Interoperability Between Institutional and Data Repositories: a Pilot Project at MIT

Katherine McNeill*

Abstract

Academic libraries are working in new areas to support the publishing activities of their institution's faculty members, including helping them to manage and archive research data that they produce. Many institutions, such as the Massachusetts Institute of Technology, have multiple locations in which faculty can deposit their data. Yet this distributed arrangement presents challenges for searching, unifying collections, and archiving. In order to foster some interoperability between these multiple data repositories, the MIT Libraries developed a prototype system to bring studies between two such systems, DSpace and the Institute for Quantitative Social Science Dataverse Network, by enabling the harvesting and replication of metadata and content across the two systems. This paper will discuss the motivation for this project, details and challenges of the system, and future goals for enhancing interoperability among the two systems.

Literature Review

Many academic library systems, such as the one at the Massachusetts Institute of Technology (MIT), have been developing more services in recent years to support the publishing activities of their faculty. Developing institutional repositories (IRs) for housing and disseminating the digital research materials produced by an institution is a main area of work. Academic librarians play a key role in promoting and facilitating the use of IRs (Bailey 2005). These activities provide new opportunities for librarians to become partners in publishing with their faculty, which can enrich their relationships and increase the library's relevance (Buehler and Boateng 2005; Bell, Foster, and Gibbons 2005). However, many IRs are experiencing low rates of faculty contribution (McDowell 2007). In order to enhance participation, many librarians are working to evaluate the utility of their IR from their faculty's perspective. Some institutions have undertaken projects to study faculty work practices in order to design the repository system which best meets faculty needs. One such project discovered that faculty members must be able to personalize their presence in the IR in order for it to provide them with significant value (Foster and Gibbons 2005).

A recent study indicates that datasets comprise only a very small percentage of items in IRs (McDowell 2007). In this context, many librarians assist faculty members in publishing their datasets, whether it is in their IR, a domain-specific data repository, or another location. For example, Purdue University library has established the Distributed Data Curation Center (D2C2) to support the curation and archiving of faculty-produced data.[1] Success in this work requires an understanding of the needs of individual faculty members in order to recommend to them an appropriate system for managing and archiving their data (Witt and Carlson 2007). Moreover, a viable data archiving system must be of tangible benefit to the depositor, not just the secondary data user. One study argues that a requirement for citation of datasets by secondary users would be the best incentive for faculty to prepare their data appropriately for deposit in a data archive (Niu 2006). For several years, members of the social science data community have been promoting the need for standards for citing data. Some have developed specific standards recommendations designed to interoperate with data repository systems (Altman and King 2007). All these studies shed light on how to design data repositories in alignment with the needs of faculty and researchers.

A range of different kinds of digital repositories exists: "individual, discipline-based, institutional, consortial, and national" (Peters 2002). Given this landscape, there often are multiple locations where an individual faculty member can publish and archive data, each of which may have its own approach to and policies regarding archiving and management. These variations in service may make one repository more appealing to a faculty member, and thus implicates the choices she must make (and thus the availability of her research data). How might two kinds of repositories, IRs and domain-specific data repositories, come together? Green and Gutmann envision a collaborative system whereby the IR facilitates communication and exchange of data between the researcher and the domain repository (Green and Gutmann 2007). The ability for different repositories to exchange metadata and content would provide an important service to enable faculty data to be housed and discovered in more than one system.

Working examples of systems that can exchange metadata and content among data and/or institutional repositories exist in the field. One such project, BibApp, developed by the University of Wisconsin-Madison and University of Illinois Urbana-Champaign, has designed software to index, search, and harvest from the web (for one's local IR) publications of university faculty.[2] Other repositories[3] have the capability to harvest metadata and content from other repositories compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[4] Other funded projects aim to develop models for transporting metadata and digital objects among repositories of different structures (FCLA Digital Archives).

