Case Study in Data Curation at Johns Hopkins University

G. Sayeed Choudhury

Abstract

At Johns Hopkins University, the institutional repository (IR) is being developed as a component of an overall digital library architecture that will emphasize long-term preservation. The IR represents a set of services that will be developed to support the identified needs or requirements of faculty and students. Given the research-intensive environment at Johns Hopkins, one particular area of interest relates to data sets from a diversity of disciplines ranging from the humanities to the sciences. Essentially, the IR is being developed as a "gateway" to the underlying digital archive that will support data curation as part of an evolving cyberinfrastructure featuring open, modular components. In addition to this technological framework, Johns Hopkins is developing new roles and relationships between the library and the academic community, most notably through the development of "data scientists" or "data humanists." These developments reflect the realization that the IR is the first step in a longer journey and that for institutional efforts to be successful, they must be integrated into a larger landscape of repositories that serve a distributed and diverse academic community.

Concerning Institutional Repositories

Within only a few years, IR-related discussions have moved from great promise to the inevitable "reality check" that has left the library and academic community with a more nuanced perspective. The discrepancy between promise and reality has led some to assert that the IR movement has already failed. At a December 2007 data curation workshop sponsored by the Andrew W. Mellon Foundation and the Joint Information Systems

LIBRARY TRENDS, Vol. 57, No. 2, Fall 2008 ("Institutional Repositories: Current State and Future," edited by Sarah L. Shreeves and Melissa H. Cragin), pp. 211–220

Committee, Greg Crane, editor-in-chief of the Perseus Project and professor of Classics at Tufts University, stated "no movement has had more promise or has delivered less on that promise than institutional repositories" (Jacobs, 2007). Ultimately, this reflection will prove to be useful since initial expectations were perhaps too premature or optimistic. Institutional repositories were cited as a strategy for a number of broad ranging topics such as the university's relevance in the digital age or the scholarly communication crisis (Crow, 2002). Each of these beliefs or hopes reflected an earnest attempt to address these important topics. With the benefit of hindsight, it seems that some institutions hoped that simply installing an IR would result in transformative effects. Technology alone cannot engender transformation. Additionally, many of the earlier conversations regarding IRs tended to focus on benefits to the host institutions, rather than the scholars who would use IRs.

Even if the IR movement has not failed, it is at a critical juncture. There is notable support and interest within the library and university community for IRs. Perhaps more importantly, there remains a great deal of trust, even faith, associated with IRs as the means to regain control of the scholarly process and to support new forms of research, learning, dissemination, and preservation. A successful trajectory of the IR movement will depend on a healthy balance between the original ideas of great promise and a pragmatic, honest assessment of current practices, especially as they compare to the evolving needs of scholars. Given that IRs are sometimes defined as a means for capturing the intellectual output of a specific institution, it may seem ironic to assert that it will be necessary to move away from a collection or institution-centric view for IRs to serve scholars most effectively.

Though this approach may seem contradictory, Crow (2002) stated that "While institutional repositories centralize, preserve, and make accessible an institution's intellectual capital, at the same time they will form part of a global system of distributed, interoperable repositories that provides the foundation for a new disaggregated model of scholarly publishing." Lynch (2003) provided a useful definition of an IR that emphasizes services:

> a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.

These definitions emphasize IRs as vehicles for delivery of services such as long-term preservation rather than systems for collection development, and as part of an overall landscape, the latter point being reinforced by Duranceau's article in this issue. This viewpoint of IRs offers a useful framework for future development. In a fundamental sense, it will be

important to return to the first principles that prompted interest in the IR movement, rather than focus on the current practices that have defined the IR movement to date.

Incentives and Motivation

One of the main reasons for disappointment with the IR movement is the slow rate at which individual IRs have become populated with content. Others have discussed possible reasons for this low level of engagement (see Salo's and Duranceau's articles in this issue). This is especially disconcerting if one considers that some libraries have viewed IRs as a means for (digital) collection development, at least within their host institutions. There is at least anecdotal evidence that deposit rates increase with appropriate outreach, awareness building, and marketing efforts. Many IR proponents advocate mandatory deposit requirements as a means to increase participation. While these approaches increase the amount of content within IRs, they do not address the question of why *voluntary* submission rates have been low. When left to their own decisions about how to allocate a scarce resource, many scholars have seemed reluctant to choose submission to the IR as good use of their time.

