01 Feb 2015

Data Management and Curation Practices: The Case of using DSpace and Implications

Yin Zhang

Hsin-Liang Chen
Missouri University of Science and Technology, chenhs@mst.edu

Recommended Citation

Zhang, Y., & Chen, H. (2015). Data Management and Curation Practices: The Case of using DSpace and Implications. *Proceedings of the 78th Annual Meeting of the Association for Information Science and Technology (2015, St. Louis, MO), 52*(1) John Wiley & Sons Inc..
The definitive version is available at https://doi.org/10.1002/pra2.2015.1450520100109

Data Management and Curation Practices: The Case of Using DSpace and Implications

Yin Zhang
School of Library and Information Science
Kent State University
yzhang4@kent.edu

Hsin-liang Chen
Palmer School of Library and Information Science
Long Island University
Hsin.Chen@liu.edu

ABSTRACT
Data management and curation is a new challenge with the emerging trend of data-dependent scholarly research. Due to the lack of common standards and best practices, current data management and curation practices have been varied. This poster presents a project that examines the common practices of data management and curation that helps understand the scope of and factors behind such variations. The sample of this study consists of 171 unique data repositories created by 164 institutions from 95 countries worldwide. The preliminary results indicate that data management and curation is a global issue. Currently, academic institutions and government agencies are the leading force in contributing and sharing data. Data repositories are used for various purposes with international repository and learning resources being the most common use cases. Additionally, system functions used to manage data repositories vary to a great extent with statistics and OAI harvesting being the most common ones.

Keywords
Data management, data curation, institutional repository, DSpace.

INTRODUCTION

Scholarly research has seen a new paradigm characterized by the massive scale of data creation and accumulation, as well as scientific discovery based on intensive data (Hey, Tansley, & Tolle, 2009; Jahnke, Asher, Keralis, 2012). Major funding agencies such as National Science Foundation (NSF) have imposed requirements for data sharing and management plan for funded projects. As a response to the challenge for research and scholarship, more institutions and libraries have started implementing data management and curation programs. However, as noted in the Council on Library and Information Resources report *The Problem of Data*, there has been a lack of common standards and best practices to meet the challenge of data management and curation (Jahnke, Asher, Keralis, 2012).

Current data management and curation practices have been varied. There is a need for understanding the scope of and factors behind the variations in practice. The aim of this research is to examine the common practices of data management and curation using DSpace. DSpace was chosen for several considerations: 1) it has the largest digital repository user community and developers worldwide; 2) it is free open source software; 3) it was initially developed by and for academic institutions and now is most commonly used by research libraries to manage digital contents; and 4) it is completely customizable to meet needs of individual institutions and repositories (DSpace, 2015).

LITERATURE REVIEW

Related Background

Joyce (2012) described the development of "digital curation" and "cyberinfrastructure" since the end of the 20th century and discussed how government agencies and research communities have embraced the concepts with funding and research activities. Some of those key achievements are the required data management plans by NSF and the Institute of Museum and Library Services (IMLS); IMLS' *A Framework of Guidance for Building Good Digital Collections*; interoperable standards (e.g., Open Archive Initiative Protocol for Metadata Harvesting, OAI-PMH); and institutional repository systems (e.g., Purdue University Research Repository, PURR).

Major funding agencies now require data sharing and management plan for funded projects. For example, beginning January 18, 2011, National Science Foundation (NSF) required all grant proposals to include a two-page "Data Management Plan." Other U.S. federal funding agencies such as National Institutes of Health (NIH) and National Aeronautics and Space Administration (NASA) also implemented similar requirements.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

Standard Development

Academic libraries and librarians have been identified as curatorial liaisons on campus in the data curation movement due to their long-standing history, credentials and commitments (Fox, 2013; Heidorn, 2011; Lyon, 2012; Schubert, Shorish, Frankel, & Giles, 2013). As a result, several metadata standards for data management and curation have been developed to manage massive large-scale data sets (Ogier, Hall, Bailey, & Stovall, 2014; Weber, Palmer, & Chao, 2012).

Weber, Palmer and Chao (2012) emphasized the importance of discipline-specific data practice and data privacy and ownership policies in developing interoperable standards. Lyon (2012) proposed a research data management (RDM) model in the UK environment. Ogier, Hall, Bailey and Stovall (2014) applied the Data Asset Framework (DAF) methodology to audit and evaluate the electronic resources data at the Virginia Tech Libraries. Currently many data curation standards are still under development.

System Implementation

Initially, many academic libraries used institutional repository (IR) systems as their research data management systems. MIT's DSpace is a popular IR system adopted by global institutions. Tansley et al. (2003) summarized DSpace's initial functions as a data model, metadata, e-people, authorization, ingesting, workflow, CNRI Handle system, search and browsing, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), subscription, and Web user interface. Baudoin and Branschofsky (2004) noted that implementation of DSpace changed how MIT researchers think about the lifecycle of scholarly research and the operating definitions of units of the scholarly enterprise. Additionally, DSpace is increasingly seen as an active player in developing technical infrastructure at MIT.