The Harvard-MIT Data Center (HMDC), a member of the Institute for Quantitative Social Science at Harvard University, has significant experience in developing systems that enable the exchange of metadata and content among data repositories. HMDC has developed and deployed two successive systems of open-source data repository software, the Virtual Data Center and Dataverse Network Software systems (Altman et al. 2001 and King 2007). HMDC has utilized these systems to harvest metadata and content from partner data archives, such as ICPSR and the Roper Center for Public Opinion Research. Continuing these projects, HMDC now operates the shared catalog for the Data Preservation Alliance for the Social Sciences (Data-PASS) (Altman et al. 2009).

Supporting MIT Faculty as Data Producers
How do the MIT Libraries support their faculty members so that they can archive and publish their research data? Within the MIT Libraries Social Science Data Services program,[5] the Data Services Librarian encourages faculty to archive and disseminate data that they produce and helps them to do so. As discussed earlier, the ability to support faculty in their publishing efforts provides new opportunities to be of utility to faculty and share with them the Libraries' expertise in this area (Buehler and Boateng 2005).

MIT faculty members in the social sciences have three main options for where they can deposit data that they have produced. DSpace, MIT's institutional repository is a Dublin-Core-based IR system utilizing software developed jointly by the MIT Libraries and Hewlett Packard Laboratories (Smith 2002).[6] DSpace is committed to preserving not only data sets but also any MIT-produced material (e.g., working papers, images, etc.). The Harvard-MIT Data Center (HMDC) provides MIT with its own customized data repository at the Institute for Quantitative Social Science (IQSS) Dataverse Network.[7] This Data Documentation Initiative (DDI)[8] -compliant system is based on the Dataverse Network Software (DVN) developed at Harvard (King 2007). MIT can load into the IQSS DVN any data that MIT licenses, purchases or produces. Lastly, MIT faculty also can deposit their data

in the Inter-university Consortium for Political and Social Research (ICPSR) data archive.[9]

Each system has its own advantages and challenges. DSpace is more likely to have an established workflow for loading items in a faculty member's department, yet lacks specific features for working with data. IQSS DVN has tools for online data manipulation and enables a high-level of control by the faculty member. Moreover, faculty members can create personalized home pages to highlight their data. This feature has been a selling point for some faculty, supporting research documenting the need for personalization (Foster and Gibbons 2005). ICPSR is a formal, full-service archive with staff that can guide depositors in preparing their data for archiving and distribution and can perform additional services such as a confidentiality review and documentation enhancement. Different MIT faculty members in departments such as economics, history, and political science, have deposited items in each of these repositories, depending upon their individual needs and preferences. In working with these systems and facilitating faculty deposit, the MIT Data Services Librarian has been involved in many tasks associated with managing IRs (Bailey 2005).

The Data Services Librarian begins a consulting arrangement with a one-on-one meeting with the faculty member to first to understand her data management and archiving needs (Witt and Carlson 2007). The Data Services Librarian highlights the benefits of archiving personal research data; discusses with the faculty member which if any of the aforementioned repositories will suit her data; answer questions; and coaches her to a decision as to if, and where, to archive her data. To support this activity, the Data Services Librarian worked with two other MIT librarians to develop a web site for faculty on Data Management and Publishing.[10] Experience has shown however, that in-person meetings (rather than the simple existence of informative web pages) are necessary to give faculty the information and incentive needed to start such a project.

While there are benefits to having multiple options for archiving faculty-produced data, this situation also can lead to certain challenges. End users have to search multiple systems and there is no unified collection for a given faculty member. Most importantly, it is a challenge to help faculty decide where to put their data. MIT has an interest in--and thereby a desire to promote--all three aforementioned systems, yet one cannot expect faculty to deposit in more than one. Each system has its strengths, yet it would be beneficial if faculty-produced data could be housed and discovered in all of them. For some time, the MIT Libraries have been considering how enabling some level of interoperability among these systems could help ease these problems.

