Postdoctoral Fellowships in Data Curation for the Sciences and Social Sciences:

Early Experiences and Contexts

Alice Bishop, Christa Williford
Lori Jahnke, Andrew Asher
Jodi Reeves Eyre

April 2019

Council on Library and Information Resources
Arlington, VA

Contents

Scientists and Social Scientists Working in Data Curation:
The Inaugural CLIR/DLF Cohorts

Alice Bishop and Christa Williford

In 2015, the U.S. National Research Council (NRC) released *Preparing the Workforce for Digital Curation*, a report examining the critical need for education and training in digital curation to address the increasing demand for access to and meaningful reuse of digital information. Prepared by the Board on Research Data and Information (BRDI), the report states, "There is no single occupational category for digital curators and no precise mapping between the knowledge and skills needed for digital curation and existing professions, careers, or job titles." One of its conclusions is that the knowledge and skills required of those engaged in digital curation are highly dynamic and highly interdisciplinary. They encompass an integrated understanding of computing and information science, librarianship, archival practice, and the disciplines and domains generating and using data. The report cites the increase in the number of domain experts learning digital curation as an important development and states, "One promising model for transitioning domain experts into curation work has been developed by the Council on Library and Information Resources (CLIR)/Digital Library Federation (DLF) as an extension of their Postdoctoral Fellowship Program" (71).

Since 2004 CLIR's Postdoctoral Fellowship Program has recruited, trained, and established cohorts of new PhDs working within the digital environment to help manage, sustain, and generate valuable information in support of higher education. Beginning in 2012 with support from the Alfred P. Sloan Foundation, the program deepened and sharpened its focus on research data curation, a focus that persists to this day. Between 2012 and 2014, CLIR and DLF collaborated to bring 23 scientists and social scientists into data curation fellowships at 20 host institutions in the United States and Canada. The goal was for these fellows to contribute to a more sophisticated understanding of data curation and its often-determining role in the conduct of scientific and social scientific research. What impact did these initial cohorts have? What kinds of ongoing needs and challenges do the fellowships reveal? And what are the implications of their experiences and potential for the stewardship of scholarly research data?

As a follow up to their 2012 CLIR report *The Problem of Data*, anthropologists Lori Jahnke and Andrew Asher interviewed data curation fellows and their colleagues to assess the impact of the initial three cohorts supported by the Sloan Foundation. In "Ongoing Challenges for Data Curation Support: A Program Assessment of the 2012–2014 CLIR/DLF Postdoctoral Fellowships in Data Curation for the Sciences and Social

Sciences," they conclude that, "considerable work is still needed to develop a system of data curation that supports the preservation and access of research data while allowing researchers to fulfill their ethical and professional obligations, and an improved understanding of research processes and their variance among disciplines is critical to this work." Clarifying tools, standards, and workflows for research data curation practice is only part of this work: Jahnke and Asher point out that shifts in academic cultures are also necessary to recognize and reward the contributions of curators who create useful and reusable datasets that are instrumental to the advancement of science. Through creating the CLIR/DLF fellowships in data curation, CLIR aspired to trigger new ways of thinking about divisions of labor and expertise across the academic professions while bringing focused attention to the development of data curation resources and services that would be responsive to the needs of current researchers.

Fellows' Disciplines

With the addition of the data curation fellowships in 2012, the proportion of scientists and social scientists in CLIR's Postdoctoral Fellowship Program increased significantly. When the program was founded, fellows came exclusively from training programs in the humanities or, occasionally, from information science. The first social scientists from other fields joined the program in 2009; in 2010, the program welcomed the first fellows trained in natural sciences.

To recruit fellows for the new data curation fellowships, current and former postdoctoral fellows helped solicit candidates through their own networks and communities, while CLIR staff reached out to programs and departments known for data-intensive research to inquire about suitable PhD candidates in the sciences and social sciences. Staff worked closely with host institutions throughout the recruitment and hiring processes.

