ISSN 2161-3974

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

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

Full-Length Paper

An Analysis of Datasets within Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the University of Illinois Urbana-Champaign Repository

Christie A. Wiley

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

Abstract

Objectives: The objective of this study is to identify: (1) how many datasets are within Illinois Digital Environment for Access to Learning and Scholarship (IDEALS); (2) which types of files are deposited in the repository; (3) which research methodologies are associated with these datasets; and (4) which research discipline or research communities are associated with these datasets within IDEALS.

Methods: Datasets collected in this study were found using the University of Illinois repository IDEALS website link https://www.ideals.illinois.edu. The keywords used were data or dataset. In order to facilitate analysis, datasets were analyzed using MS-Excel spreadsheets. They were coded by title, issue date, research methodology, research discipline, and community to explore patterns of use and the relationship to data management and research data services.

Results: There are 507 datasets in IDEALS dating from 1905-2015. Text files are the most frequently deposited file type; bibliographies represent 34% of the datasets; and, farming inventory lists are 26% of the datasets. Various research disciplines represent 18% of the datasets and research communities are associated with 78% of the datasets. 7% of the datasets are sponsored by NSF, NIH, IMLS and DOE funding agencies.

Conclusion: Understanding the file types, research methodologies, research disciplines and research communities within a university's current infrastructure, will provide a representation of the datasets and research supported within the university repository. It will enhance academic librarians and repository managers' data management conversations with researchers and provide information needed to needed to improve workflow deposit and batch loading. It will enhance research data services, meet researcher's needs, assess short-term preservation, and determine long-term preservation needs.

Correspondence: Christie A. Wiley: cawiley@illinois.edu
Keywords: research data management, institutional repository, data sets, academic library

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Introduction

Academic libraries have always been in the role of curating information and providing access to resources to support the work of parent institutions; this role is broadening to include digital resources and to accommodate data management requirements (Tarver & Phillips 2013). Many universities and libraries have infrastructures in the form of systems or educational programs for managing data sets (Carlson et al. 2010; Soehner et al. 2010). Universities provide centralized data storage for archiving, collaboratively working on, or sharing data. Data generated by university research is disseminated, managed, and preserved by the institutional repository. Some universities provide temporary storage while research projects are underway (DataStaR at Cornell University), providing collaborative open spaces for team members to communicate with each other and share project results with the science community (Steinhart et al. 2012). Academic researchers affiliated with University of Illinois Urbana-Champaign (UIUC) specified IDEALS as data deposit and sharing service (Mischo et al. 2014). This indicates that repositories have become platforms in which research team members can review and annotate data sets, share, assess, and interpret within groups.

The goal of IDEALS is to collect, disseminate, and provide persistent and reliable access to the research and scholarship of faculty, staff, and students at UIUC. The repository contains datasets that are often are supplemental files attached to theses and dissertations, as well as datasets associated with funding agency-required data management plans. Although this description of information informs users that datasets exist in the repository, the applicability of this knowledge within the context of data management, research data services, and scholarly communication is unknown. This study seeks to fill this gap. by examining the research datasets within the repository IDEALS. By analyzing datasets within the University repository, this study addresses the following research questions:

1. How many datasets are in the repository?
2. Which file types are the datasets deposited in the repository?
3. Which type of research methodologies are the datasets associated with?
4. Which research disciplines or communities do they represent?

Literature Review

Existing literature on institutional repositories (IR) defines their role as capturing, disseminating, and preserving the intellectual output of the institution (Baudoin & Branschofsky 2003). Intellectual output has been described as scholarly works and more broadly as digital materials created by institutions and community members (Aschenbrenner et al. 2008; Lynch 2003). Initially, the primary content of institutional repositories was pre- and post-prints of faculty research. Academic researchers use repositories to store theses, dissertations and technical reports as well as image and sound files (Henty 2008). Potential repository collections exist in digital formats or online, on departmental or faculty websites and in online cloud-based storage platforms.

Research data is defined beyond what is found in a spreadsheet or in a dataset and has evolved from flat files and data dumps. Data sources vary widely. For example, within the

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

physical and life sciences the data gathered and produced by researchers is observational, experimental, or models. Social science researchers' data ranges are collected within the field and produced from public services. Humanities data is obtained from records of human culture, varying from archival materials, published documents, or artifacts (Borgman 2007, 2009). The format of research data varies and includes generated digital content, special collections from the humanities and scientific disciplines, databases, spreadsheets, code, and images (Ray 2014; Steinhart et al. 2012). Datasets sometimes are dynamic, resulting from sensors (e.g. weather recording devices) automatically collecting data, and possibly requiring special interfaces between digital repositories and these data collection devices (Luce 2012). Datasets may be distributed across various national and international institutions. Making distributed data accessible requires not only new approaches for storage but also descriptors that work across disciplines and borders (Diekema et al. 2014).