This article focuses on two factors that relate to the scholarly publishing and institutional landscape, which were identified in earlier articles on IRs. In his opinion piece on institutional repositories, Johnson (2002) identified potential changes in the scholarly publishing system and increased visibility for institutions as two major motivators for IRs. He also posed the critical question "What's in It for Faculty and Researchers?" While acknowledging that faculty's reward structure is tied fundamentally to the existing publishing system, he asserted that greater visibility, increased citation, and broader dissemination of results might motivate faculty to embrace the IR movement. He also mentioned that unlike disciplinary repositories, "institutional repositories represent an historical and tangible embodiment of the intellectual life and output of an institution. And, to the extent that institutional affiliation itself serves as the primary qualitative filter, this repository becomes a significant indicator of the institution's academic quality."

Underlying these statements was an assumption that faculty and researchers felt an urgent need to change the scholarly publishing system and that they felt a need to promote their host institution. These arguments were more important for libraries and universities, rather than individual faculty and researchers, especially since they often have closer ties to their project teams or professional societies. Davis & Connolly (2007) state, "While some librarians perceive a crisis in scholarly communication as a crisis in access to the literature, Cornell faculty perceive this essentially as a non-issue. Each discipline has a normative culture, largely defined by their reward system and traditions." Noting the important caveats that the

analysis was focused on a single institution and did not address the possible impact of outreach or marketing efforts, they provide evidence of this misalignment about incentives and motivation. If IRs were advanced as a solution to problems that do not interest faculty, it is not surprising that voluntary submission rates have been low.

This misalignment has unfortunate consequences in terms of the potential value of IRs, especially since first impressions matter. Salo's article (this issue) describes a case of plagiarism that led a faculty member to question the value of the open-access movement. This reaction is perfectly understandable, especially if one views the IR as a content repository primarily for the institution's benefit. However, if one views the IR as an engine for services for the scholar's benefit, it might be possible to revisit this plagiarism concern. Sorokina et al. (2007) have developed methods for detecting plagiarism within a research document collection. With such tools and services, might it be possible to persuade faculty that depositing a paper into an IR may actually *reduce* plagiarism? Another important example relates to the National Institutes of Health mandate to deposit papers into PubMed Central (http://publicaccess.nih.gov). Faculty might be more motivated to submit papers to an IR if it meant that those papers were also automatically submitted into existing publishing workflows or repositories including PubMed Central.

For the IR movement to succeed, it will be essential to focus on well-defined *requirements*, not *assumptions*, and to recast IRs as part of an overall scholarly landscape, rather than a distinct entity unto itself. Rather than promote IRs as a cause and effect relationship, it may be preferable to think of them in a correlational manner. For instance, depositing papers into an IR will not bring about a change in the scholarly communication process, but engaging faculty in discussions about the value of IR services might help raise awareness about the issues related to the current scholarly publishing system. Duranceau's article (this issue) describes such developments at MIT through their repository outreach efforts that ultimately highlight potential value of the IR. This type of direct engagement with faculty will be critical for the long-term success of IRs.

Data Curation

Faculty at Johns Hopkins University (JHU) associated with community-wide eScience projects have identified data curation as one of the most important repository-related services. The Digital Curation Centre (DCC) (http://www.dcc.ac.uk/) defines digital curation as "maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific materials" (DCC, n.d.) While there are several major eScience projects involving JHU, the initial dialogue on data curation has focused on the Sloan Digital Sky Survey

(SDSS) (http://www.sdss.org/) and the National Virtual Observatory (NVO) (http://www.us-vo.org/). The dialogue between digital librarians and astronomers has resulted in a greater understanding of the transformative nature of data-intensive science, especially as it relates to new forms of publication, research, and learning (Choudhury, 2008). This dialogue has also revealed important insights and observations regarding IRs in the context of data curation.

First, given the scale and complexity of certain astronomy datasets, it became clear that it would not be possible to ingest these particular datasets into an IR, particularly with systems that were designed for documents rather than data. Second, even in cases where an IR could accommodate data of smaller scale, it would not be appropriate to assert specific institutional ownership given the multi-institutional nature of these projects. Third, scientists wish to produce new forms of publications that comprise both articles and data, both of which can be traced back to source data in distributed repositories or content stores.

