2012

The Digital Repository Landscape: Developing a Research Guide to Selected Digital Repositories

Zheng (Jessica) Lu
University of San Francisco, zjlu@usfca.edu

Recommended Citation

Lu, Zheng (Jessica), "The Digital Repository Landscape: Developing a Research Guide to Selected Digital Repositories" (2012). *Gleeson Library Librarians Research.* Paper 1.
http://repository.usfca.edu/librarian/1

The Digital Repository Landscape: Developing a Research Guide to Selected Digital Repositories

Professional Development Leave Report

by

Jessica Lu, August 20, 2012

Overview

I applied for my professional development leave to research the digital repository landscape with the following two goals in mind:

1. Survey and develop a research guide to digital repositories with open content that can later be used for reference or incorporated into subject guides.
2. Use my findings to inform a long-term strategy and plan for the Gleeson Library's repository development.

As a result, I'm attaching to this report (see Appendix) a research guide to selected digital repositories that could serve as general reference resources or be worked into subject resources in research guides. In this report, I will detail the research methodology I used to compile this guide, as well as other general findings from my survey of digital library literature and browsing of various repository instances. Finally, I would summarize how my research has informed me about the long-term strategy and plan for the Gleeson Library's digital repository development.

Methodology

Literature on digital library and digital repositories has been flourishing in recent years. So for my research purposes, other than identifying literature addressing general trend in repository development, I concentrate on one relatively new topic: subject repositories.

The starting points of my planned survey of digital repositories are two registries frequently referenced in digital library discussions and literature: Registry of Open Access Repositories (ROAR) http://roar.eprints.org/ and the Directory of Open Access Repositories (OpenDOAR) http://www.opendoar.org/.

Upon close study of both websites and the various ways to browse both registries, I decided to concentrate on ROAR as my primary starting point based on the following two reasons.

1. Comprehensiveness of the registered repositories.

While the majority of listings on both registries may overlap, by just looking at statistics, ROAR's 2925 listings is 33% more than OpenDOAR's 2193 records (both number retrieved from respective websites on August 14, 2012). One important reason for this difference is probably the different treatment of what is defined as an open access repositories. According to OpenDOAR: "*Open*DOAR has opted to collect and provide information solely on sites that wholly embrace the concept of open access to full text resources that are of use to academic researchers. Thus sites where any form of access control prevents immediate access are not included: likewise sites that consist of metadata records only are also declined." In comparison, ROAR does not clearly state what kind of open access repositories are listed and in practice often includes data repositories, or repositories with metadata only records that are excluded by OpenDOAR. Since my project goal is to survey the general digital repository landscape with an emphasis on compiling a list of useful sites for references, I decide to use ROAR as my starting point to get a more complete view of the field.

2. Browsing options

Both sites offer a very thorough search function to identify specific repository, but differs significantly in the browsing options. OpenDOAR only list repositories by countries and organizations. ROAR on the other hand offers four browsing options: by country, by year, by repository type and by repository software. These categorizations not only provide multiple access points to the actual listings but also offer a good overview of the general repository landscape as I'll summarize in my findings.

To identify subject repositories with significant holdings among the thousands of registry listings, I used the following approach:
First, sort all repository listings by total records held in the repository and review the first 100 results over the span of a week. While the rank of the identified repositories varied day to day, they consistently come up among the top 20 repositories with most records.

Next, select all US repositories (results are grouped by repository type) and review "Research Multi-institution Repository" and "Research Cross Institutional" as these two are most likely to be subject specific repositories. I then cursively browse the repository names under the "all Research Institutional or Departmental" type to see if it looks like a subject repository. In some cases, the nature of the institution or its specialty makes its institutional repository a subject specific repository. These types of subject repositories may be smaller in scale than a cross institutional collaboration, but often offers unique and important resources. Therefore I also include such instances in the research guide.

The repositories I identify as appropriate for the attached research guide through this approach is by no means comprehensive, as the registry itself relies on self-registering from repositories and my survey method largely overlook foreign repositories. Still, I believe all the major subject repositories with wide influence in their respective disciplines are covered, as the list corresponds well with the named subject repository in digital library literature.

There's no literature summarizing format/type specific repository in general, as each individual category (newspaper or ETD in this case) is a topic big enough to warrant individual research on the topic. I put them under the heading of format specific repository because in both digitized newspapers and ETDs, the content could be all-encompassing and diverse, but the format/type of the content is what defines these repositories. Just like traditional library not only sort its collections by subject, but also by format/type such as periodicals. It is worth monitoring if future digital repository development sees this type of mimicking of a physical library.

