IJDC | *General Article*

Open Data Meets Digital Curation: An Investigation of Practices and Needs

Christopher A. Lee
University of North Carolina

Suzie Allard
University of Tennessee

Nancy McGovern
Massachusetts Institute of Technology

Alice Bishop
Council on Library and Information Resources

Abstract

In the United States, research funded by the government produces a significant portion of data. US law mandates that these data should be freely available to the public through ‘public access’, which is defined as fully discoverable and usable by the public. The U.S. government executive branch supported the public access requirements by issuing an Executive Directive titled ‘Increasing Access to the Results of Federally Funded Scientific Research’ that required federal agencies with annual research and development expenditures of more than $100 million to create public access plans by 22 August 2013. The directive applied to 19 federal agencies, some with multiple divisions. Additional direction for this initiative was provided by the Executive Order ‘Making Open and Machine Readable the New Default for Government Information’ which was accompanied by a memorandum with specific guidelines for information management and instructions to find ways to reduce compliance costs through interagency cooperation.

In late 2013, the Institute of Museum and Library Services (IMLS) funded the Council on Library and Information Resources (CLIR) to conduct a project to help IMLS and its constituents understand the implications of the US federal public access mandate and how needs and gaps in digital curation can best be addressed. Our project has three research components: (1) a structured content analysis of federal agency plans supporting public access to data and publications, identifying both commonalities and differences among plans; (2) case studies (interviews and analysis of project deliverables) of seven projects previously funded by IMLS to identify lessons about skills, capabilities and institutional arrangements that can facilitate data curation activities; and (3) a gap analysis of continuing education and readiness assessment of the workforce. Research and cultural institutions urgently need to rethink the professional identities of those responsible for collecting, organizing, and preserving data for future use. This paper reports on a project to help inform further investments.

Received 19 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Christopher Lee, 216 Lenoir, Chapel Hill, NC, 27599, USA. Email: callee@ils.unc.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v11i2.403
DOI: 10.2218/ijdc.v11i2.403

Introduction

Research data is a valuable resource for a variety of stakeholders across all sectors of society. In the United States, there is a legal mandate for research funded by the federal government to be fully discoverable and usable by the the public which is referred to as 'public access'. An Executive Directive titled 'Increasing Access to the Results of Federally Funded Scientific Research' (22 Feb 2013), requires federal agencies with annual research and development expenditures of more than $100 million to create public access plans by 22 August 2013. The directive applies to nineteen federal agencies, some with multiple divisions. Additional direction for this initiative was provided by the Executive Order 'Making Open and Machine Readable the New Default for Government Information' (9 May 2013), which was accompanied by a memorandum (OMB m-13-13) with specific guidelines for information management and instructions to find ways to reduce compliance costs through interagency cooperation.

At the same time, there has been substantial activity – much of it within academia and the cultural heritage sector – to both define and support the competencies required for digital curation. Although some cultural heritage institutions had been actively caring for digital materials for many years, most of these institutions had relatively little experience with scientific data until quite recently. In the United States, the Institute of Museum and Library Services (IMLS) is a key player in the development of conceptual and professional approaches to digital curation. Building on the 21st Century Librarian program that it began in 2003, IMLS issued a 2006 call for grant proposals to develop educational programs in digital curation and has since funded numerous projects and programs in this space (Ray, 2009).

There has been a significant move toward providing public access to data that has been created with public sector funds, and there has also been considerable progress on the definition and development of professional capabilities to care for such data to ensure continuing access. However, neither one of these streams of activity has a single clearly defined professional home. Both are undertaken by individuals with a vast array of disciplinary backgrounds, job titles and institutional contexts.

In late 2013, IMLS funded the Council on Library and Information Resources (CLIR) to conduct a project to help IMLS and its constituents understand the implications of the US federal OA mandate and how needs and gaps in digital curation can best be addressed. This study is intended to help span the boundaries between the arenas of OA and digital curation.

Purpose and Background

Given the importance of the federal investment in public access to data and publications and in the significance of the content they comprise, the standards, practices, and guidelines emerging from the agency plans will have a notable impact on the standards, practices, and guidelines that libraries and cultural heritage organizations need to adopt. Meeting the technical, structural, and interpersonal demands required to manage digital research data is a shared responsibility of researchers, information professionals, cultural heritage professionals, academic and cultural heritage institutions, and publishers.