Returning to the original ideas about IRs and scholarly communication, it is important to note institutional repositories did not inspire changes in scholarly communication, but they could play an important role in supporting new forms of data-intensive scholarship. Scientists' incentives for changing the scholarly communication process do not relate to institutional needs, but rather the reality that data have become a new form of publication, which are critical for their research and teaching purposes. Promoting IRs as a solution to problems that may not concern faculty has been unproductive. However, presenting IRs as a mechanism for housing certain data as part of a compound object publication could be more productive. Perhaps more importantly, IRs could become an important component in a data curation strategy.

These important characteristics can be demonstrated through a current project at JHU. The JHU Sheridan Libraries and the NVO are working with the American Astronomical Society (AAS) and its publishing partner to develop a data curation prototype system that connects digital archiving and electronic publishing systems (Choudhury et al., 2007). For this project, the compound object publication is being modeled using the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) protocol (Van de Sompel et al., 2006). OAI-ORE "defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video" (Open Archives, n.d.). OAI-ORE specifies the concept of Resource Maps (ReMs) to describe these aggregations. At some fundamental level, OAI-ORE acknowledges the realization that repositories are not an end, but rather a means to participate in a distributed network of content and services.

Figure 1 depicts an OAI-ORE based model for an astronomy article and data. The ReM (marked by the exterior dashed line) encompasses both an aggregation of multiple objects (marked by the interior dashed line) and additional objects beyond the confines of the aggregation. The article ("A-1") comprises text, tables, figures, and associated data (such as "DR-5" or "DR-3"). These directly cited data are derived from other sources such as the data released from the SDSS (those data identified outside the boundaries of the aggregation such as SDSS DR4). Without even delving into the details of this figure, it is apparent that there is a complex web of potential connections. An individual article may cite multiple datasets, or multiple articles may cite an individual dataset. A researcher who reads an article may wish to review the directly cited datasets, but also know the source from which those datasets were derived and their provenance.

It is unrealistic, even impossible, to imagine any single IR housing all of the objects depicted in this model. However, an IR might contain an individual element or some of the elements within this compound object. Documents such as scholarly articles, electronic theses and dissertations, and grey literature can be easily accommodated within IRs. Derived astronomy datasets and other datasets from "small science" (e.g., individual researchers or small teams conducting bench or laboratory science) are appropriate candidates for depositing into IRs. These documents and data could be conceived of as part of a network of compound objects instead of an institutional asset—a view that resonates more readily with faculty associated with the community-wide astronomy projects.

Most importantly, these astronomers have identified the preservation aspect of data curation as a critical requirement. While there are mechanisms in place for large datasets associated with specific projects, the most highly processed datasets that are derived by individuals from analyses of large databases often reside on websites, individual workstations, etc. Without a systematic effort to capture and archive them, these datasets (like those from "small science" projects) remain at risk. There is a real urgency on the part of the astronomers related to this particular topic. For this reason, this topic of preserving and representing these aggregations of articles and derived datasets has proven useful for promoting the use of repositories, including the institutional repository. While depositing these objects into a repository does not constitute preservation, it is an important step toward systematically capturing objects and attaching or generating preservation metadata (e.g., checksums).

Institutional Repositories and Infrastructure

As the community considers the trajectory and fate of IRs, it is worth considering them in the context of infrastructure building efforts. Edwards, Jackson, Bowker, and Knobel (2007) described the historical trends of in-