Campus IRs are suited for discrete, static, and processed datasets (Mischo et al. 2014). IR contents contain access to administrative records, dissertations, grey literature, monographic records, and small datasets. This content expansion can be explained in various ways including institutional goals, motivations, data management funding agency requirements, and academic researchers' needs. The contents of IR consists of primarily journal articles, but many see the addition of a datasets as a logical fit (Alvaro et al. 2011; Hey & Hey 2006; Mullins 2009; Ramírez 2011). Extensive data management transcends local storage and is often outside the experiences of scientists who have seen changes within information environments and data requirements (Dikema et al. 2014). These changes will require librarians and archivists to learn about three important aspects of data management: the data life cycle, technical aspects (storage, indexing, retrieval), and social and policy issues (Qin & D'Ignazio 2010).

Recent IR literature examines motivation of use, usage statistics, workflow deposit models, content description, and assessment. Although this literature is valuable, it does not indicate the impact of research data within existing campus repositories. A description of the research datasets within IRs will provide librarians the ability to determine patterns-of-use and its applicability to evaluate how current structures is used by researchers.

Methodology

The datasets collected in this study were found using the University of Illinois repository IDEALS website link: https://www.ideals.illinois.edu. From this website link the advanced search feature was selected, then type was selected from the search type drop, and dataset or data were the keywords typed into the search box. In order to facilitate analysis, datasets were analyzed using MS-Excel spreadsheets.

The author assessed each dataset within the repository and coded them using the terms title, issue date, research type, research discipline, research publisher, publication status, sponsor, and peer review. Title represents the name of the dataset. Issue date indicates the date of the data. Research type indicates the type of research associated with the datasets. Research discipline indicates the research department associated with depositing data. Research community represents the publishers of datasets deposited by groups or organization not associated with an academic discipline. National Science Foundation (NSF), National Institute of Health (NIH), Department of Energy (DOE), Institute Museum Library Science (IMLS)

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

associated datasets were identified by the descriptor Sponsor within the repository. Publication and peer review indicates datasets associated with journal articles and technical reports; these were not peer reviewed. The categories were then organized and analyzed to explore patterns-of-use and the relationship to data management and research data services.

Results

Number of datasets

The number of datasets within this study can be found using the method illustrated in Figure 1. Searching the website link https://www.ideals.illinois.edu returned results of 507 datasets.