There are a number of assumptions embedded in the digital curation efforts in the cultural heritage community over the past couple decades that have informed and influenced the nature of digital curation programs and services:

- **Libraries have been doing data curation a long time:** There have been data archives and data librarians in the social sciences and in other fields for decades in the US and elsewhere. The perception that libraries have been doing data curation for a long time seems to result from the presence of data librarians in some libraries and services some libraries provided to assist users by providing secondary data and support for statistical analysis. In practice, data curation is more active and comprehensive than the services most libraries provided prior to the mid-2000s, though there were some standout programs.
- **The digital curation services to provide and the skills needed are already known:** Services and programs are typically characterized as being responsive to the needs and requirements of users. As most research domains and disciplines are still being transformed by digital practices and protocols and will continue to evolve, the possibilities of services to provide could not yet be fully realized and the digital curation skills needed to build programs and provide services must continue to evolve in response.
- **It is essential to understand and address domain-specific curation practices:** Monitoring emergent practices across the range of research disciplines is important, and it will also be beneficial to look for commonalities (often a strength of information professionals) to also develop common services to serve domains.

Fixating too heavily on these assumptions may be slowing the progress of the cultural heritage community in developing truly innovative services and programs in collaboration with and in the service of researchers in the changing landscape of domains that interdisciplinary research practices, by their nature, encourage and enable. Research and cultural institutions urgently need to rethink the professional identities of those responsible for collecting, organizing, and preserving data for future use. While the primary focus of our study has been on US government agencies and the implications for IMLS programs and constituencies, we believe the findings are relevant more broadly.

Study Design

Our project has three research components: (1) a structured content analysis of federal agency plans supporting public access to data and publications; (2) case studies (interviews and analysis of project deliverables) of seven projects previously funded by IMLS to identify lessons about skills, capabilities and institutional arrangements that can facilitate data curation activities; and (3) analysis of continuing education and readiness assessment of the workforce.

Content Analysis

We conducted a content analysis of the open data and public access plans created by the federal agencies. When the U.S. government was shut down by the budget sequester through early 2014, the process originally outlined by OSTP for the submission of the public access plans, was delayed. As of December 6, 2015, 16 agencies (or at least one of their operating units) have made their public access plans available to the public. We analysed 21 plans from these 16 agencies. Based on the Executive Directive, the plans focus on two separate but related components: access to research data and access to the products of analysis based on these data in the form of peer-reviewed articles. Our study focused primarily on the former. We reviewed the federal documents to determine the appropriate level of analysis to be used and concentrated on aligning similar concepts. We assessed the plans for their similarities and differences, which led to identifying important themes for understanding the emerging government data management environment.

Case Studies of IMLS Digital Curation Projects

We identified seven recent (2010-2013) IMLS-funded projects that included significant digital curation objectives, which could include management, preservation or provision of access to digital information. The sampling frame aimed for diversity of project objectives, curation functions and data types. This investigation – focusing on the experiences of professionals who have engaged in digital curation work – complements the content analysis discussed above, which focused on the aspirations of government agencies based on the text of their plans. Our investigation was based on multiple data sources. The primary data source was a set of semi-structured interviews with key project personnel. We conducted one interview per project for a total of seven interviews. Six of the interviews were conducted with a single individual (usually the project's principal investigator), but one interview involved two individuals. All interviews were recorded and transcribed.[1] They lasted between 20 and 49 minutes (average of 37 minutes). Table 1 summarizes the seven projects. In addition to the interviews, we analysed project documentation and (when applicable) online products of the projects.

[1] We would like to thank Fawnie Erickson for transcribing all of the interview recordings.

Table 1. IMLS-funded projects investigated.

Project	Primary Focus
Creating a Better World by Sharing Research Online	Institutional repository (IR) to provide access to the university's research output
Databib	Annotated online bibliography of research data repositories
Datastar	Study researchers' data sharing and discovery needs and enhance a linked data platform to meet those needs
ETD [Electronic Theses and Dissertations] Lifecycle Management	Guidance documents and software tools for life-cycle data management and preservation of ETDs
Improving Data Stewardship with the DMPTool	Identify and propose strategies to address challenges to adopting the Data Management Planning Tool (DMPTool)
Virtual Archiving for Public Opinion Polls	Demonstrate and promote streamlined workflows for getting research data into data archives
What's on the Menu? – From Software to Funware	Support crowdsourcing of menu transcriptions

Analysis of Capacity Building: Curriculum, Competencies, and Careers

Our approach for this part of the study was to review digital curation and related curriculum projects, identify strengths and gaps in relevant continuing education programs, and assess the readiness in the current professional workforce.