The candidates hired for the 2012-2014 CLIR/DLF Postdoctoral Fellowships in Data Curation for the Sciences and Social Sciences came from the following 16 disciplines:

Anthropology	2
Archaeology	2
Biology/Biological Sciences	2
Biomedical Informatics	1
Chemistry	1
Comparative Literature	1
Economics	1
Educational Research and Policy Analysis	1
Environmental Studies	1

Geological Sciences	1
History of Consciousness	1
Information Science	4
Neuroscience	1
Political Science	1
Psychology	1
Sociology	2

The high number of candidates in information science suggests a greater awareness of the need for new interdisciplinary work in data curation among iSchool PhDs than among those in the sciences and social sciences. In many disciplines with a strong tradition of postdoctoral research and training, such as in laboratory-based sciences, pursuing a library-based fellowship following the degree was, and still is, unusual. As a result, CLIR continues to invest time and energy in recruiting more disciplinary perspectives into the fellowships, since it is the fellows' academic grounding that helps the program facilitate meaningful, productive connections between scholarly and professional communities of practice.

The attraction of the program for talented but diversely trained young scholars has remained strong since the CLIR/DLF data curation fellowships were created. Of the 15 institutions that hosted fellows through CLIR's program for the first time during the first three years of Sloan Foundation support, 8 have since chosen to host a second fellow, and 4 of these have hosted a third. These numbers suggest that the variety of perspectives the program incorporates is an asset for the ongoing development of research data collections and services.

The value of disciplinary experience for success in a professional role in research data curation can be unpredictable. While hosts identify the specific qualifications they seek in their position descriptions, CLIR and DLF have found that when reviewing the applicant pool, hosts sometimes hire fellows with disciplinary backgrounds that are different from those originally anticipated. That was the case in 2014 when one host institution chose a fellow with a comparative literature degree. The university had to obtain an exemption from its human resources administrative office before hiring the fellow since the original fellowship description had called for a PhD in the social sciences. Regardless of their degrees, many candidates in CLIR's applicant pool stand out not just for their strong disciplinary and methodological skill sets but also for their contextual understanding and desire to make data useful and reusable across disciplines and institutional boundaries.

Fellows' Career Outcomes

Regardless of how fellows come to the program, they generally report that they gained valuable experience that helped them build careers in both academic libraries and on the tenure track. Of the 23 data curation fellows brought into CLIR/DLF fellowships through the Sloan Foundation's support between 2012 and 2014, more than half now hold permanent positions in academic libraries, either at their original host institutions or elsewhere. The majority of the remaining former fellows from these first three cohorts now hold positions in the field of data management and are engaged in closely related work in research and research support, software creation, assessment, or user experience design.

2012-2014 Sloan-Supported Fellows' Career Outcomes (as of April 2019)

Library	56% (13/23)
Tenure track faculty	4% (1/23)
Other non-academic jobs	13% (4/23)
Other academic jobs	13% (3/23)
Other postdoctoral fellowship	4% (1/23)
Unknown	4% (1/23)

The number of these 23 fellows who were hired by their host institutions immediately following the fellowship into permanent roles (39%) is healthy and the regular submission of new job descriptions to CLIR with requests to post specifically to the postdoctoral fellowship community indicates that these fellows share skills and perspectives that are in high demand. However, given the relatively small sample and the uniqueness of the circumstances affecting each fellow's and each host's choices, it is difficult to draw conclusions about which factors are most closely associated with the conversion of data curation fellowships into longer term careers in data curation. CLIR's Postdoctoral Fellowship Program remains a small part of a much larger narrative involving many other organizations, funders, institutions, and individuals. It is a persistent challenge to distinguish between individual fellows' influence on changes in research and research support practices, and the influence of other, larger networks of professionals to which fellows have become connected.