PLEDGE project
In 2005, the MIT Libraries had an opportunity to work on a project to develop limited interoperability, in the form of metadata and content exchange, between two of these three systems for data deposit: DSpace and IQSS DVN. The PLEDGE (PoLicy Enforcement in Data Grid Environment) Project, a partnership with the San Diego Supercomputer Center and University of North Carolina, Chapel Hill, explored the use of data grid technology (i.e., distributed data storage infrastructure) for replication of content across systems for preservation purposes.[11] As part of this grant, MIT worked on a specific project to exchange metadata and content between DSpace and IQSS DVN.

In order to address the challenges posed by housing data in multiple locations, MIT took on this project with the goal to develop a mechanism to archive, preserve, and provide access in DSpace to MIT-authored studies in DVN. Such a system would allow DSpace (with its mission to preserve MIT-produced material) to archive the studies and enable users to find the studies from within DSpace, while still being able to access the studies separately via DVN, which features unique services tailored to manipulating data. As a result, this system demonstrates how DSpace can archive MIT content while at the same time allow MIT-produced research to be discovered and housed in other specialized systems. In line with the goals of PLEDGE, this project enhances the preservation of these data files through the replication of content. It is important to understand that the interoperability achieved in this project is limited to the exchange of metadata and content, and does not extend to other possible services such as integrated access or shared interfaces.

Within the past couple of years, MIT faculty members have begun to load data that they produce into HMDC. The first MIT group to do so was the Abdul Latif Jameel Poverty Action Lab (J-PAL), a research lab associated with the Department of Economics.[12] By putting their data in DVN, J-PAL could take advantage of all of its data-specific features. But its data was neither housed nor discoverable in DSpace, the system for preserving MIT-produced research. This set of data was the test case for the PLEDGE project, which was designed to bring the MIT-authored studies housed in DVN, such as those from J-PAL, into DSpace. Note: overall, MIT-authored studies comprise only a subset of studies that MIT stores in HMDC, which also houses materials licensed or purchased (but not produced) by MIT.

DSpace and DVN were selected for the

project because they are both home-grown and have MIT involvement. The main MIT staff member working on the project was the DSpace System Manager, who previously was a developer of software at HMDC (and thus had a key intersection of skills for the project); the Data Services Librarian provided information and advice on the project.

How the System Works
The DSpace System Manager designed an agent (working with manual involvement by both the System Manager and the Data Services Librarian) to harvest and replicate metadata and content across DVN and DSpace. In order to accomplish these tasks, the agent converted metadata and information packages between different formats used by the two systems. Therefore, the agent was designed to transform DDI metadata (used by DVN) into METS (Metadata Encoding and Transmission Standard), a submission package standard that allows for the exchange of both content and metadata.[13] The METS package includes specific items from the DDI record, including MODS (Metadata Object Description Schema) descriptive metadata[14] and PREMIS (Preservation Metadata) technical metadata.[15] This new set of metadata then needs to be converted into a form that can be understood and processed by DSpace. DSpace ingests digital objects in submission information packages (SIPs), based on the Reference Model for an Open Archival Information System (OAIS) (Consultative Committee for Space Data Systems 2002). Next, the agent produces a stand-alone, self-describing zip package that then is ingested into DSpace, creating an item (and associated catalog record) in the repository.

The system works in four main steps (see Figure 1).