Higher education institutions, research centers, and government agencies have adopted DSpace. From its initial success, DSpace has grown into a worldwide community. For example, Chen and Hsiang (2009) used DSpace to implement The National Taiwan University Repository (NTUR) with several modifications of its functional modules to fulfill the requirements of Chinese users. The content acquisition of NTUR was carried out by a machine-aided manual approach, which quickly accumulates the volume of registered digital objects in NTUR.

With the active and continued contributions from its user communities, DSpace has been expanded with growing functions. For example, the Texas Digital Library team introduced Manakin for specialized user interfaces (Philips, Green, Maslov, Mikeal, & Leggett, 2007), added a customized workflow management system and Open Archives Initiative Object Reuse and Exchange (OAI-ORE) (Mikeal et al., 2009; Maslov et al., 2010; Lagoze et al. (2012); and created a Web 2.0-based interface for a map collection (Maslov, Mikeal, Weimer, & Leggett, 2009) to the DSpace system. Semantics is another emerging development area in DSpace functions that aims to facilitate more efficient search processes among DSpace members and their collections (Kruk & McDaniel, 2009; Usman & Khan, 2012; Cherukodan, Kumar, & Kabir, 2013). Additionally, Cherukodan, Kumar, and Kabir (2013) applied Google Analytics to evaluate the distribution of the digital items and usage of an academic DL implemented by DSpace.

Some research extensive universities opt to develop their home-grown repository systems. Rolando, Doty, Hagenmaier, Valk, and Parham (2013) presented an internal study on research data assessment at Georgia Institute of Technology and recommended to develop a research data repository to support data management. On the other side, Purdue University developed its own research repository, PURR, for its faculty, students and staff (Matthews & Witt, 2013). Purdue researchers use PURR, a web-based platform powered by HUBzero, to share data and collaborate on research online.

Research Gap

In recent years, there have been active exploration and development on the design and implementation of data management repository systems to meet the needs of data intense scientific research and discovery. Because of the evolving nature of this new trend, related standards and practices are still being developed. There is a need to survey and understand current data management and curation practices. This research was designed to fill this gap.

RESEARCH QUESTIONS

- RQ1: What types of institutions currently have data repositories?
- RQ2: What do institutions use their data repositories for in terms of use cases?
- RQ3: What are the most adopted system functions that are commonly used for data repositories?

METHODOLOGY

The data about DSpace data repositories were collected from the DSpace User Registry (http://registry.duraspace.org/registry/dspace) during November 2014 to January 2015. A total of 205 repositories in the registry designated "data sets" as their file type or content type, among which 34 designated "data sets" as both their file type and content type while 171 only designated "data sets" as their file type. This sample represents 171 unique repositories created by 164 institutions from 95 countries worldwide, with Turkey, United States, India, and United Kingdom being the top

four countries with at least 10 data repositories. This sample also reflects the global nature of data management and curation issues. Table 1 summarizes the most representative countries in the sample.

Country	# of Data sets repositories	Percent
1. Turkey	18	11%
1. United States	18	11%
3. India	16	10%
4. United Kingdom	10	6%
5. Brazil	6	4%
5. Taiwan	6	4%
7. Canada	5	3%
7. Colombia	5	3%
7. Sri Lanka	5	3%
7. Vietnam	5	3%
11. Germany	4	2%
11. Greece	4	2%
11. Kenya	4	2%
14. France	3	2%
14. Indonesia	3	2%
14. Mexico	3	2%
14. Spain	3	2%
14. Ukraine	3	2%

Table 1. Most representative countries with data sets repositories in DSpace (N=171).

The following data elements were collected for each of the repositories and the data were processed using Microsoft Access and SPSS for analysis:

- institution affiliation,
- institution type,
- country,
- use case type(s),
- content type(s) in the repository,
- file type(s) in the repository, and,
- system implementation integrations/customizations.

FINDINGS

RQ1: What types of institutions currently have data repositories?

As shown in Table 2, 70% of the data set repositories are affiliated with academic institutions, about 10% affiliated with government, and 6% with nonprofit organizations. This result echoes the research needs and funding requirements for data management and curation for academic communities.

RQ2: What do institutions use their data repositories for in terms of use cases?

A repository may be used for multiple purposes as indicated in its use case in the registry. In this sample, a total of 354 use case instances are reported for the 171 data repositories. Data repositories are most commonly used as institutional repositories (69% of the data repositories), learning resources (33%), subject repositories (25%), and image repositories (22%).

Type of Institution	Repository count	Percent
Academic	119	70%
Government	17	10%
Nonprofit	10	6%
Personal	6	4%
Research Center	6	4%
Commercial	5	3%
Archive / Public Library	3	2%
Consortium	2	1%
Medical Center / Hospital	2	1%
Other	1	1%
Total	171	100%

Table 2. Data repositories and institution types (N=171).