The Larger Data Job Context

As a way to contextualize the earliest CLIR/DLF data curation fellows in the sciences and social sciences within the broader academic library job market, CLIR approached research consultant and current CLIR Program Officer Jodi Reeves Eyre—who was

herself among the 2013–2015 fellows—to undertake an analysis of the responsibilities and qualifications listed in data job descriptions posted by academic libraries between 2013 and 2017, the years during which CLIR's earliest data curation fellows entered the job market. In her paper, she discusses the ways that the qualifications and perspectives of recent PhDs in the sciences and social sciences may or may not suit the positions as they were described. Her findings indicate that if research and subject knowledge, experience, and skill requirements are the only factors taken into account, then recent PhDs in the sciences and social sciences are strong candidates for data jobs. However, many data jobs listed during this period required a library degree, library experience, or both; these stipulations can pose insurmountable barriers for recent PhDs interested in research data curation. Reeves Eyre's research also indicates that emphasizing experience in a service or consulting role in data job listings poses additional challenges for recent PhD scientists and social scientists who are more likely to have worked collaboratively in laboratories, in the field, or in the classroom teaching.

What's Next?

Sustaining the Postdoctoral Fellowship Program long-term has never been a major objective for CLIR. The program was designed to help academic library leaders envision a sustainable future, expanding the pool of talent available to replace those leaders in the coming decades. CLIR intends to remain involved in the program only as long as it genuinely serves these purposes and CLIR's mission.

Expanding the range of roles and disciplines in recent years has dramatically increased the size of the program, as well as the breadth of stakeholders who provide support to the program and receive its benefits. There is no sign of waning interest, and for the near term CLIR expects to attract healthy numbers of host institutions and candidates. By continuing to cultivate cohorts of differently trained yet like-minded researchers who work together to explore the challenges that the collection, organization, and preservation of data pose for the future of scientific research, CLIR aspires to help build more robust shared information systems, inculcate more deeply interdependent relationships among institutions of higher learning, and develop a more flexible and multifunctional human infrastructure to maintain and strengthen those relationships in the future.

Ongoing Challenges for Data Curation Support:

A Program Assessment of the Early CLIR/DLF Postdoctoral Fellowships in Data Curation for the Sciences and Social Sciences, 2012–2016

Lori M. Jahnke, Emory University
ljahnke@emory.edu

Andrew Asher, Indiana University Bloomington
asherand@indiana.edu

Introduction

In 2011, we undertook a study of data curation practices among scholars at several higher education institutions. In this study, we set out to document workflows related to the creation, management, and preservation of research data, with the ancillary goal of identifying unmet researcher needs. That study was published in the CLIR report *The Problem of Data* (Jahnke et al. 2012). Following this initial phase of research, CLIR implemented the CLIR/DLF Postdoctoral Fellowship Program in Data Curation for the Sciences and Social Sciences. Both the research and the program were supported by the Alfred P. Sloan Foundation.

This report is a follow up to our 2011–2012 study. Here we focus on the 2012, 2013, and 2014 cohorts of CLIR/DLF postdoctoral fellows, their work, and the host institutions. Following a qualitative interviewing methodology similar to the one we employed for *The Problem of Data*, we interviewed fellows, their coworkers and team members, supervisors, and researchers, and in some cases we visited host institutions. Employing this contextual approach to both studies enabled us not only to assess the impact of the initial cohorts of the CLIR/DLF postdoctoral program, but also to further characterize ongoing needs and challenges for data curation in the university environment and to compare the experience of fellows and the fellowship community with that of the researchers whose practices and workflows had helped shape the CLIR/DLF program. In this second study, we set out to answer the following questions:

1. In what types of projects and research teams have CLIR/DLF fellows participated during their fellowships? What effect have the fellows had on the data practices of these research teams?