Findings

Literature findings:

- Subject librarian not yet to view IR as valuable information resources, though IR is good for finding theses. (Dorner)
- Faculty familiar with subject repository not necessarily more apt to archive in IR (Xia)
- "The field of economics has characteristics that contribute to the success of a subject repository, such as a pre-print culture and an interest in intellectual property and the economics of publishing." (Kelly J)
- "Subject repositories are frequently cited as highly successful scholarly communication initiatives, especially in relation to institutional repositories. The lack of subject repository recognition within the literature and among commonly used repository tools may be attributed to the isolated development of the largest subject repositories and a general lack of awareness about small-scale subject repositories." (Adamick)
- Collaboration between IR (example Economists online subject repository) (Puplett)

ROAR findings

Although repositories are found all over the world, U.S. leads the field with a dominating 417 registered repositories and several European nations come next. Outside North America and Europe only Japan and Brazil have over 100 registered repositories. On the one hand, this indicates the very unbalanced production of digital information among the continents; on the other hand it also signifies great potential for repository development or collaboration in those less developed regions.

Among the top 100 repositories (by record numbers), subject repositories with multi-institutional participation ranks highest, most of them in top 10-20 range, with the rest of the 100 mostly institutional repositories. Majority are university repositories hosted by libraries, some are hosted by research organizations or government. This confirms libraries as the central force in pushing the open access movement and librarians as the logical choice as repository custodian.

The top 3 repository software by popularity are D-Space (1176 implementations), E-Prints (458 implementations) and BePress (137 implementations). The former two are both open source repository software that have been in play in the field for over 10 years (Dspace since 2002 and E-Prints since 1999). BePress's repository software Digital

Commons is a hosted system and relative new comer, launch to the wider customer base only in 2007. Its fast take-up indicates a flourishing of institutional repositories among smaller institutions and organizations that do not want to take on the burden of up keeping an open source system.

When reviewing repository data by year, it's easy to see the number of new repositories launched entered hundreds starting in 2004, increased every year until peaked in 2010 and then decreased by about 100 annually in 2011 and 2012. Couple this data with the fact that many new repositories since 2007 are on hosted Digital Commons platform, the conclusion is repository development is probably entering a mature stage, and many newer and smaller repositories are probably not registered in ROAR.

Historic newspapers digitization findings

While many operating commercial newspapers have digitized archival collections available through major news databases, digitization of smaller or discontinued historic newspaper often relies on libraries and archives that still hold physical copies. Therefore the digitization project and the resulting digital collections are often scattered among different institutional digital repositories. However, spearheaded by state universities, state library or archives, several states have made a central repository of digitized regional historic newspapers available to the public online. Therefore I have included a number of large statewide digital historic newspaper collections in the research guide.

Implication for Gleeson Library Digital Collections development

1. The above findings confirm that we are on the right path in establishing an open-access institutional repository to host faculty and student scholarly works and the choice of Digital Commons as our repository software.
2. We need to register our IR and digital collections on both registries, not only to publicize the existence and content of our collections, but also to become an active member of the repository community.
3. Other than maintaining an institutional repository, we also need to actively seek opportunity for collaboration in subject repository, esp. in humanities. CRRA (Catholic Research Resources Alliance) is a perfect venue for a start.

References

Adamick, J., & Reznik-Zellen, R. (2010). *Trends in large-scale subject repositories D-Lib Magazine, 16*(11/12) doi: 10.1045/november2010-adamick

Adamick, J., & Reznik-Zellen, R. (2010). Representation and recognition of subject repositories. *D-Lib Magazine, 16*(9), 1-10. doi:10.1045/september2010-adamick

Armbruster, C., & Romary, L. (2010). Comparing repository types - challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. Retrieved from http://0-search.ebscohost.com.ignacio.usfca.edu/login.aspx?direct=true&db=edsarx&AN=1005.0839&site=eds-live&scope=site; http://arxiv.org/abs/1005.0839

Dorner, D. G., & Revell, J. (2012). Subject librarians' perceptions of institutional repositories as an information resource. *Online Information Review, 36*(2), 261-277. doi:10.1108/14684521211229066

Kelly, J., & Letnes, L. (2010). AgEcon search: A case study on the differences between operating a subject repository and an institutional repository. *JODI: Journal of Digital Information, 11*(1), 9p. Retrieved from http://0-search.ebscohost.com.ignacio.usfca.edu/login.aspx?direct=true&db=rzh&AN=2010641374&site=eds-live&scope=site

Puplett, D. (2010). The economists online subject repository-using institutional repositories as the foundation for international open access growth. *New Review of Academic Librarianship, 16*, 65-76. doi:10.1080/13614533.2010.509490

Xia, J. (2008). A comparison of subject and institutional repositories in self-archiving practices. *The Journal of Academic Librarianship, 34*(6), 489-495. doi:10.1016/j.acalib.2008.09.016

Appendix

Research Guide to Selected Digital Repositories

Subject Repository

PubMed Central
http://www.ncbi.nlm.nih.gov/pmc/
PMC is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM)

CiteSeerx
http://citeseerx.ist.psu.edu/index
CiteSeerx is an evolving scientific literature digital library and search engine that focuses primarily on the literature in computer and information science. CiteSeerx aims to improve the dissemination of scientific literature and to provide improvements in functionality, usability, availability, cost, comprehensiveness, efficiency, and timeliness in the access of scientific and scholarly knowledge.