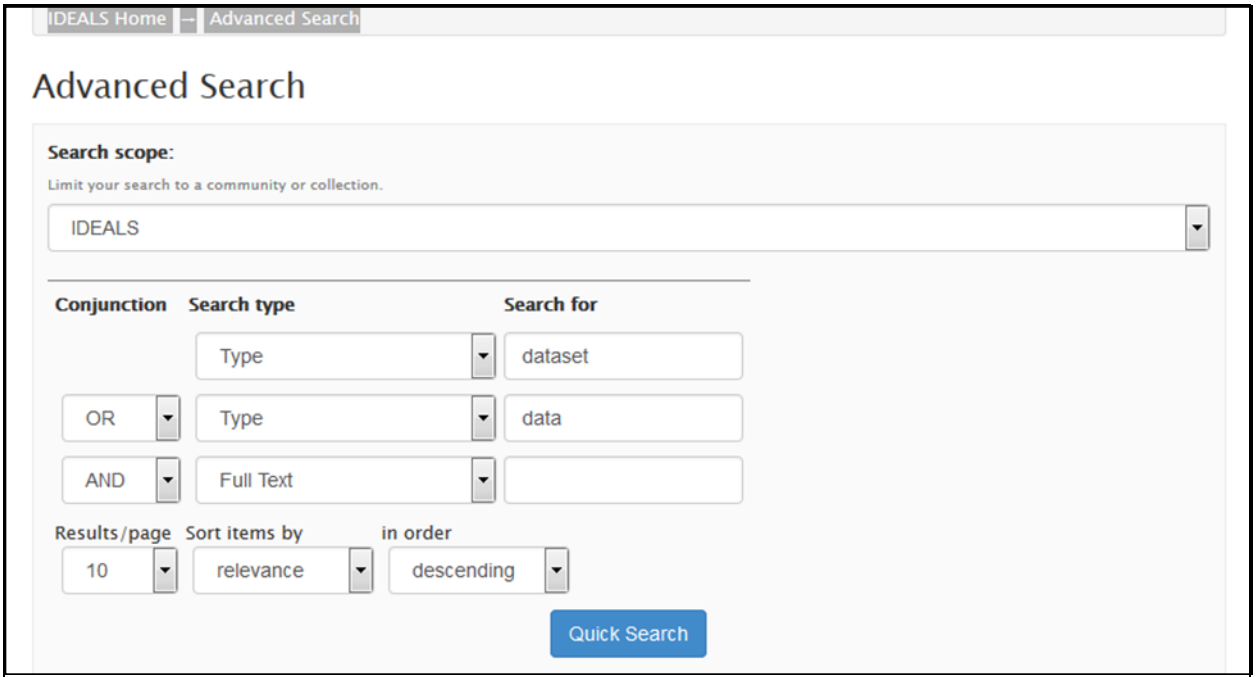


Figure 1: Searching the IDEALS Repository

Datasets in IDEALS are described by the following Dublin Core descriptors: title, author, contributor, subject, geographic coverage, issue date, type, language, description, uniform resource identifier (URI), publication status, peer reviewed, sponsor and date available (Figure 2). Information regarding best metadata practices are provided to researchers submitting data for deposit. However the amount of information used to describe the datasets is determined by the researcher.

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Description

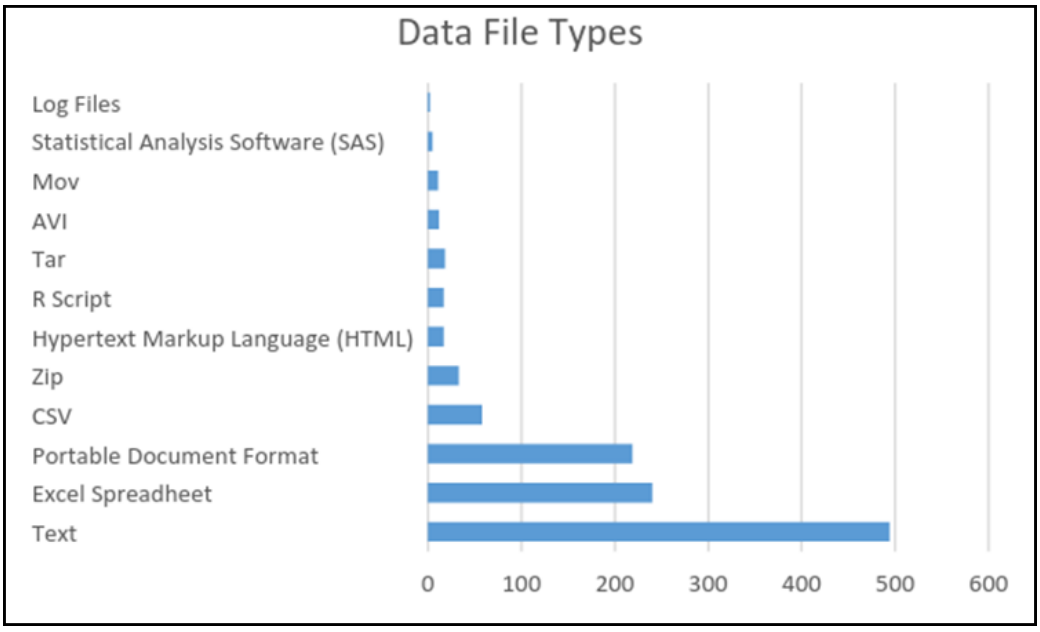
Title:	Film control data 20120601-07
Subject(s):	superfluid helium
Issue Date:	2013-11-06
Type:	Dataset / Spreadsheet
URI:	http://hdl.handle.net/2142/45966
Date Available in IDEALS:	2013-11-06

Figure 2: Metadata associated with datasets

Types of datasets in the repository

The second research question concerns which file types of datasets are represented. Table 1 lists the file type and number of these file types associated with the datasets in this study. This table indicates 494 files deposited are text files; 240 are *Microsoft Excel* files; and 219 PDFs are represented in the repository. Text files are the most frequently deposited file type within IDEALS. UIUC faculty and staff can deposit single and multiple files into the repository: 85 % of these file types correspond to one item being deposited, and 15% of the datasets contain multiple file deposits.

