 datasets averaged 3.2, with a range of 0 – 63.9 and a median of 2.1. The top 10 datasets with the highest number of average downloads per month per file are listed in Table 2. All but one of the top 10 datasets are comprised of just one file. When the number of files included in the dataset is not taken into consideration, the top dataset is a physics dataset deposited in May of 2015 titled "Boloscope Scans" (Handle 2142/78815). This dataset contains 1,097 individual files and accumulated 19,098 downloads by September 2016. However, when compared with other datasets' average downloads per month per file, the Boloscope dataset ranked 467th, demonstrating the importance of how downloads are counted for datasets.

Table 2: Top 10 datasets with the highest number of Average Downloads/Month/File

Rank	Handle	Title	Total Downloads	Time Available (Months)	Total Files & Composition	Average Downloads / Month / File
1	2142/45709	List of stop words used in topic modeling journals, summer 2013	2475	36	1 - Single Text	68.8
2	2142/88828	Telecommunication Services Provision in Nigeria	511	8	1 - Single PDF	63.9
3	2142/8775	Hard times farming: Farming during the Great Depression of 1930-1940	5734	99	1 - Single HTML	57.9
4	2142/3495	Identifying the Core Periodical Literature of the Agricultural Communications Documentation Center	5398	104	1 - Single PDF	51.9
5	2142/3483	What Can I Expect to Earn: Information Sources for Library Salary Negotiations	8680	105	2 - Multiple Mixed	41.3
6	2142/13183	Modeling and Comparison of Primary and Secondary Refrigeration System Performance	2424	86	1 - Single PDF	28.2
7	2142/78980	ARL Academic Library Directors Study	311	13	1 - Single Excel	23.9
8	2142/54952	Michigan Top 3 Treasures Results	570	24	1 - Single Excel	23.8
9	2142/88863	Urbana Irrigation Controls Study	134	7	1 - Single PDF	19.1
10	2142/50053	Bibliography of Electro-Organic Chemistry Part I	459	25	1 - Single PDF	18.4

One benefit of examining usage statistics for datasets in a repository that has a long history is the ability to start looking at trends over time. As libraries and other organizations take on long-term responsibility of creating and stewarding data collections, it would be helpful to know what kind of patterns of use may occur. For example, through the statics report graphs, we can look at trends of top 10 heavily downloaded datasets that have been available for a long period of time. Three of these datasets that have been available for greater than eight years are used as an example in Figure 2 and show very different usage patterns. One dataset shows high downloads at first with less activity in later years, another shows steady use during the entire time available, and yet another shows a very abrupt increase in use a few years after initial ingest. Understanding these patterns, or at least being aware of them, will be important to future management of data repositories.
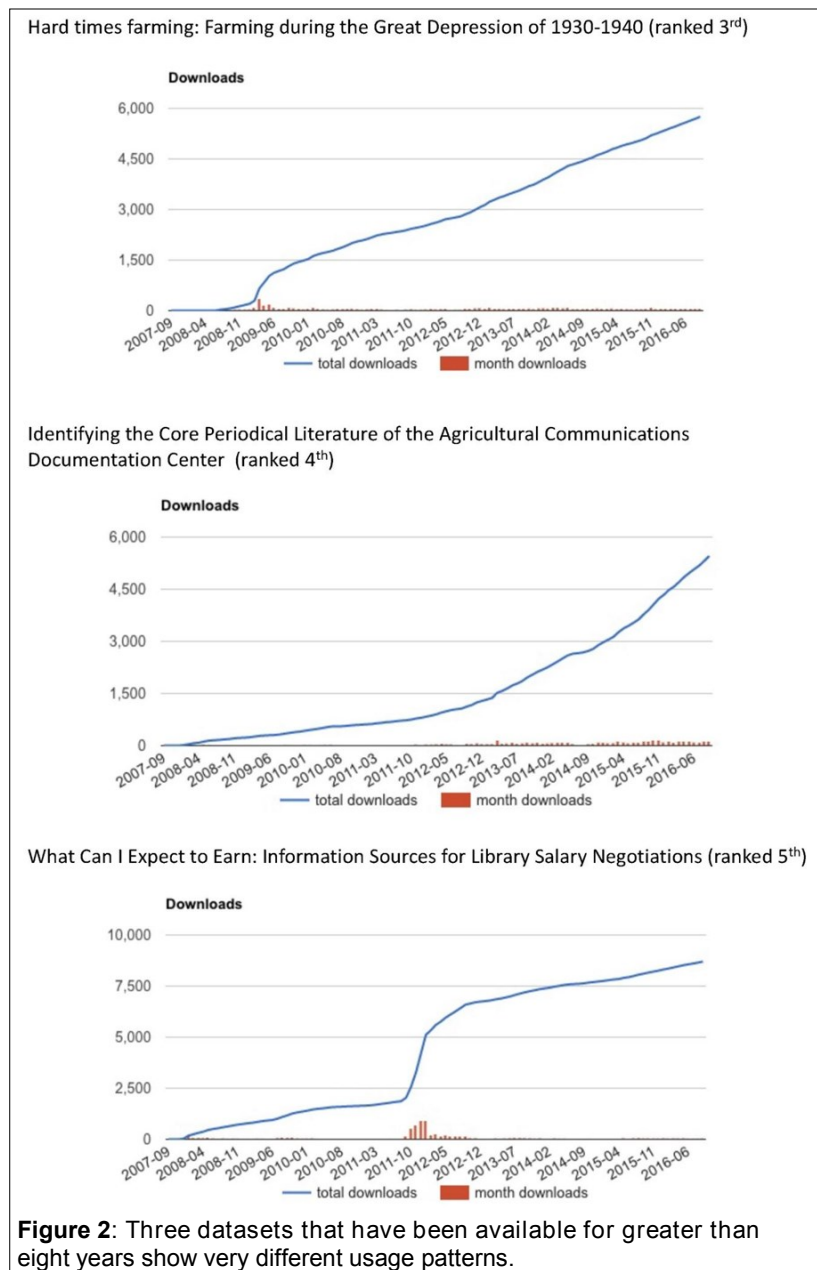
Figure 2: Three datasets that have been available for greater than eight years show very different usage patterns.