Humanities Text Initiative
http://www.hti.umich.edu/
HTI includes individual text collections & **Making of America**.
Making of America (MoA) is a digital library of primary sources in American social history from the antebellum period through reconstruction. The collection is particularly strong in the subject areas of education, psychology, American history, sociology, religion, and science and technology. The collection currently contains approximately 10,000 books and 50,000 journal articles with 19th century imprints.

RePEc (Research Papers in Economics)
http://www.repec.org/
RePEc is a collaborative effort of hundreds of volunteers in 75 countries to enhance the dissemination of research in Economics and related sciences. The heart of the project is a decentralized bibliographic database of working papers, journal articles, books, books chapters and software components, all maintained by volunteers. The collected data is then used in various services such as:

- **EconPapers** http://econpapers.repec.org/
  EconPapers use the RePEc bibliographic and author data, providing access to the largest collection of online Economics working papers and journal articles. The majority of the full text files are freely available, but some (typically journal articles) require that you or your organization subscribe to the service providing the full text file.
- **IDEAS** http://ideas.repec.org/
  The largest bibliographic database dedicated to Economics and available freely on the Internet. Over 1,200,000 items of research can be browsed or searched, and over 1,100,000 can be downloaded in full text! This site is part of a large volunteer effort

to enhance the free dissemination of research in Economics, RePEc, which includes bibliographic metadata from over 1,400 participating archives, including all the major publishers and research outlets.

UKPMC – UK PubMed Central
http://ukpmc.ac.uk/
UK PubMed Central (UKPMC) offers free access to biomedical literature resources including:

- PubMed abstracts (about 22 million)
- UKPMC full text articles (about 2.2 million, of which over 400,000 are Open Access)
- Patent abstracts (over 4 million European, US, and International)
- National Health Service (NHS) clinical guidelines
- Agricola records (500,000)
- Supplemented with Chinese Biological Abstracts and the Citeseer database.

arXiv e-Print archive
http://arxiv.org/
Open access to 623,317 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

PANGAEA
http://www.pangaea.de/
Data Publisher for Earth and environmental science. The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. PANGAEA is a designated archive for the journal Earth System Science Data (ESSD)

NASA Technical Reports Server
http://ntrs.nasa.gov/search.jsp
The NTRS is a valuable resource for researchers, students, educators, and the public to access NASA's current and historical technical literature and engineering results. Over 500,000 aerospace-related citations, over 200,000 full-text online documents, and over 500,000 images and videos are available.

BioMed Central
http://www.biomedcentral.com/
BioMed Central is an STM (Science, Technology and Medicine) publisher of 220 open access, online, peer-reviewed journals. The portfolio of journals spans all areas of biology and medicine and includes broad interest titles, such as *BMC Biology* and *BMC Medicine* alongside specialist journals, such as *Retrovirology* and *BMC Genomics*. All original research articles published by BioMed Central are made freely and permanently accessible online immediately upon publication.

FAOBIB

http://www4.fao.org/faobib/
FAOBIB is a multilingual, on-line catalogue of documents and publications produced by FAO (*Food and Agriculture Organization of the United Nations)* since 1945, books added to the library collections since 1976, and serials held in the FAO library.

Biodiversity Heritage Library
http://www.biodiversitylibrary.org/
Biodiversity Heritage Library, a consortium of natural history and botanical libraries that cooperate to digitize and make accessible the legacy literature of biodiversity held in their collections and to make that literature available for open access and responsible use as a part of a global "biodiversity commons." BHL also serves as the foundational literature component of the Encyclopedia of Life (EOL).

National Agricultural Library Digital Collections
http://naldc.nal.usda.gov/
The National Agricultural Library Digital Collections (NALDC) includes historical publications, U. S. Department of Agriculture (USDA) research, and more.
The scope of the NALDC includes the following:

- Items published by the United States Department of Agriculture (USDA) and clearly intended for public consumption;
- Scholarly and peer-reviewed research outcomes authored by USDA employees while working for USDA; and
- Other items selected in accordance with the subjects identified in the NAL Collection Development Policy.