Table 1: Types of data files

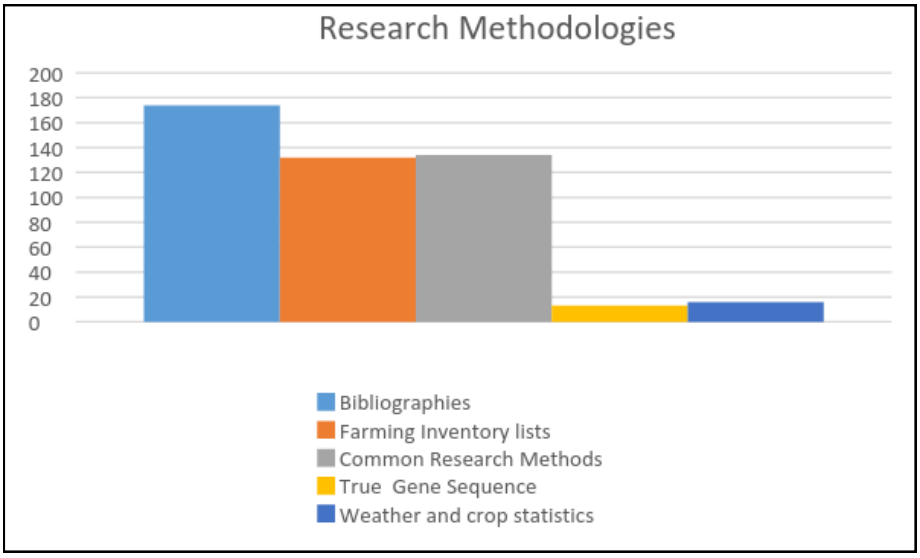


JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Research Methodologies associated with datasets

The third research question concerns the types of research methodologies associated with datasets. Table 2 indicates the types of research methodologies associated with datasets in IDEALS.

Table 2: Research Methodologies



Common research methodologies indicated in this table are the total number of datasets that referenced case studies, surveys, questionnaires, conceptual analysis, statistical analysis, framework models, technology reviews, experiments, data files, and databases within their description — 26% of these types of datasets are represented in the repository. However, the repository also includes datasets that are associated with bibliographies, farming inventory lists, weather and crop-related statistics, and true gene sequence. Bibliographies represent 38% of the datasets and farming inventory lists represent 29% of the datasets.

Peer Reviewed and funding agency related datasets

Four percent of the datasets were associated with peer-reviewed publications published within library, atmospheric, and biology disciplines. The farming inventory associated datasets are not peer reviewed but contributed to annual summary reports within the department of Crop Sciences and Illinois Department of Agriculture. Technical reports were published within the water, natural history, and engineering communities and disciplines are associated with 2% of the datasets in the repository. Seven percent of the datasets are associated with National Science Foundation (NSF), National Institute of Health (NIH) and Institute of Museum and Library Service (IMLS) grants.

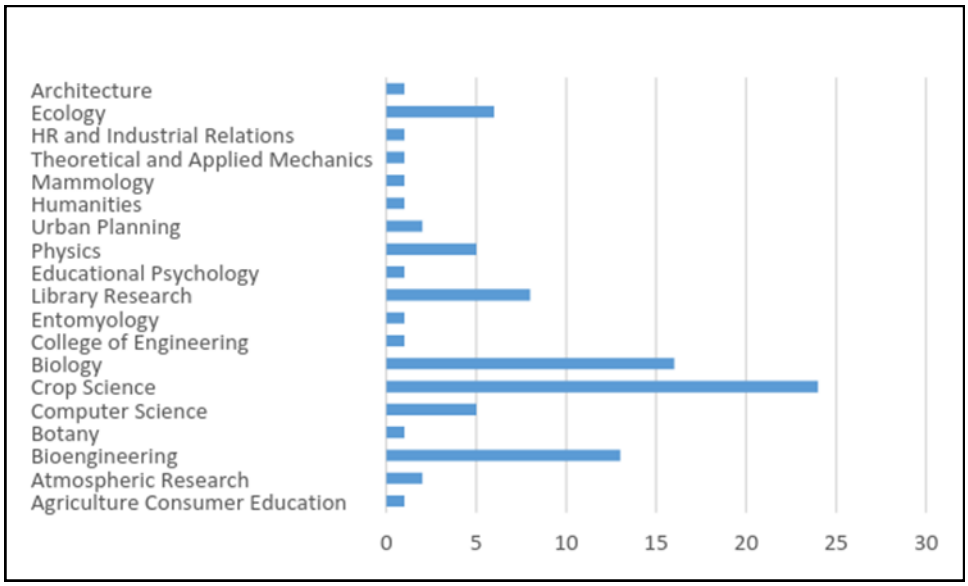
Research Discipline and Community

The fourth question concerns what research discipline or research communities are associated with the datasets.