Discussion

The first question of this study sought to determine the file composition of datasets ingested into the campus repository. This study revealed that there were a high number of PDFs categorized as datasets, which are heavily accessed and several appear in the top 10 within the campus repository. Can they be considered data? Some people find this questionable (Anonymous, 2013). However, researchers have shown a strong preference for the portable document format, and PDF use is continually expanding in new ways. For example, in the early 1990s, it was suggested that extensive tagging and indexing articles would lead to more targeted reading (Anonymous, 2013). Although this expectation was unfulfilled, publishers are enhancing their online platform, improving data representation and display within PDFs. This has created various opportunities in which publishers provide various opportunities for users to visualize data.

For example, Journal of Cell Biology has been providing access to raw image data through data viewer thus allowing improved access to the presentation of data within articles. Nature Methods began integrating supplementary videos into the main text of manuscripts by using embedded website links that open a pop-up window for viewing the video without disrupting the reading process (Anonymous, 2013). Thus, a primary value of PDFs are that they can be stored and transported between devices and read without access to specialized software or the internet. This appears to be another surprising way PDFs are being used and even categorized as "data" during deposit.

Another question from this study inquired if datasets are more likely to be single-file or multiple-file items. Complexity can manifest in a number of ways, with the number of files being just one example. While not all types of complexity could be accommodated in IDEALS, multiple file items are readily ingested in this system. However, somewhat surprisingly, this work shows that majority of data contain just a single file. In collaboration with the Research Data Service (RDS), librarians at UIUC provide data management instructional sessions to environmental science, aerospace and engineering graduate study groups. More multiple file datasets, especially over time, could be anticipated due to the importance of documentation being taught in data management workshops. Although we could look at this more carefully by examining the 73 Multiple-Mixed datasets to see if documentation files are included, the prominence of single-file items suggests that documentation is not being regularly included in data deposits into this IR.

The second question in this study inquired about the usage data associated with these datasets. This study shows that download counts are useful, but have to be calculated carefully. This is because counting data is more complicated than journal counts of articles. Data many have multiple files and numerous versions. One dataset can also be part of or derived from another dataset. This is a known problem, but a solution is being addressed through the Counter for Data Usage standard currently being developed as of August 2017 (Data Cite Blog, 2017). The first draft of the release of this standard specifically targets research data usage. Overall the goals of COUNTER are to limit the usage data to human users and filters out all known robots, crawlers and spiders, include the volumes of data reported being transported through the variations of e-resources, and enable the reporting of usage stats by different data repositories.