2. What skills, training, or other factors have contributed to the perceived success or failure of individual CLIR/DLF fellows and their collaborative efforts in the domain of research data curation? What roles do institutional contexts play in these outcomes?
3. What are the component processes and workflows in the creation, management, and preservation of scholarly research data?
4. How do these processes and workflows compare with those observed in the 2011–2012 study, *The Problem of Data*?
5. What are the unmet researcher needs within these processes?

Our approach included a series of semi-structured interviews with more than 30 individuals from 13 of the host institutions, as well as site visits and information gleaned from exit interviews conducted with each of the fellows. Upon completion of their fellowship, either a CLIR staff member or an external consultant conducted an exit interview with each fellow to gather information about work undertaken during the fellowship with particular attention to the challenges encountered. We (the authors) developed the exit interview protocol in collaboration with members of the CLIR staff to ensure that the data obtained would be compatible with this study, as well as with the needs of long-term program assessment.

Researcher Priorities and Research Data Curation

Although the open data movement has evolved rapidly since our 2011–2012 study (Allard et al. 2016), researchers, institutions, and fellows continue to confront many of the same challenges. In our initial study, several research faculty spoke candidly of difficulties they faced in trying to balance the demands of high expectations for publication output while meeting their professional and ethical obligations of data stewardship and intellectual transparency. In this environment, producing publications often took priority over other activities. The decision to prioritize publications is unsurprising given the influence a researcher's publication record has on a number of career- and life-altering decisions such as hiring, promotion, funding, tenure, and salary (Nosek et al. 2010; Ostriker et al. 2009).

Our 2011–2012 study participants are not alone in expressing concern over the ever-increasing expectations for research productivity and the detrimental effects of this pressure on the quality of research (Fischer, Ritchie and Hanspach 2012a; Ioannidis 2014; Nosek, Spies and Motyl 2012; Sarewitz 2016; Sills 2016; Smaldino and McElreath 2016).

> Scrupulous research on difficult problems may require years of intense work before yielding coherent, publishable results. If shallower work generating more publications is favored, then researchers interested in pursuing complex

> questions may find themselves without jobs, perhaps to the detriment of the scientific community more broadly. (Smaldino and McElreath 2016:29)

The trend of increasing competition for jobs[1], funding, and other resources in the research sector may have greater consequences than a poor employment outlook for young researchers. There is a growing concern that the enormous pressure to produce consistently novel and positive findings[2] promotes bias in research design and statistical analysis that increases the chances of false-positives (Eisner 2018; Simmons, Nelson and Simonsohn 2011). Among scientific publications, multiple authors have commented on the increasing rate of article retractions (Brembs, Button and Munafo 2013; Cokol et al. 2007; Fang and Casadevall 2011), and some authors have found further evidence that the rate of retractions as a result of misconduct, rather than error, is increasing even more rapidly (Fang, Steen and Casadevall 2012; Steen 2011). The likelihood that an article will be retracted is also strongly correlated with the journal impact factor (Brembs, Button and Munafo 2013; Cokol et al. 2007; Fang and Casadevall 2011; Fang, Steen and Casadevall 2012). Although Cokol et al. (2007) attribute this correlation to the greater visibility enjoyed by high-impact journals, Brembs, Button and Munafo (2013:5) find that the relationship cannot be attributed to visibility alone, but it is also the result of intense social pressure to publish in the best-known, most frequently cited, and highest-impact factor journals, which increases unreliability of the submissions.

In a recent analysis covering 60 years of publications in the behavioral sciences, Smaldino and McElreath (2016) showed how the emphasis on high publication output inevitably leads to poorer methods and increasing rates of false-positives. In their analysis, the methodological deterioration does not require any strategizing or conscious misbehavior on the part of individuals or labs. It is simply that quantity of publication is so strongly rewarded that the penalty of failed replication studies and retractions is not sufficient to overcome this benefit (see also Eisner 2018). In other words, the volume of publications and their appearance in high-impact journals is so highly incentivized that researchers gravitate, consciously or unconsciously, toward strategies that support high volume rather than quality. While replication studies do help identify false positives and slow the progress of methodological degradation, Smaldino and McElreath (2016:26–28) note that all studies would need to be replicated multiple times in order to detect researchers that cut corners, a scenario that is unrealistic.