Bepress Legal Repository
http://law.bepress.com/
The bepress Legal Repository offers working papers and pre-prints from scholars and professionals at top law schools around the world. 126039 papers to date (August 14, 2012)

Comprehensive Digital Repository

Hathitrust Digital Library
http://www.hathitrust.org/
HathiTrust Digital Library is a digital preservation repository and highly functional access platform. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives.

Internet Archive
http://archive.org/
The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library. Now the Internet Archive includes texts, audio, moving images, and software as

well as archived web pages in our collections, and provides specialized services for adaptive reading and information access for the blind and other persons with disabilities.

Other Noteworthy Digital Collections

Digitized Full-text Historical Newspaper

***Chronicling America:* Historic American Newspapers**
http://chroniclingamerica.loc.gov/

California Digital Newspaper Collection
http://cdnc.ucr.edu/about_us.html

Colorado Historic Newspapers Collection
http://www.coloradohistoricnewspapers.org

Historic Kentucky Newspapers (Part of the Kentuckiana Digital Library)
http://kdl.kyvl.org/cgi/t/text/text-idx?xg=0;page=simple;g=news

Louisiana Newspaper Access Program
http://louisdl.louislibraries.org/cdm4/browse.php?CISOROOT=/LSU_LNP

Missouri Digital Heritage
http://www.sos.mo.gov/mdh/browse.asp?id=12.6

North Carolina Newspaper Digitization Project
http://www.archives.ncdcr.gov/newspaper/search.html

Historic Oregon Newspapers
http://oregonnews.uoregon.edu/

Utah Digital Newspaper
http://digitalnewspapers.org/

Washington Historic Newspaper
http://www.sos.wa.gov/history/newspapers.aspx

Theses and Dissertations

Networked Digital Library of Theses and Dissertations Union Catalog
http://www.ndltd.org/find
The NDLTD Union Catalog contains more than one million records of electronic theses and dissertations. For students and researchers, the Union Catalog makes individual collections of NDLTD member institutions and consortia appear as one seamless digital library of ETDs.

Other useful ETD search tools (copied from http://www.ndltd.org/find)
ADT (Australiasian Digital Theses Program) This search portal provides searching, browsing, and access to ETDs produced in Australia.

Biblioteca Digital de Teses e Dissertacoes A search tool for accessing ETDs produced in Brazilian universities.

Cybertesis A portal developed jointly by the University of Chile, the Universites de Lyon, Montreal, and Alexandrie, and the University of Geneva for accessing full text ETDs from many countries, including Bolivia, Brazil, Canada, Chile, France, Hong Kong, Mexico, Peru, Spain, and the United States.

DART-Europe E-theses Portal A discovery service for open access research theses awarded by European universities.

Deusche National Bibliothek Dissertations since 1998 are available via search in the German National Library.

DiVA This portal provides access to ETDs and research publications written at 26 institutions in Scandinavia.

EThOS Electronic Theses Online Service (EThOS) offers free access, in a secure format, to the full text of electronically stored UK theses--a rich and vast body of knowledge.

NARCIS This search portal provides access to ETDs produced in the Netherlands, as well as access to a variety of other research and data sets.

National ETD Portal (South Africa) This search portal provides access to ETDs produced in South Africa.

ProQuest Dissertations and Theses Many university libraries provide password access to this commercial database, but the link above also provides access for individuals without a connection to a research library. The collection includes most recent North American ETDs and selective coverage for other regions of the world.

RCAAP - Repositório Científico de Acesso Aberto de Portugal
The RCAAP's mission is to promote, support and facilitate the adoption of the open access movement in Portugal. RCAAP The project aims to: increase the visibility, accessibility and dissemination of academic activity and Portuguese scientific research, facilitating the management and access to information about scientific production and integrate Portugal into a set of international initiatives. This portal offers a union catalog with digital contents from more than 30 institutions.

Theses Canada A union catalog of Canadian theses and dissertations, in both electronic and analog formats, is available through the search interface on this portal.

WorldCat NDLTD Union Catalog hosted by OCLC.

Foreign Comprehensive/Gateway Repositories

Gallica
http://gallica.bnf.fr/
Supported by French National Library with multiple European languageoptions for interface

Hispana
http://hispana.mcu.es/en/estaticos/contenido.cmd?pagina=estaticos%2Fpresentacion
(English language interface)
Supported by Spanish Ministry of Culture with multiple European language options for interface

Pacific Rim Digital Library Alliance
http://prl.lib.hku.hk/exhibits/show/prdla/home
Supported by twenty-eight academic libraries surrounding the Pacific with multiple Asian language options and English for interface.

Trove
http://trove.nla.gov.au/
Supported by National Library of Australia. Include online resources like
books, images, historic newspapers, maps, music, archives and more.