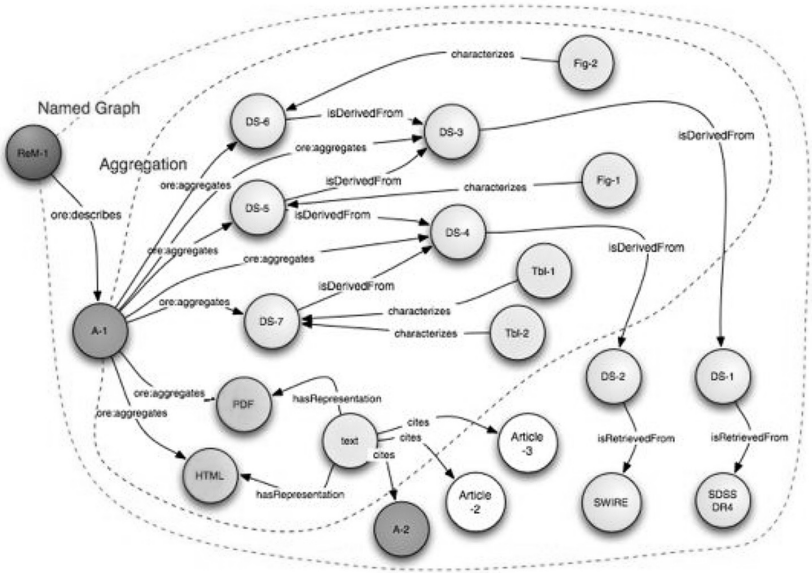

Figure 1. OAI-ORE based model for astronomy data and article

frastructure building efforts in a report of a National Science Foundation sponsored workshop. This report states that an

> initial stage in infrastructure formation is system-building, characterized by the deliberate and successful design of technology-based services. Next, technology transfer across domains and locations results in variations on the original design, as well as the emergence of competing systems. Infrastructures typically form only when these various systems merge, in a process of consolidation characterized by gateways that allow dissimilar systems to be linked into networks.

With this insightful characterization, IRs can be considered one of many systems that will ultimately come together into an overall infrastructure, which will include both technology and human components. Duranceau's article (this issue) points out the importance of new library professionals who can gather requirements, explain the value of IRs, and advance IR development accordingly. The data curation activities at JHU have highlighted the importance of new roles of "data scientist" or "data humanist." These individuals, who are few in number at the moment, possess domain-specific knowledge and data management expertise. They act as the human interface between the library and the eScience projects. In a fundamental sense, they may represent the future of subject librarianship and help craft a new relationship between the library and scientists.

Scientific datasets may be thought of as the "special collections" of the digital age (Choudhury & Stinson, 2007). Libraries' role in preserving and curating special collections has led to a deep engagement with humanists who see the library as an objective, trusted laboratory. As libraries develop data curation programs, scientists may start to see libraries in a similar manner, leading to greater advocacy and support within and without the university.

An example of this support is the newly formed Institute for Data Intensive Engineering and Science (IDIES) at JHU. IDIES was formed as an organizational cluster for eScience activities at JHU. Its charter document states: "A unique opportunity exists, therefore, to coalesce data-intensive science efforts at Johns Hopkins into a well-focused center of activity, which would then propel various fields towards new discoveries and breakthroughs." This institute was launched with support from the Krieger School of Arts and Sciences, the Whiting School of Engineering, and the Sheridan Libraries. The library was included as a charter member in this important development because of its data curation program, of which the IR is a component. Most recently, the founders of IDIES have discussed the possibility of extending its capabilities and reach into digital humanities. The Sheridan Libraries acted as the bridge between the scientists and humanists to initiate this dialogue. Arguably, the library, with its mandate to support all disciplines through services such as the IR, is uniquely positioned to make such connections.

Conclusions

It is tempting to reach conclusions regarding the value of IRs, but it is important to realize that these are the early days. There will be inevitable "growing pains" as the library and academic community experiments with IRs. While current repository platforms may not be ideal in terms of ease of use or functionality, it is critically important to note the open-source nature of DSpace (http://www.dspace.org), ePrints (http://www.eprints.org), and Fedora (http://www.fedora-commons.org) supports experimentation and development. There is no doubt that it takes effort to customize these software platforms, but it is effort well worth making. And it is similar to other infrastructure development efforts that required patience before reaching maturity, ubiquity, and seamlessness. It is also worth noting that faculty do not always act rationally or even toward their best interests. Instead of viewing faculty engagement through new roles or relationships as an indicator of shortcomings in IR software, perhaps it might be viewed as a sign of healthy professional development. As new systems evolve into infrastructure, it will be essential to develop trust. Friedlander (2008) correctly states that trust develops when infrastructures "will do what we expect them to do and not do what we expect them not to do." For this reason, it will be essential to understand requirements, set

expectations appropriately, and build on the foundation established by existing IRs. Earlier attempts to advance IRs seemed to focus on an exclusive role, rather a complementary or integrated role within a network or ecosystem. Rather than casting the IR as an all or nothing proposition, its future lies in determining its appropriate place in the newly evolving data-intensive scholarly landscape.