[1] In an analysis of 56 “junior” evolutionary biologists, Brischoux and Angelier (2015) found that to get hired into tenure track positions young biologists needed to have published twice as many articles and have approximately three years more experience in 2013 than in 2005.

[2] Publication bias, also referred to as the “file drawer problem,” is a well-known phenomenon in which surprising or novel results are more likely to be published than studies showing no effect. For discussion and examples of this phenomenon, see Fanelli (2010); Lane et al. (2016); Schooler (2011); van Dongen (2011); Young, Ioannidis and Al-Ubaydli (2008).

In the context of data curation, we must ask the bigger question of how open data fits into aligning the incentive and reward structures of scholarship with the values of scrupulous research. The final research publication is only the narrowest representation of the scholarship, and in most cases the publication does not include the methodological detail needed to adequately evaluate the work[3]. While the benefits of publishing research data have been widely discussed (e.g.,Voytek 2016), these arguments rarely account for the costs to the individual researcher in terms of the time spent on activities that are unrewarded professionally. Slowing the publication cycle may be an additional unrealized benefit of data curation for improving the health of the current research system (Fischer, Ritchie and Hanspach 2012a, 2012b; Halme, Komonen and Huitu 2012; Ioannidis 2014), but the incentive structure must also change to reward researchers for their efforts in promoting intellectual transparency. In other words, change must also come from the administrative, institutional, and national policy levels (Fischer, Ritchie and Hanspach 2012b; Halme, Komonen and Huitu 2012; Ioannidis 2014).

While researchers have more options than ever for depositing and publishing their data, such as data repositories[4], data journals[5], and several hybrid journals that accept and encourage the submission of datasets along with the research article (e.g., Nature, Wiley journals, PLoS), we have barely scratched the surface of the deeper issues related to supporting intellectual transparency in the research process and building a durable record of knowledge. In our 2012 report we highlighted the opportunities for fostering intellectual transparency along with the challenges of balancing access with privacy and confidentiality issues and supporting the ethical responsibilities of researchers (Jahnke et al. 2012:5). Although there has been more dialogue around these issues (e.g., RDA/NISO Privacy Implications of Research Data Sets IG[6]), this is one area where there is still much work to be done.

Summary of the Fellows' Projects

The 2012–2014 cohorts consisted of 23 fellows who were hosted at 20 institutions. We conducted interviews with individuals from 11 of these institutions and did site visits at

[3] Donoho et al. (2009:9) and Stodden (2011) paraphrase Stanford professor Jon Claerbout in describing the research article as merely the advertisement of scholarship rather than the work itself.
[4] A few field general options include Dataverse Network Project http://dataverse.org, Dryad https://datadryad.org, Qualitative Data Repository https://qdr.syr.edu, and Figshare https://figshare.com. In the social sciences, options include Archaeology Data Service http://archaeologydataservice.ac.uk, OpenICPSR https://www.openicpsr.org/openicpsr, Open Science Framework https://osf.io, and UK Data Service https://www.ukdataservice.ac.uk. For a list of recommended repositories, see PLoS http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories, or refer to the Registry of Research Data Repositories for an extensive list http://www.re3data.org.
[5] Candela et al. (2015) discuss the growth of operational data journals in recent decades.
[6] https://www.rd-alliance.org/groups/rdaniso-privacy-implications-research-data-sets-wg.html

two additional institutions from the 2015 cohort. According to information gathered during their exit interviews, fellows engaged in approximately 30 projects across the host institutions, ranging from assessment and planning activities to education and training to infrastructure development and implementation (Appendix A).