Use Case	Repository count	% used
Institutional Repository	118	69%
Learning Resources	56	33%
Subject Repository	42	25%
Image Repository	38	22%
Audio/Video Repository	31	18%
Government Records/Reports	23	13%
Museum/Cultural Heritage	21	12%
Federated Repositories/Networked Instances	14	8%
Other	11	6%

Table 3. Data repositories and use cases (N=171).

RQ3: What are the most adopted system functions that are commonly used for data repositories?

Among the 32 unique DSpace system functions available, the most commonly used ones for data repositories are summarized in Table 4. The most used function is statistics (35%), which tracks repository usage and repository visits. The next most commonly used function is OAI Harvester Plugin (23%) that facilitates data sets harvesting and sharing across systems. The next two (tied for the third) most popular functions are Google Analytics Tracking Code (19%) and Manakin Themes (19%). It is interesting to note that although Google Analytics Tracking Code is a relatively new feature compared to other ones, it gains popularity for data repositories for evaluating the distribution of the digital items and usage. The other popular functions that make the top ten are adopted by at least 10% of the repositories in registry include Dublin Core Meta Toolkit, Language Packs, Google Indexing, Creative Commons Open URL, Websites, and Embargo.

Integration and Customization	Count	% used
1. Statistics	60	35%
2. OAI Harvester Plugin for Dspace	40	23%
3. Google Analytics Tracking Code	33	19%
3. Manakin Themes	33	19%
5. Dublin Core Meta Toolkit	31	18%
5. Language Packs	31	18%
7. Google Indexing of DSpace Instances	26	15%
8. Creative Commons Open URL	20	12%
9. Websites	18	11%
10. Embargo	17	10%

Table 4. Data repositories and most commonly used system functions used (N=171).

DISCUSSION AND CONCLUSIONS

The preliminary results of this study show that data management and curation is an issue shared globally. Previous research revealed that about 21% of the digital repositories are for data sets (Chen & Zhang, 2014), this study shows that academic institutions and government agencies are taking a lead in making their data repositories available. Due to lack of standards, current practices of data management and curation vary significantly by institutions, use cases, and system functions in implementations. Further data analysis is underway to examine the factors behind such variations. The results will help institutions make informed decisions as they create their data repositories based on their institutional needs while learning from their peers. The results will also facilitate the development of related standards and best practices in the context of institutional needs, purpose of data repositories, and system functions.

REFERENCES

Baudoin, P., & Branschofsky, M. (2004). Implementing an institutional repository: the DSpace experience at MIT. *Science & Technology Libraries*, *24*(1-2), 31-45.

Chen, K. H., & Hsiang, J. (2009). The unique approach to institutional repository: Practice of National Taiwan University. *The Electronic Library*, *27*(2), 204-221.

Chen, S.L., & Zhang, Y. (2014). Functionality analysis of an open source repository system: current practices and implications. *The Journal of Academic Librarianship*, *40*, 558-564.

Fox, R. (2013). The art and science of data curation. *OCLC Systems & Services*, *29*(4), 195-199.

Heidorn, P. (2011). The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration*, *51*(7/8), 662-672. doi:10.1080/01930826.2011.601269

Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond,WA: Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

Jahnke, L., Asher, A., & Keralis, S.D.C. (2012). The problem of data. Washington, DC: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub154/pub154.pdf

Joyce, R. (2012). The rise of digital curation and cyberinfrastructure: From experimentation to implementation and maybe integration. *Library Hi Tech*, *30*(4), 604-622.

Lyon, L. (2012). The informatics transform: Re-engineering libraries for the data decade. *International Journal of Digital Curation*, *7*(1), 126-138.

Mallon, M. (2012). Data curation. *Public Services Quarterly*, *8*(4), 326-337. doi:10.1080/15228959.2012.730400

Matthews, C. E., & Witt, M. (2013). The Purdue University Research Repository (PURR): Providing institutional data services with a virtual research environment, data publication, and archiving. *Open Repositories*. Charlottetown, Prince Edward Island. Retrieved January 12, 2015 from http://works.bepress.com/courtney_earl_matthews/3

Ogier, A., Hall, M., Bailey, A., & Stovall, C. (2014). Data management inside the library: Assessing electronic resources data using the data asset framework methodology. *Journal of Electronic Resources Librarianship*, *26*(2), 101-113. doi:10.1080/1941126X.2014.910406

Rolando, L., Doty, C., Hagenmaier, W., Valk, A., & Parham, S. W. (2013). Institutional readiness for data stewardship: Findings and recommendations from the research data assessment. Retrieved March 11, 2015 from https://smartech.gatech.edu/bitstream/handle/1853/48188/Research%20Data%20Assessment%20Final%20Report.pdf

Schubert, C., Shorish, Y., Frankel, P., & Giles K. (2013). The evolution of research data: Strategies for curation and data management. *Library Hi Tech News*, *30*(6), 1-6.

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., & Smith, M. (2003, May). The DSpace institutional digital repository system: current functionality. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries* (pp. 87-97). IEEE Computer Society.

Weber, N. M., Palmer, C. L., & Chao, T. C. (2012). Current trends and future directions in data curation research and education. *Journal of Web Librarianship*, *6*(4), 305-320.