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Table 3 lists a breakdown of individual research disciplines and the total number of datasets deposited. Biology, Bioengineering and Crop sciences are the disciplines with the highest number of datasets. Figure 3 compares the disciplines as a sum versus the research communities. Based on these results, 21% of the datasets within the repository are associated with research disciplines. They are listed as a total because their individual percentage is less than 1%. Rare Books represents 38 %, Illinois Department of Agriculture 35 % and Special Collection are 7%. 2.8% of the datasets represents disciplines or communities that are unknown. Research communities represent 78% of the datasets within the repository.

Table 3: Research Discipline associated Datasets



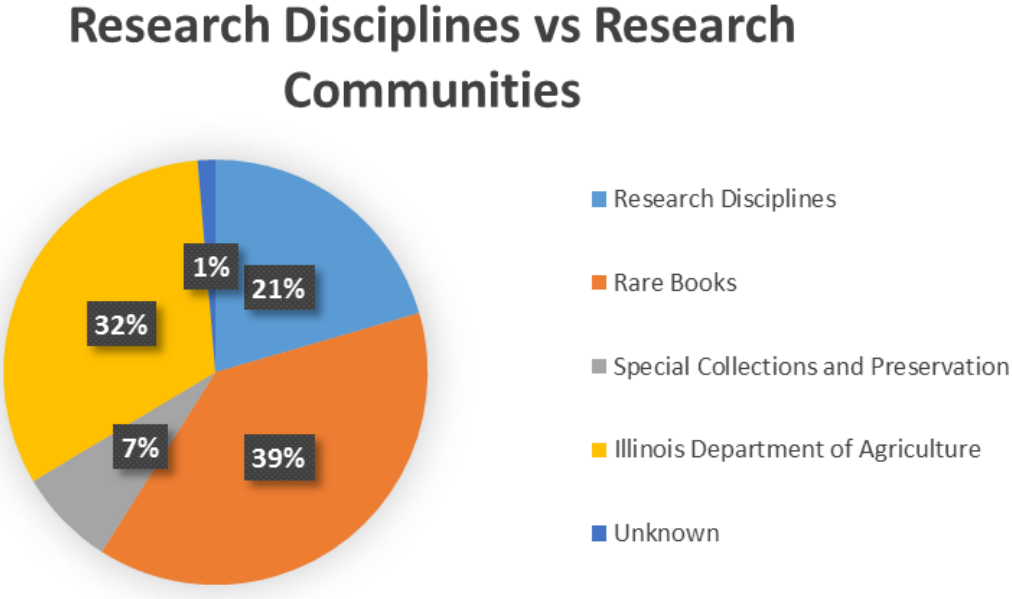


Figure 3: Research Disciplines vs Research Community

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Discussion

This study revealed special collection, rare books, and Illinois Department of Agriculture are the most frequent publishers of research data in the repository. This indicates the repository's greatest content is annual reports about farming, indexes of publications, and surveys. Research discipline representation indicates the relationship to funding agency requirements and data files associated with peer-reviewed research publications. The research discipline-related datasets indicates some faculty researchers are depositing data. Awareness efforts have been increased by subject librarians, research data interest groups, and committees on the University campus. The University of Illinois Research Data Service was created in 2013; future work could examine faculty awareness of the University having a repository to manage data.

Repository Issues

The result of using dataset provides 456 returns and the result of using data returns 265. Examining the 265 results indicates 204 of these are the same datasets listed within the search type dataset. Search results within Figure 4 are a comprehensive list of data and datasets of research disciplines and research communities. Whereas, search results within Figure 5 is comprised of data files for research article, supplementary data files, and farming inventory data.