Most fellows participated in multiple projects simultaneously, and a few individuals worked on as many as four projects during their tenure. In some cases, this reflects the fellow's diverse interests, but in other instances it also reflects the institutional environment and the level of experimentation surrounding data curation and related services. To understand the nature of the fellows' work, we classified the projects into six types according to the primary activities or goals described during the exit interview: assessment and planning, education and training, software/tools development, infrastructure, collections, and outreach (table 1). Many projects included elements from several project types so these categories are not mutually exclusive.

Most of the projects fit into the assessment and planning category. These projects typically focused on needs assessment, requirements gathering, and other planning activities rather than program assessment. Projects in this area included grant writing, research around various aspects of data curation (e.g., infrastructure, publishing, restricted data services), compiling user profiles, and assessing needs at the university and disciplinary level. The abundance of assessment and planning activity is perhaps another reflection of the very early stage of development around data curation services and infrastructure at many of the host institutions. Implementing data curation support requires understanding one's institutional environment in new ways and many institutions are still undergoing a process of self-discovery.

Education and training projects included activities such as developing curricula and educational frameworks, managing working groups, providing consultation services, and developing peer mentoring or networking groups. Much less common were projects focusing on infrastructure, collections, or software development, although there were a few examples of each. Interestingly, fellows rarely described their work as outreach explicitly, but a key activity of nearly all projects included interacting with previously isolated departments, divisions, or programs. This was also an activity that nearly all study participants, including the fellows themselves, regarded as very successful.

Regardless of project type, fellows typically undertook multiple responsibilities, such as project management, research and development, and bringing their scholarly or disciplinary perspective to the work (table 2). In the following section, we discuss how the fellows' roles relate to organizational placement in more detail.

Table 1. *Types of projects reported during exit interviews listed by host institution (N = the number of projects reported by fellows per institution, not the number of fellows per institution). Some host institutions are missing due to incomplete reporting.*[7]

	N	Assessment and Planning	Collections	Education and Training	Infrastructure	Outreach	Software/tools development
Arizona State University	2	0	2	0	0	0	0
California Digital Library	3	2	0	0	1	0	0
Indiana University-Bloomington	2	1	0	0	1	0	0
Lehigh University	1	0	0	0	1	0	0
Pennsylvania State University	2	1	0	1	0	0	0
Purdue University	3	1	0	2	0	0	0
University of Alberta	2	2	0	0	0	0	0
University of California Davis	2	2	0	0	0	0	0
University of California Los Angeles	3	0	0	2	0	0	1
University of Colorado-Boulder/National Snow and Ice Data Center	2	1	0	0	1	0	0
University of Michigan	3	1	0	2	0	0	0
University of Minnesota	2	1	0	0	0	1	0
University of New Mexico	1	1	0	0	0	0	0
University of Notre Dame	2	0	0	1	0	0	1
Total	**30**	**13**	**2**	**8**	**4**	**1**	**2**

[7] CLIR staff were unable to schedule exit interviews with two of the fellows initially targeted for this study.

Table 2. *Roles occupied by fellows according to project type (N = the number of projects).*

Project Type	N	Project Management	Needs Assessment	Research and Development	Implementing IT Tools	Scholarly Perspective	Other
Assessment and Planning	13	11	9	11	4	9	3
Collections	2	2	1	2	0	1	1
Education/Training	8	3	3	4	2	3	2
Infrastructure	4	2	2	3	3	2	1
Outreach	1	1	1	0	0	1	0
Software/tools development	2	0	1	2	1	1	0
Total	30	19	17	22	10	17	7

Role of the CLIR/DLF Fellow

The majority of the CLIR/DLF fellows were housed in a university library. This placement was reflected in the fellows' responsibilities, which focused largely on environmental scans, needs assessment associated with research data management (RDM), and library support services for research data. As noted earlier, many fellows were tasked specifically with assessing faculty needs for RDM services and support in specific disciplines, often in relation to the NSF's data management plan requirements.