A number of projects have focused on defining digital curation skills and the curricula to develop those skills. For example, we identified 24 projects between 2004-2015 funded by IMLS that have addressed digital curation curriculum and skills development, costing approximately $14 million in total. This extensive investment has resulted in a significant set of resources – including e-certificate programs, workshops, and online resources and tutorials – aimed at developing digital curation skills and competencies. Our analysis focused primarily on the results of four specific competency-based projects within the digital curation and preservation community, three of which were developed in the US: the Digital Curation Curriculum (DigCCurr) Matrix from the University of North Carolina at Chapel Hill, the Digital Curator Vocational (DigCurV) Education Europe Project Curriculum Framework, the Preparing the Workforce for Digital Curation report's distinct and essential knowledge and skill areas, and the National Digital Stewardship Alliance (NDSA) Staffing for Effective Digital Preservation report's skills survey instrument.

We also investigated job postings as a further source of data on competencies for digital curation. One of the primary challenges is to define the proper scope for selection of the postings. Digital curation is an arena that is still undergoing rapid change and spans the boundaries of a variety of existing professions and types of institutions. Any decision to select particular job posting venues or search terms is loaded with potential biases. With full recognition of these considerations and the implications for the limitations of our findings, we analyzed 120 job postings from the DigiPres mailing list.[2]

Findings

This paper provides a high-level summary of findings from the three areas of our study: content analysis, case studies of IMLS digital curation projects, and gap analysis of continuing education and readiness assessment. The full findings are presented in the project's final report (Allard et al., 2016).

Content Analysis

The review of the government documents generated 12 high-level findings grouped in three general areas: open data infrastructure, roles and responsibilities, and making data public.

Open Data Infrastructure

In the broader community discussion of research data, the question often arises, "What do we mean by data?" These plans suggest that for scientific data the answer to this question has been adequately defined in OMB Circular A-110 so that it can be used by myriad agencies holding diverse and heterogeneous data.

Setting up a coordinated framework that works across agencies is hard even when there are specific directives. While structure is important to provide a framework, it takes strong interagency connections to make it work. Ultimately, success is likely to rely on effective interpersonal communication and vibrant community engagement. Flexibility for agencies so they can best serve their communities is a considerable strength for responsible data management and for adoption of the required behaviors by researchers. However, there is a need to balance this with the ability to cross-reference on-going activities to facilitate collaboration between government agencies, and non-governmental partners.

The public access documents suggest that agencies see the data generated by their researchers as part of a larger corpus. In the framework, collaboration was identified as an important component for the future. Cooperation among agencies is discussed and one platform – PubMed Central, which was developed by the National Library of Medicine as a repository for publications – has emerged as a significant point of collaboration. A proposed 'Research Data Commons' would provide tools to facilitate the discovery, access and use of data from across multiple agencies. One agency notes that the Commons would operate on the FAIR principle – Find, Access, Interoperate, Re-use. The challenges surrounding interoperability both technically and socio-culturally, and the issues surrounding reuse including data citation make this a difficult proposition, but it is a promising part of the formal discussion.

[2] DigiPres Mailing List: http://lists.ala.org/sympa/info/digipres

Roles and Responsibilities

Although 14 of the 16 agencies have libraries, archives or information centers, the role of the agency library or data center was only expressly stated in the public access plans for six of the 16 agencies. Many agencies' libraries and data centers may already be engaged in public access activities, but their role is not explicitly recognized.

The role of education is specifically noted in the framework but only a limited number of agencies have explicit plans and six do not address it at all. The agencies that discuss education approach it either in terms of (1) compliance focusing on educating the agency employee so that the policy can be efficiently and correctly implemented; or (2) outreach for researcher training and eventually as a means of moving science forward. Adopting best practices in data management behavior requires education of both agency employees and data creators. These findings suggest this is an area that may need more attention.

Many plans fail to address in a meaningful way the cost of creating and maintaining open data and how this may be recovered. While the cost of data management is an essential consideration in designing an open data plan that is sustainable, six of the 16 plans do not address cost or only briefly mention the need to consider the monetary and administrative burden. Only five agencies suggest that researchers could or should include a budget item for the cost of data management.

Taking the Data Public

Even though making research results publicly accessible is a key point in the OSTP mandate, there is a range in how agencies address the topic. Most simply meet the mandate by noting that peer-reviewed articles will be freely available in a repository no more than 12 months after publication, and that data supporting that publication would also be made available in 12-39 months. However, few agencies move beyond a discussion of simple discoverability and accessibility and address the need to build an environment to interact meaningfully with data.

Metadata are an essential element. Thirteen plans note the importance of metadata for discovery and access and outline plans in this area ranging from general to quite specific. The following are recurring themes regarding metadata: the data management plan must identify standards used for the metadata, the data set must have a formal metadata document, the metadata for the data set must include the common core from the schema used by the Federal government[3], and there must be metadata supplied for publications. For agencies that have a broader research spectrum, the reference is to having metadata meet appropriate industry standards. Some plans call for developing modules or services to manage metadata generation, acquisition and quality control.