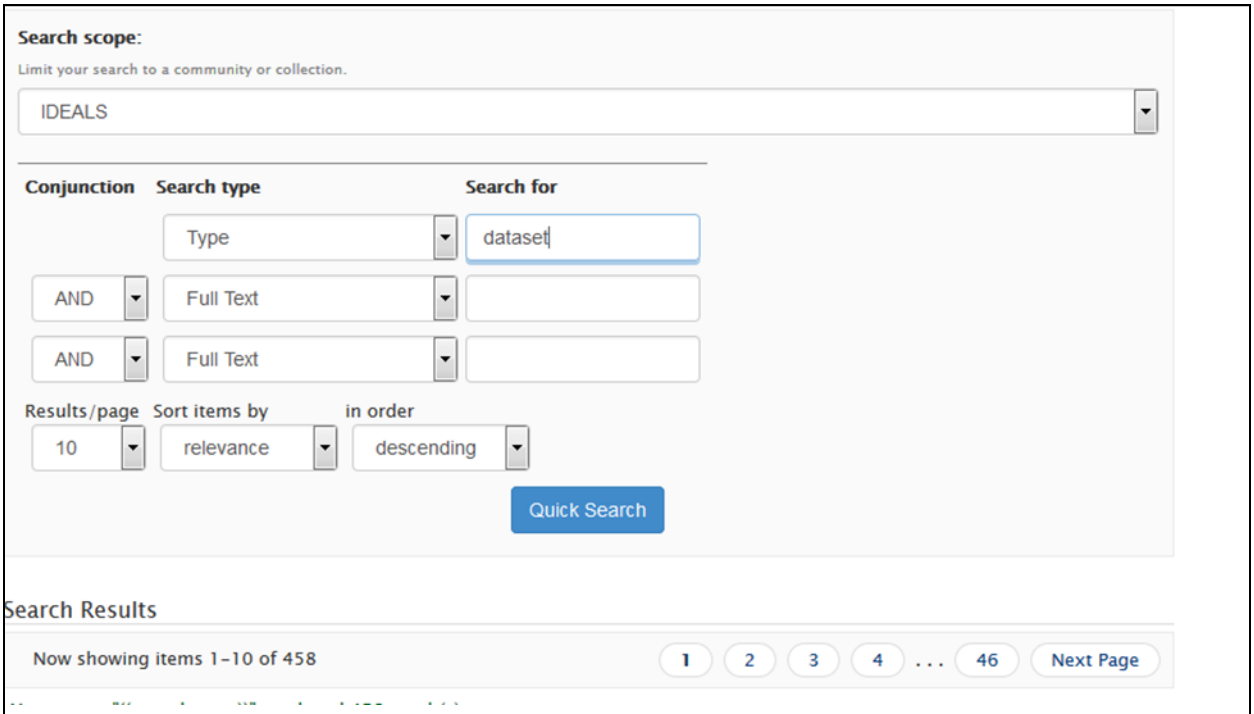


Figure 4: Search type: Dataset

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

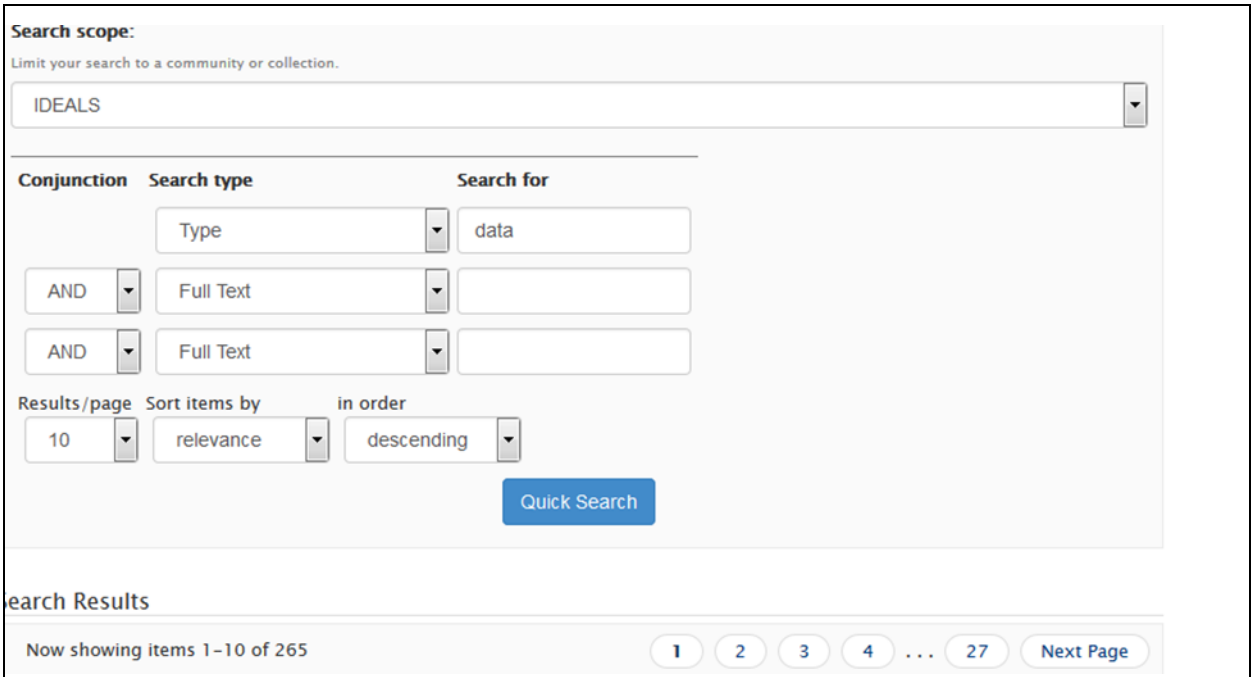


Figure 5: Search type: Data

These varied results indicate two problems defining data within the contents of a repository and the identifier issue within the campus repositories. The first is more difficult because data is not defined by a set definition. Borgman states data can take many forms in physical and digital content (Borgman 2012). Buckland states, defining data is a difficult task because it is an artifact, or observation at best of "alleged evidence" to use (Buckland 1991).

However, the identifier issue could be resolved if a uniform method of finding data and datasets within the repository was created. Other suggestions are examining the datasets to determine their classification. Bibliographies and attendance lists are valuable information but their classification as a dataset is uncertain. Current content policy ensures relevance and quality of material (Riddle 2015). In the context of datasets, relevance and quality are not enough. Campus repositories are seeing more datasets based on data management requirements (Ray 2014). But the implications of this when datasets are not related to data management requirements or non-peer reviewed publications is unknown. Best practices could include guidelines on classification, collection, and descriptions of datasets when it is not related to data management grants or peer reviewed research.

Another issue is in identifying a correlation — if any — between research methodology and research discipline. There is no direct correlation between research methodologies and research disciplines in this study. Survey research methods are associated with special collection and preservation datasets. None of these produced peer review or published work. Weather and crop statistics were among the Illinois Department of Agriculture and were annual summaries or internal: no published reports. Case studies, questionnaires, conceptual analysis, statistical analysis, framework models, technology reviews, experiment, and data files varied among the research disciplines. The College of Engineering and Theoretical Applied

JeSLIB 2015; 4(2): e1081
doi:10.7191/jeslib.2015.1081

Mechanics datasets contributed to technical reports. Future work could explore correlations between research disciplines, research methodology, and research datasets in the repository.

This study provided information about the number, file types, research disciplines, and research communities represented in the University repository. The context of this information can be applied to data management discussions with researchers. The results of this study indicate some researchers are using the campus repository to share data. This enhances data management conversations by allowing librarians to share the types of data and disciplines represented in the repository. This