This emphasis on the assessment and planning phase of RDM support services suggests that many host institutions were in the formative stages of RDM programming and may have seen CLIR/ DLF fellows as a way to initiate these services. Researcher needs assessment and the development of institutional structures and services is valuable and important work, but from the programmatic standpoint of the postdoctoral fellowship it risks not fully utilizing fellows' disciplinary expertise and duplicating research efforts and outcomes between institutions. However, as the RDM field continues to mature, we expect this type of work to diminish for future cohorts.

Very few fellows were placed within disciplinary teams that were actively collecting research data. Instead, fellows who worked with primary datasets typically worked retroactively to curate already-collected sources, often via digitization initiatives. One project collaborator underscored the importance of embedding disciplinary expertise on active research teams while describing difficulties in working with historical archeological data and trying to determine after the fact which files were important and should be retained, explaining, "The people who do the actual archeological project are the ones who should be deciding, 'this [output] is really important work product and this is something that we just did as a note that we turned into this other file so you don't need to keep it.' That kind of cleaning out should really be happening when the project is active. And so one of our biggest problems was trying to sort out after a project was already done, 'how much of this stuff do we actually need to keep?' because the default position for a lot of people was just to keep everything. For us to go back retroactively not knowing why a file existed to figure out if it was worth keeping or not that's a problem."

Within their institutions, fellows often provided a type of connective tissue between functional areas, which assisted in filling gaps in responsibilities for data management services and provided new paths of communication between libraries and academic departments. Many fellows' supervisors and collaborators characterized one key aspect of fellows' work as acting as translator or interpreter between librarians and disciplinary faculty. For example, one project collaborator in archeology explained: "[the fellow] had

all that library and digital knowledge, but she also understood enough about archeology that we could talk like normal people and I wasn't like 'I don't understand what you mean by all this IT stuff.' And I think she fundamentally understood why we keep the records we keep, which I think really varies by discipline."

As one supervisor noted, the nature of many of the problems associated with data management, curation, and preservation are sociological as much as technical, and require a process of building trust between researchers in different disciplines. Another supervisor echoed this observation, commenting that within the realm of RDM, libraries must demonstrate themselves to be credible and trustworthy partners so that researchers can be confident that the library can handle the ethical and legal requirements required for long-term data management and curation of data. CLIR/DLF Fellows are perhaps uniquely suited for helping to establish and maintain this type of trust by helping to broker relationships between and among researchers and libraries.

Nevertheless, the experience of working simultaneously in a discipline and a library can be a source of tension for some fellows. One fellow remarked, "It's important to recognize the distinctions between 'library work' and 'research work' and for fellows to clearly articulate what it is they want, and for hosts/supervisors to articulate what they need." Another fellow responded to this tension by identifying explicitly as a data curation specialist and not as a librarian.

A number of fellows observed that they felt their librarian colleagues did not understand the value of their research activities, or in some cases resented time and resources provided for this work. These instances of identity politics can have significant negative effects on fellows' experience of the CLIR/DLF program, their successful integration within their institutional management structures, and their future career decisions with regard to continuing in data management fields.

Characteristics of Successful Fellows

Fellows, supervisors, and project team members almost universally cited communication, collaboration, and project management skills as vital to the success of a CLIR/DLF fellow, as well as flexibility and a facility for learning new tools and approaches to problems. Skills for negotiating the complex social organization of academic libraries and universities were also regularly mentioned, such as the ability to identify effective partners and recognize gatekeepers among librarians and faculty who often operate outside of formal reporting structures, and to understand the multifaceted relationships among and between disciplinary faculty members and librarians.

In particular, fellows were often asked to manage and respond to institutional change. For example, when asked about the skills a fellow needed to be successful, one

supervisor responded: “Working in team environments and [knowing] how to negotiate unpopular change well. . . . [Fellows] are being plunked down into a world in which they are not familiar and expected to do change management with a group of people that they have no experience with. . . . It’s a very different skill set to be successful as an academic than it is to be a leader of people and a change manager.” Another collaborator noted the importance of fellows exhibiting empathy for (especially librarian) colleagues who may be experiencing and negotiating significant structural changes in their work roles, and who may view fellows as instruments of these changes.