IDEALS launched in 2007, before research data management and funding agency mandates were at the forefront, and many institutional repositories allow researchers to deposit any format of research data. This idea is understandable because it provides a flexible system to meet researchers' varied needs, so it is not surprising nor incorrect that the content is varied. Yet future directions should continue to explore what data within institutional repositories "should look like" and why this is sufficient and useful. This area is ripe for further analysis. Research is needed to understand trends of data deposit and use in order to accurately articulate resource needs to library and campus administration and inform future accession and preservation policies and practices. For example, in 2016 UIUC launched a dedicated data repository, the Illinois Data Bank as a sibling repository to IDEALS (Fallaw et al. 2016). While the Illinois Data Bank was developed to better accommodate datasets specifically, analysis of datasets deposited previously in IDEALS provides a useful reference point.

Conclusion

While this is just one study that represents a snapshot in time, academic librarians, repository managers, and research data services staff can use the results presented here to begin anticipating the nature of research data that may be deposited within institutional repositories. With increased awareness, content recruitment, and improvements, IRs can provide a viable cyberinfrastructure for researchers to deposit data. As data services continue to grow in academic libraries, much can be learned from the data already deposited to inform future practices. Being aware of trends can help librarians facilitate discussions with researchers about research data deposits as well as better tailor their services to address short- and long-term research needs.

Data Availability

Data collected for this study is available in the Illinois Data Bank at
https://doi.org/10.13012/B2IDB-1235375_V1

Disclosure

The author reports no conflict of interest.

References

Akers, Katherine G. and Jennifer Doty. 2013. "Disciplinary differences in faculty research data management practices and perspectives." *International Journal of Digital Curation* 8(2): 5-26.
http://dx.doi.org/10.2218/ijdc.v8i2.263
Anonymous. 2013. "Beyond the PDF." *Nature Methods* 10(2). http://dx.doi.org/10.1038/nmeth.2363

Ball, Alexander, Mansur Darlington, Thomas Howard, Chris McMahon and Steve Culley 2012. "Visualizing research data records for better management." *Texas Digital Library* 13(1).
https://journals.tdl.org/jodi/index.php/jodi/article/view/5917

Bruns, Todd and Harrison W. Inefuku. 2015. "Purposeful Metrics: Matching Institutional Repository Metrics to Purpose and Audience." In *Making Institutional Repositories Work*, edited by Burton B. Callicott, David Scherer and Andrew Wesolek, 213-234. West Lafayette, IN: Purdue University Press, 2015.

Cragin, Melissa H., Carole L. Palmer, Jake R. Carlson, and Michael Witt. 2010. "Data sharing, small science and institutional repositories." *Transactions of the Royal Society*. 368(1926): 4023-4038. http://www.dx.doi.org/10.1098/rsta.2010.0165

Cullen, Rowena, and Brenda Chawner. 2010. "Institutional repositories: Assessing their value to the academic community." *Performance Measurement and Metrics* 11:131-147. http://dx.doi.org/10.1108/14678041011064052

Cullen, Rowena, and Brenda Chawner. 2011. "Institutional Repositories, Open Access, and Scholarly Communication: A Study of Conflicting Paradigms." *Journal of Academic Librarianship* 37(6):460-470. http://dx.doi.org/10.1016/j.acalib.2011.07.002

Data Cite (Blog). *COUNTER Code of Practice for Data Usage Draft Update*. Last modified August 23. https://blog.datacite.org/counter-code-practice-for-data/

Fallaw, Colleen, Elise Dunham, Elizabeth Wickes, Dena Strong, Ayla Stein, Qian Zhang, Kyle Rimkus, Bill Ingram, and Heidi J. Imker. 2016 "Overly Honest Data Repository Development." *Code4Lib Journal* 34: 2016-10-25. http://journal.code4lib.org/articles/11980