Case Studies of IMLS Digital Curation Projects

This part of our study generated nine high-level findings:

1. **Successful initiatives are part of ongoing capacity building activities:** Many of the successful projects built upon lessons and capabilities that were established in previous activities, including previously-funded projects. Interview participants often found it difficult to speak exclusively of the work they had done on the specific IMLS-funded project in question, because it was

[3] See the Project Open Data policy at: https://project-open-data.cio.gov/

often so closely tied to work they had done in earlier projects. In turn, the IMLS-funded projects investigated in our current study have themselves often provided an important foundation for future work.

2. **Digital curation requires control over software:** Managing and providing access to digital data requires a variety of software elements. Professionals responsible for digital curation must establish proper control over that software. This can involve development of new software, customization of existing code, using existing tools as they are, and various aspects of configuration management. Building upon existing software (open-source tools, commercial tools and online services) was a major theme from the interviews. Regardless of what combination of software is used, setup and integration can often be quite resource intensive. In some cases, existing tools and systems served as important models and sources of ideas, even if they were not incorporated directly into the project's own software products.
3. **Effective digital curation involves not only working with data but also active engagement with relevant stakeholders:** Leaders of the projects under investigation had a strong sense of who their primary stakeholders were and made concerted efforts to engage with them. The primary stakeholders that they expressed by the end of the project were not always the same ones that they started with. Building an effective system requires not just technical development, but also marketing and outreach. Projects not only engaged with relevant stakeholders themselves but also generated resources that professionals can use to support their own engagement activities. A particular type of stakeholders are those involved with allied projects and initiatives.
4. **Making the case to resource allocators is a key factor in many settings:** One of the key categories of stakeholders in most digital curation initiatives is institutional leaders who make resource allocation decisions. However, it is also important to recognize the essential role of line staff to carry out the work.
5. **Releasing early prototypes can be beneficial, in order to test with real data:** As discussed above, there are various forms of stakeholder engagement that can be essential to the success of digital curation efforts. One particularly valuable form of engagement is to have potential users interact with the intended deliverables, whether those are systems, applications or documents. Self-reported needs (e.g. those elicited from surveys, interviews or focus groups) can be revealing, but they are not always accurate representations of user behaviors. Early prototyping and testing can help to ensure that development is moving in a direction that is likely to benefit users.
6. **Meeting user needs involves many inferences about their behaviors and expectations:** As discussed above, analyzing user needs often involves mechanisms such as user testing, interviews, surveys and focus groups. Such methods can be very valuable ways to test assumptions, identify design priorities and identify opportunities for improvement. However, even with such resources at hand, it is rarely possible to directly elicit data on all aspects of system or process. One way to make inferences about user needs is to rely on information professionals who can serve as proxies for users, based on their experiences in working with given populations. One way in which information professionals with knowledge of information practices within a domain can serve as proxies

for users is by providing 'reality checks' on what sorts of actions users would likely be willing to engage in. This comes up frequently in terms of how much and what types of metadata users will generate, as well as what types of documentation they would be willing to read.

7. **Metadata satisficing[4] is essential:** There is significant value in defining clear metadata conventions (e.g. schemas, ontologies, data dictionaries), and this is something that information professionals are very well positioned to do. Metadata enhancement, clean-up and transformation can require substantial resources. No project or institution has unlimited resources, so it can be important to maintain the flexibility to accommodate metadata that does not fully conform to the ideal. Digital curation professionals must make numerous decisions about metadata trade-offs. One fundamental choice is between the following three options: (1) insist that those submitting data to their systems conform to strict metadata conventions when they submit the data, (2) accept 'sloppy' metadata but then engage in substantial clean-up activities in order to ensure that the metadata ultimately conform to strict metadata conventions, or (3) establish metadata conventions that are more flexible and tolerant of variance within the values. Participants in our study conveyed approaches that involved various combinations of all three options. A common strategy is to identify a relatively limited, core set of metadata elements that can then be extended in particular cases. It is important to determine not only what metadata should be captured/created but also what subset should be exposed to users.
8. **Public access involves not just enabling discovery of data but also enabling new forms of interaction with and among users:** The access provision duties of digital curation are not exhausted by putting the data and associated metadata on the Web (no matter how good the metadata might be). Effective data use can involve a variety of interaction mechanisms. In addition to allow search and navigation through an institution's web site, interview participants cited mechanisms including RSS feeds, Twitter (which "drives traffic to the record"), and Google spreadsheets populated with data. One potential form of user interaction is the generation of additional metadata and documentation. Several of the interview participants also pointed out the potential for facilitating further interaction between users.
9. **There is value to pushing into Producer practices and behaviors:** An essential aspect of digital curation that relates to many of the above findings is interjecting digital curation knowledge and methods into the information lifecycle as early as possible.