knowledge can lead to discussions with researchers on discerning researchers' current needs and determining whether the campus repository is a good fit or if a disciplinary repository would be a better option. The accessibility of the current file types illustrates the ability of the current structure to support and manage research data services. The research data service is launching a data repository in May 2016. The current structure meets short-term preservation needs, but it does not meet long-term preservation needs relating to active data or large datasets.

Limitations

There are limitations in the accuracy of quantifying the number of file types. Zip files are included in single item deposited because it only lists one item within the submission record. Yet a zip file consists of more than one file. The analysis of this occurrence is outside the scope of this study.

Another limitation is the accuracy of counting submissions records with multiple files deposited. For example, within the context of one record is a submission of 449 text files, five r-script files, and an html file; whereas another contributor deposited a single pdf file.

Conclusion

This type of case study provides a broad understanding of the representation of data within campus repositories. This knowledge allows subject specialists to have informed data management conversations with researchers and speak with greater knowledge about how the current infrastructure support researchers' data needs. The results of this study revealed that while the repository supports research communities, research disciplines, and features various types of data, there is significant room for improvement. In particular, establishing a policy for what research data gets collected, improving the issues with describing data, and creating a long-term preservation policy.

Understanding the file types, research methodologies, discipline, and communities are of significant value to academic librarians and repository managers. This understanding can strengthen data management consultations and needs assessments and will create collaborations with researchers. It allows subject specialists and data managers to quantify the number of research disciplines using the repository, make suggestions for changes in classification, collection, and description of datasets. Since the data management mandates were created, more opportunities exist for librarians to assist faculty beyond the data management plan. Some of these areas include creating and making tools available to share data, assisting with finding research data, and providing information on copyright and ownership issues associated with datasets. Librarians can also help with implementing

metadata standards for datasets. The information determined in the study can help librarians assist researchers with describing their contributed datasets so they are findable and usable.

Disclosure

The author reports no conflict of interest.