Fralinger, Lauren and Jonathan Bull. 2013. "Measuring the international usage of US institutional repositories." *OCLC Systems & Services: International Digital Library Perspectives* 29(3): 134-150. http://doi.org/10.1108/OCLC-10-2012-0039

Jinha, Arif E. 2010. "Article 50 million: an estimate of the number of scholarly articles in existence." *Learned Publishing* 23(3): 258-283. http://dx.doi.org/10.1087/20100308

Kratz, John E. and Carly Strasser. 2015. "Making data count." Scientific Data 2:150039. http://dx.doi.org/10.1038/sdata.2015.39

Kelly, Brian, Nick Sheppard, Jenny Delasalle, Mark Dewey, Owen Stephens, Gareth J. Johnson, and Stephanie Taylor. "Open Metrics for Open Repositories." Paper Presented at *OR2012: the 7th International Conference on Open Repositories*, Edinburgh, United Kingdom, July 9-13, 2012. http://opus.bath.ac.uk/30226/

Kunda, Sue. and Mark Anderson-Wilk. 2011. "Community Stories and Institutional Stewardship: Digital Curation's Dual Roles of Story Creation and Resource Preservation." *portal: Libraries and the Academy* 11(4): 895-914. http://dx.doi.org/10.1353/pla.2011.0047

Lagzian, Fatemeh, A. Abrizah, and Mee Chin Wee. 2015. "Critical success factors for institutional repositories implementation." *The Electronic Library* 33(2): 196-209. http://doi.org/10.1108/EL-04-2013-0058

Marsh, Rebecca Mary. 2015. "The role of institutional repositories in developing the communication of scholarly research." *OCLC Systems & Services: International digital library perspectives* 31(4): 163-195. http://dx.doi.org/10.1108/OCLC-04-2014-0022

National Science Board. 2005. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." *National Science Foundation*, Arlington, VA. https://www.nsf.gov/pubs/2005/nsb0540/

Plale Beth, Robert H. McDonald, Kavitha Chandrasekar, Inna Kouper, Stacy Konkiel, Margaret L. Hedstrom, Jim Myers, and Praveen Kumar. 2013. "SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long Term Data Preservation." Practice Paper presented at the *8th International Digital Curation Conference*, Amsterdam, Netherlands, January 14-16 2012. http://hdl.handle.net/2022/15247

Ray, Joyce M., editor. 2014. *Research Data Management: Practical Strategies for Information Professionals.* Purdue University Press. http://www.thepress.purdue.edu/titles/format/9781557536648

Towns, John, David Gerstenecker, Laura Herriott, Ashley Hetrick, Heidi J. Imker, Chris Larrison, Marshall Scott Poole, Eric Gene Shaffer, Tracy Smith, Chuck Thompson, and Rob Watson. 2013. "University of Illinois Year of Cyberinfrastructure Final Report." *Illinois Research and Scholarship (Open Community).* https://www.ideals.illinois.edu/handle/2142/88444

White, Hollie C. 2014. "Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors." *Journal of Library Metadata* 14: 24-51. http://dx.doi.org/10.1080/19386389.2014.891896

Wiley, Christie 2015. "An Analysis of Datasets within Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the University of Illinois Urbana-Champaign Repository." *Journal of eScience Librarianship* 4(2): e1081. http://dx.doi.org/10.7191/jeslib.2015.1081

Forthcoming Wiley, Christie and Margaret Burnett. 2017. "Assessing Data Management Support Needs of Bioengineering and Biomedical Research Faculty."

Wiley, Christie and William H. Mischo. 2016. "Data Management Practices and Perspectives of Atmospheric Scientists and Engineering Faculty." *Issues of Science and Technology Librarianship* 85, Fall 2016. http://dx.doi.org/10.5062/F43X84NJ

Xia ,Jingfeng and David B. Opperman. 2010. "Current Trends in Institutional Repositories for Institutions Offering Master's and Baccalaureate." *Serials Review* 36:10-18. http://doi.org/10.1016/j.serrev.2009.10.003

Yang, Zheng Y (Lan) and Li, Yu. 2016. "University Faculty Awareness and Attitudes towards Open Access Publishing and the Institutional Repository: A Case Study." *Journal of Librarianship and Scholarly Communication* 3(1): p.eP1210. http://dx.doi.org/10.7710/2162-3309.1210