Figure 1: System Overview
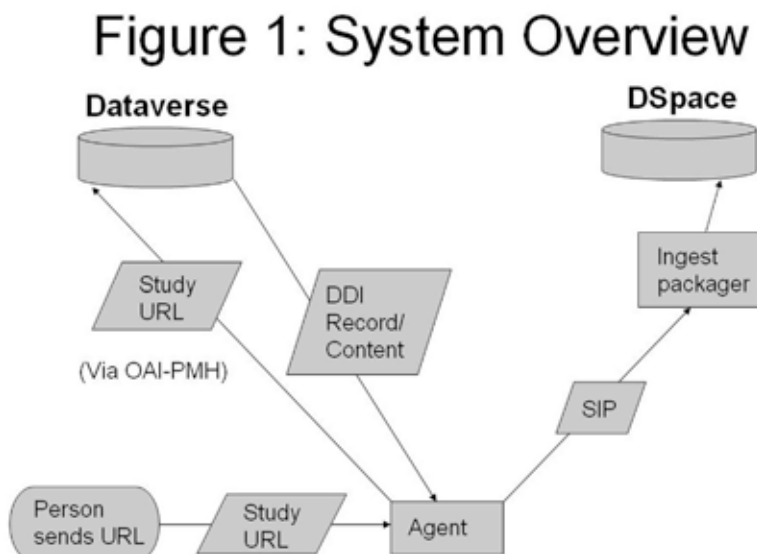
1. The System Manager sends the URL for a particular DVN study to the agent.

2. The agent then, utilizing a given URL for a study, harvests a DDI record and all the appropriate study content (i.e., data and related files such as a codebook) from DVN, via OAI-PMH, the protocol for metadata harvesting.[16]

3. The agent packages the content into a SIP zip file containing:

- METS file (including MODS descriptive metadata and PREMIS technical metadata)

- DDI file

- Content file(s) (data file and others associated with a given study, e.g., codebook or other documentation)

4. The agent then sends the SIP to the DSpace DDI Ingest Packager. This tool processes the package to produce a DSpace item described in a Dublin Core metadata record. It takes all other items in the package (DDI, data and other files) and attaches them as files associated with the DSpace item.

Figures 2 and 3 are screenshots of a sample data file that has been brought through this system. Figure 2 shows the catalog record for one of the data files that J-PAL had submitted to DVN. The study record in DVN has some DDI-specific metadata fields: e.g., geographic coverage, unit of analysis, etc. In that system, online subsetting and statistics feature are also available. Figure 3 shows the item in DSpace after it was processed by the agent. The metadata record is simpler because it is described in the Dublin Core format. The DDI metadata from DVN now is available as an additional file associated with the item in DSpace. This particular study is included in the Datasets Collection within the J-PAL Community in DSpace. In the future, J-PAL could load into DSpace other content types that they produce, such as papers, images, etc. (see Figures 2 & 3)

Discussion
Despite its benefits, the system is not without its challenges. One is the inefficiency of the workflow for selecting and processing studies. Since only the MIT-authored data files in DVN are being harvested for DSpace, a librarian needs to make a manual selection decision to start the process. For now the process begins with the Data Services Librarian notifying the DSpace System Manager which studies are MIT-authored, and then the latter executes the agent. This is a functional, but not particularly efficient or scalable, system. However, with further effort, one might be able to design a more automated system. This system could take the shape of a form in DSpace in which one enters the URL for a study at DVN, initiating a set

of automated processes that execute the workflow fully without further manual intervention. Alternatively, the DVN software is designed so that an administrator can create a custom OAI set based on metadata criteria such as author or DVN collection. An OAI client could be developed that would monitor this set and feed new or updated studies into the DSpace ingest system, creating a certain level of synchronization of DVN to DSpace.

The interaction between these two systems also has implications for licensing. Normal DSpace loading workflow includes some licensing screens. The MIT submission process includes a click through agreement makes MIT distribution policy and contributor responsibilities explicit; the author grants the MIT Libraries the right to distribute his/her content. In the prototype system at MIT, an agent loads the studies in the back end without the author viewing those screens (in this test case, agreement to the terms was obtained through an informal email exchange with the author). Moreover, DVN has specific pop-up windows whereby secondary data users must agree to text included in the terms of use element in the DDI metadata before the system allows them to download the data file. In DSpace, these terms of use are simply described within the DDI file attached to the record of the item. The Libraries are considering the implications of these issues.