References

Alvaro, Elsa, Heather Brooks, Monica Ham, Stephanie Poegel, and Sarah Rosencrans. 2011. "E-Science Librarianship: Field Undefined." *Issues in Science and Technology Librarianship* 66:28-43. http://dx.doi.org/10.5062/F46Q1V55

Aschenbrenner, Andreas, Tobias Blanke, David Franders, Mark Hedges, and Ben O'Steen. 2008. "The Future of Repositories? Patterns for (Cross-)Repository Architectures." *D-Lib Magazine* 14. http://www.dlib.org/dlib/november08/aschenbrenner/11aschenbrenner.html

Baudoin, Patsy and Margaret Branschofsky. 2003. "Implementing an Institutional Repository: The DSpace Experience at MIT." *Science & Technology Libraries* 24:31-45. http://dx.doi.org/10.1300/J122v24n01_04

Christine L. Borgman. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/books/scholarship-digital-age

Borgman, Christine L. 2009. "The digital future is now: A call to action for the humanities." *Digital Humanities Quarterly* 3. http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Borgman, Christine L. 2012. "The conundrum of sharing research data." *Journal of the American Society for Information Science and Technology* 63:1059-1078. http://dx.doi.org/10.1002/asi.22634

Carlson, Jake, Alexis E. Ramsey, and J. David Kotterman. 2010. "Using an institutional repository to address local-scale needs: a case study at Purdue University." *Library Hi Tech* 28:152-173. http://dx.doi.org/10.1108/07378831011026751

Diekema, Anne R., Andrew Wesolek, and Cheryl D. Walter. 2014. "The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories." *Journal of Academic Librarianship* 40:322-331. http://dx.doi.org/10.1016/j.acalib.2014.04.010

Hey, Tony and Jessie Hey. 2006. "e-Science and its implications for the library community." *Library Hi Tech* 24:515-528. http://dx.doi.org/10.1108/07378830610715383

Henty, Margaret. 2008. "Dreaming of Data: the Library's Role in Supporting E-Research and Data Management." Paper presented at the *Australian Library and Information Association Biennial Conference*, Alice Springs, NT Australia, September 2–5. http://apsr.anu.edu.au/presentations/henty_alia_08.pdf

Luce, Richard E. 2012. "Grand challenges and new roles for the twenty-first-century research library in an era of e-science." In *The data deluge: Can libraries cope with e-science*, edited by D.B. Marcum and G. George. California: Libraries Unlimited.

Lynch, Clifford. 2003. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *portal: Libraries and the Academy* 3:327-336. http://dx.doi.org/10.1353/pla.2003.0039

Mullins, James L. 2009. "Bringing Librarianship to E-Science." *College & Research Libraries* 70:212-213. http://dx.doi.org/10.5860/crl.70.3.212

Mischo, William H., Mary C. Schlembach, and Megan N. O'Donnell. 2014. "An analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals." *Journal of eScience Librarianship* 3:e1060. http://dx.doi.org/10.7191/jeslib.2014.1060

Qin, Jian and John D'Ignazio. 2010. "The Central Role of Metadata in a Science Data Literacy Course." *Journal of Library Metadata* 10:188-204. http://dx.doi.org/10.1080/19386389.2010.506379

Ramírez, Marisa. 2011. "Whose Role Is It Anyway?: A Library Practitioner's Appraisal of the Digital Data Deluge." *Bulletin of the American Society for Information Science* 37:21-23 http://works.bepress.com/marisa_ramirez/20

Ray, Joyce M., editor. 2014. *Research Data Management: Practical Strategies for Informational Professionals* (Charleston Insights in Library, Archival, and Information Sciences). Indiana: Purdue University Press.

Riddle, Kelly. 2015. "Creating policies for library publishing in an institutional repository: Exploring purpose, scope, and the library's role." *OCLC Systems & Services: International digital library perspectives* 31:59-68. http://dx.doi.org/10.1108/OCLC-02-2014-0007

Soehner, Catherine, Catherine Steeves, and Jennifer Ward. 2010. "E-Science and Data Support Services: A Study of ARL Member Institutions." *Association of Research Libraries*. http://www.arl.org/storage/documents/publications/escience-report-2010.pdf

Steinhart, Gail, Eric Chen, Florio Arguillas, Dianne Dietrich, and Stefan Kramer. 2012. "Prepared to Plan? A Snapshot of Researcher Readiness to Address Data Management Planning Requirements." *Journal of eScience Librarianship* 1:e1008. http://dx.doi.org/10.7191/jeslib.2012.1008

Tarver, Hannah and Mark Phillips. 2013. "Integrating Image-Based Research Datasets into an Existing Digital Repository Infrastructure." *Cataloging & Classification Quarterly* 51:238-250. http://dx.doi.org/10.1080/01639374.2012.732203