Analysis of Continuing Education and Readiness Assessment of Workforce

The importance of continuing education in advancing digital curation within the cultural heritage community and for researchers in the growing range of disciplines that are engaged in digital curation activities is evidenced by the significant number of

[4] Satisficing is a term introduced by Herbert Simon in the 1950s to characterize a decision-making process that involves settling on an option that is 'good enough' to meet a certain threshold of acceptability (called an 'aspiration level'), rather than attempting to find a single optimal solution to a problem. It applies particularly well to decisions about metadata, because it is impossible to predict precisely which metadata elements will be most valuable in the future, but it is possible to make educated guesses about the types of metadata that are likely to be valuable.

community projects and reports that highlight the need for education and to make progress in competency building, curriculum development, and support for lifelong learning. The following is a summary of high-level observations from our gap analysis:

- **Curriculum development and programs:** The commitment to developing training programs and building competencies is evidenced by the funded projects that have resulted in some progress in developing continuing education and academic training programs.
- **Competencies:** There has been great interest in and focus on defining requisite skills and competencies for digital curation that has resulted in a set of proposed frameworks for defining and developing skills.
- **Job postings and titles:** The range of job postings and titles in areas relating to digital curation and preservation reflect the evolution of the skills and roles involved.

Though there has been a significant investment and interest in continuing education, curriculum development, and competency building for digital curation and preservation, the community's resources do not yet include sufficient data, qualitative or quantitative, to monitor, analyze, or assess the impact of the programs.

Conclusions and Implications

In all aspects of cultural heritage – including the curation of scientific data – libraries, librarians, and information professionals have an important role to play. We see great potential for further collaboration and integration of efforts. Professionals engaged in public access initiatives (most often conducted in government agencies) have much to gain from learning about the work (most often conducted within academic institutions) in developing and implementing data management plans. Similarly, experience in public access initiatives can help to inform data management plans, so that their provision for access are most likely to be viable and sustainable. There is great potential for strategic connections between government public access efforts and digital curation work underway in cultural institutions. With the existing competency models in place as a foundation for understanding and building digital curation competencies, future work on competencies should at a minimum use the available foundation as a starting point. It would be valuable to develop projects and initiatives to collaborate with researchers in relevant domains and projects on the development of competencies.

Acknowledgements

We would like to acknowledge support for carrying out and reporting on this work provided by the Institute of Museum and Library Services (IMLS) and the Council on Library and Information Resources (CLIR).

References

Allard, A., Lee, C., McGovern, N., & Bishop, A. (2016). The open data imperative: How the cultural heritage community can address the federal mandate. Retrieved from https://www.clir.org/pubs/reports/pub171/pub171

Burwell, S., et al. (2013). *Open data policy: Managing information as an asset.* Memorandum for the Heads of Executive Departments and Agencies. Executive Office of the President, Office of Management and Budget, May 9. Retrieved from https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Holdren, J. (2013). *Increasing access to the results of federally funded scientific research.* Office of Science and Technology Policy, Memorandum for the Head of Executive Departments and Agencies, February 22. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

National Science Foundation, Blue-Ribbon Advisory Panel on Cyberinfrastructure. (2003). Revolutionizing science and engineering through cyberinfrastructure. Retrieved from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203

Office of Management and Budget. (1999). Circular A-110, to the heads of executive departments and establishments: Uniform administrative requirements for grants and agreements with institutions of higher education, hospitals, and other non-profit organizations. Retrieved from https://www.whitehouse.gov/omb/circulars_a110

Ray, J. (2009). Sharks, digital curation, and the education of information professionals. *Museum Management and Curatorship 24*(4). doi:10.1080/09647770903314720

White House, Office of the Press Secretary. (2013). Executive order: Making open and machine readable the new default for government information, May 9. Retrieved from https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Whitmire, A., Briney, K., Nurnberger, A., Henderson, M., Atwood, T., Janz, M., Kozlowski, W., Lake, S., Vandegrift, M., & Zilinski, L. (2015). A table summarizing the federal public access policies resulting from the US Office of Science and Technology policy memorandum of February 2013. Retrieved from https://figshare.com/articles/A_table_summarizing_the_Federal_public_access_policies_resulting_from_the_US_Office_of_Science_and_Technology_Policy_memorandum_of_February_2013/1372041