Updating is one of the challenges of replication of content. Given the fact that data and related files have copies at multiple locations, currently no process exists to update all copies of files simultaneously. For example, if a faculty member corrects and error in a DVN submission, updating a file, there is no automated method of tracking and providing notification of this change to trigger an update of the file in DSpace, accordingly. Currently, Libraries' staff would need to somehow learn about this event and manually update the file in DSpace. Investigation may yield a more automated alternative in the future, such as the synchronization system described earlier. In addition, it should be noted that the agent was designed to work with the systems as they were configured at a particular point in time. As each system gains new features, moves to a different platform or infrastructure, or is revised to be compliant with a new version of its metadata standard, someone will need to re-program the agent to keep it up-to-date.

Future Work
The next step in this project, after the success of the prototype, is to formalize the service. Currently the DSpace System Manager is the only one who knows how to run the agent. The MIT Libraries therefore now are working to document the system and integrate the service into the normal workflows of DSpace staff members in charge of the local repository. In addition, the Libraries hope in the future to design more automation around the use of

Figure 2

Figure 3

the system. However, given the number of MIT social scientists known to produce data, the Libraries expect to receive a relatively low volume of studies to be processed by this system each year, so manual execution certainly is feasible in the short-term.

Thinking farther into the future, a major improvement to the system would be to enable data exchange in the other direction, i.e., bringing studies housed in DSpace into IQSS DVN. Many faculty members choose to house their data in DSpace given existing workflows in their departments for loading other materials. Therefore, allowing access in DVN to studies from DSpace would further enhance services and might enable the files originally only housed in DSpace to utilize the data-specific features in DVN. However, this is a more complex challenge because of the different metadata specifications used by the systems: Dublin Core in DSpace and DDI in DVN. Taking records based on simpler metadata (Dublin Core) into a system with more complex metadata (DDI) poses a difficult challenge.

In one scenario, DVN could simply harvest the Dublin Core metadata records and map them to a limited set of DDI elements. In this model, studies from DSpace could be found via DVN searches; DVN then could direct the user to DSpace to download the study. However, more robust data-specific searches would require this metadata to be extended to further elements of the DDI, requiring additional cataloging. Moreover, to import data files from DSpace into DVN, the format of those data files would determine whether or not they would be compatible with the online analysis features of DVN. More complete interoperability, such as integrated access or shared interfaces, could be explored in the future as well.

In addition, it would be a worthwhile effort to explore if this service could be expanded to exchange metadata and content with other data repositories, such as ICPSR. This would further extend discovery and utility of MIT faculty-produced data and work towards a system of partnerships between local and domain-specific repositories (Green and Gutmann 2007).

Conclusion
Despite these challenges, PLEDGE project success can be of benefit to other institutions. While this system was built to address a particular local need, it can inform other projects to share and replicate content across repositories. The prototype demonstrates the use of packaging standards and strategies for delivery of data to exchange metadata and content between two systems. In addition, this project has lessons for other sets of information. The agent designed is not DVN-specific but can work with any DDI-based system and thus could be used by others. Moreover, the project illustrates the ability to harvest both metadata and content across systems based on different metadata conventions, i.e., was not limited to those in DDI format. Many

researchers now are working with diverse groups of content files and corresponding metadata culled from the web, utilizing the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) standard (Open Archives Initiative a). This system demonstrates how to bring these complex digital objects into a repository. In conclusion, this project devised a prototype system that accomplishes several goals. It enhances the discovery and preservation of MIT-created data files in these two systems used and maintained locally. In addition, it demonstrates how to package and import DDI and related data into a greater variety of systems, holding the promise for more interoperability among systems in the future.

* Katherine McNeill, Data Services and Economics Librarian, Massachusetts Institute of Technology. E-mail: mcneillh@MIT.EDU

References
Altman, Micah, Margaret Adams, Jonathan Crabtree, Darrell Donakowski, Marc Maynard, Amy Pienta, and Copeland Young. 2009. Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. The American Archivist (Forthcoming, 72 (1) (Spring/Summer 2009)).