Acknowledgments

I am deeply grateful to Alexander Szalay and Robert Hanisch for the ideas related to SDSS and NVO, and to Timothy DiLauro and David Reynolds for the OAI-ORE data model described in this article. The Institute of Museum and Library Services (NLG Grant # LG0606018206) and Microsoft have provided funding for the NVO data curation prototype development. Finally, I thank the reviewers and editors of this issue of *Library Trends*.

References

Choudhury, G. S. (2008). The virtual observatory meets the library. *Journal of Electronic Publishing, 11*(1). Retrieved August 8, 2008, from http://hdl.handle.net/2027/spo.3336451.0011.111

Choudhury, G. S., & Stinson, T. L. (2007, December 16). The virtual observatory and the Roman de la rose: Unexpected relationships and the collaborative imperative. *Academic Commons*. Retrieved August 8, 2008, from http://www.academiccommons.org/commons/essay/VO-and-roman-de-la-rose-collaborative-imperative

Choudhury, G. S., DiLauro, T., Szalay, A., Vishniac, (2007). Digital data preservation for scholarly publications in astronomy. *International Jounral of Digital Curation, 2*(2). Retrieved August 8, 2008, from http://www.ijdc.net/ijdc/article/view/41/48

Crow, R. (2002). *The case for institutional repositories: A SPARC position paper*. Washinton, D.C.: Scholarly Publishing and Academic Resources Coalition. Retrieved August 8, 2008, from www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

Davis, P. M., & Connolly, M. J. (2007, March/April). Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine*. Retrieved August 8, 2008, from http://www.dlib.org/dlib/march07/davis/03davis.html

Digital Curation Centre (DCC). (n.d.). About the DCC. Retrieved October 29, 2008, from http://www.dcc.ac.uk/about/

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. B. (2007, January). Understanding infrastructure: Dynamics, tensions, and design. *Deep Blue*. Retrieved August 8, 2008, from http://hdl.handle.net/2027.42/49353

Friedlander, A. (2008). The triple xelix: Cyberinfrastructure, scholarly communication, and trust. *Journal of Electronic Publishing*. Retrieved August 8, 2008, from http://hdl.handle.net/2027/spo.3336451.0011.109

Jacobs, N. (2007). *Report of a workshop on research and development priorities to support research data curation*. Retrieved August 8, 2008, from http://infteam.jiscinvolve.org/files/2008/05/datacurationwshop20071214.pdf

Johnson, R. (2002, November). Institutional repositories: Partnering with faculty to enhance scholarly communication. *D-Lib Magazine*. Retrieved August 8, 2008, from http://www.dlib.org/dlib/november02/johnson/11johnson.html

Lynch, C. (2003, February). Institutional repositories: Essential infrastructure for scholarship in the Digital Age. *ARL: A Bimonthly Report*, no. 226. Retrieved August 8, 2008, from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

Open Archives Initiative. (n.d.). Object reuse and exchange. Retrieved October 29, 2008, from http://www.openarchives.org/ore/

Sorokina, D., Gehrke, J., Warner, S., & Ginsparg, P. (2007, February). Plagiarism detection in arXiv. *arxiv.org*. Retrieved August 8, 2008, from http://arxiv.org/abs/cs.DB/0702012

Van de Sompel, H., Lagoze, C., Jeroen, B., Liu, X., Payette, S., & Warner, S. (2006). An interoperable fabric for scholarly value chains. *D-Lib Magazine*. Retrieved August 8, 2008, from http://dlib.org/dlib/october06/vandesompel/10vandesompel.html

---

G. Sayeed Choudhury serves as principal investigator for projects funded through the National Science Foundation, Institute of Museum and Library Services, and the Mellon Foundation. Choudhury has oversight for the digital library activities and services provided by the Sheridan Libraries at Johns Hopkins University. He has published numerous articles in the *International Journal of Digital Curation*, *D-Lib*, the *Journal of Digital Information* and *First Monday*. He has served on committees for the Digital Curation Conference, Open Repositories, Joint Conference on Digital Libraries, and Web-Wise, and has presented at various conferences including Educause, CNI, DLF, ALA, and ACRL.