Altman, Micah, Leonid Andreev, Mark Diggory, Gary King, Akio Sone, Sidney Verba, Daniel L. Kiskis, and Michael Krot. 2001. A digital library for the dissemination and replication of quantitative social science research: The Virtual Data Center. Social Science Computer Review 19 (4): 458-70.

Altman, Micah, and Gary King. 2007. A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine 13 (3/4) (March/April 2007), http://www.dlib. org/dlib/march07/altman/03altman.html.

Bailey, Jr, Charles W. 2005. The role of reference librarians in institutional repositories. Reference Services Review 33 (3): 259-67.

Bell, Suzanne, Nancy Fried Foster, and Susan Gibbons. 2005. Reference librarians and the success of institutional repositories. Reference Services Review 33 (3): 283-90.

Buehler, Marianne A., and Adwoa Boateng. 2005. The

evolving impact of institutional repositories on reference librarians. Reference Services Review 33 (3): 291-300.

Consultative Committee for Space Data Systems. 2002. Open archival information system (OAIS). [cited September 16 2008]. Available from http://public.ccsds.org/publications/archive/650x0b1.pdf.

FCLA Digital Archives. Archives for: September 2008. [cited September 19 2008]. Available from http://blogs.fcla.edu/index.php/digitalarchive/2008/09.

Foster, Nancy Fried, and Susan Gibbons. 2005. Understanding faculty to improve content recruitment for institutional repositories. D-Lib Magazine 11 (1) (January 2005), http://www.dlib.org/dlib/january05/foster/01foster.html.

Green, Ann G., and Myron P. Gutmann. 2007. Building partnerships among social science researchers, institution-based repositories and domain specific data archives. OCLC Systems & Services 23 (1): 35-53.

King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. Sociological Methods & Research 36 (2): 173-99.

McDowell, Cat S. 2007. Evaluating institutional repository deployment in american academe since early 2005. D-Lib Magazine 13 (9/10) (September/October 2007), http://www.dlib.org/dlib/september07/mcdowell/09mcdowell.html.

Niu, Jinfang. 2006. Reward and punishment mechanism for research data sharing. IASSIST Quarterly 30 (4) (Winter 2006): 11-5.

Open Archives Initiative. a. OAI-ORE. [cited September 16 2008]. Available from http://www.openarchives.org/ore.

Peters, Thomas A. 2002. Digital repositories: Individual, discipline-based, institutional, consortial, or national? The Journal of Academic Librarianship, 28 (6): 414-7.

Smith, MacKenzie. 2002. DSpace: An institutional repository from the MIT libraries and Hewlett Packard Laboratories. In Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries (ECDL'02), LNCS 2458, Rome, Italy, ed. Agosti, M. and Thanos, C., 543-549. Berlin: Springer.

Witt, Michael, and Jake R. Carlson. 2007. Conducting a data interview (conference poster). Washington DC, USA. [cited September 16 2008]. Available from http://www.dcc.ac.uk/events/dcc-2007/posters/data_interview.pdf.

Footnotes

1. http://d2c2.lib.purdue.edu.

2. http://code.google.com/p/bibapp.

3. http://www.policyarchive.org.

4. http://www.openarchives.org/OAI/openarchivesprotocol.html.

5. http://libraries.mit.edu/guides/subjects/data.

6. http://dspace.mit.edu.

7. http://dvn.iq.harvard.edu/dvn/dv/mit.

8. http://www.ddialliance.org.

9. http://www.icpsr.umich.edu.

10. http://libraries.mit.edu/guides/subjects/data-management.

11. http://pledge.mit.edu.

12. http://www.povertyactionlab.org.

13. http://www.loc.gov/standards/mets.

14. http://www.loc.gov/standards/mods.

15. http://www.loc.gov/standards/premis.

16. http://www.openarchives.org/OAI/openarchivesprotocol.html.