While these institutional contexts are out of fellows’ control, they can nevertheless have profound effects on the success of the postdoctoral fellowship for both parties. As one supervisor observed, fellows need the ability to recognize where their organizations are located on a broader spectrum of technological and cultural change within academic libraries and higher education, as well as the skills to cope with changes. This is an area that CLIR can emphasize or expand in its orientation for fellows and ongoing support programming.

Project team members said that fellows contributed valuable knowledge of the ethical norms and research practices of disciplinary scholars, and they particularly appreciated fellows who combined subject-area depth of understanding with technical abilities. Both project team members and supervisors observed that fellows’ knowledge of disciplinary norms, processes, and practices was extremely useful in bridging communication gaps between faculty members and librarians. Given the wide variation in disciplinary RDM needs in areas such as confidentiality and security, these skills are especially beneficial for project teams.

For example, at the University of Alberta one project team processed and curated images donated by the family of Otto Schaefer, a physician who worked with Inuit peoples in northern Canada and elsewhere in the 1950s and1960s. While these images were digitized to enable greater access and use, the collection contained culturally sensitive images, such as photographs of ceremonies, as well as clinical images of medical conditions in which individuals were readily identifiable, especially to members of the community where the photographs were taken. Further complicating matters, it was not always clear which photographs were taken as part of Schaefer’s clinical practice, which were part of his research interest in anthropological matters, and which were taken as someone who lived in the community. Working according to the principle that the community should have input, and sometimes the final say, in who should be able to access these images and for what purpose, the CLIR/DLF fellow worked with the university archivist and privacy officer to develop a methodology for evaluating images’ subject matter and obtaining community review of images flagged for potential sensitivity. Other types of RDM projects that required fellows to help implement levels of

access control included projects centered on health data, which is subject to complex ethical and legal disclosure and use requirements, and projects dealing with archeological data, which collect information about site locations and artifact inventories that must be kept confidential in order to protect against looting and theft.

Finally, project team members often noted the usefulness of fellows' methodological skills in areas such as survey analysis, statistics, and research design—expertise that is often in demand within libraries.

Challenges and Obstacles

Fellows reported institutional and organizational structures as some of the most significant challenges that they faced during their fellowships. Fellows described difficulties including conflicting goals between fellows and supervisors, a lack of support from library and institutional administrators, and insufficient institutional understanding of the fellow's role. The time horizon of a two-year fellowship within institutions that often change very slowly also presented difficulties for some fellows. The success of fellowship depends on managing expectations about what can be accomplished within these constraints. Fellows, supervisors, and project team members all reported the need for short-term, attainable goals during the fellowship, since unrealistic goals and "mission-creep" can significantly limit the fellow's effectiveness. As one supervisor noted, postdoctoral fellows are in a fairly weak structural and political position within their institutions, and require substantial administrative support to be fully effective.

Fellows typically reported a fairly steep learning curve acclimating to institutional culture, resulting in a slower work pace during the first year and a significantly accelerated work pace during the second year. Building networks quickly and learning the internal politics and cultural norms of complex institutions was demanding and time consuming for many new fellows, especially for those with no experience working in libraries. Despite these difficulties, one supervisor observed that from an organizational management perspective, a two-year postdoctoral appointment can sometimes serve as a useful bridge to creating a permanent position since those years can provide time to plan and secure funding.

Structural and organizational change also presented significant obstacles for some fellows, especially those who arrived at their host institutions when libraries or other divisions were reorganizing. Unfortunately, fellows were sometimes regarded as symptoms or agents of these changes, particularly when the newly implemented structures were viewed negatively by library faculty and staff with longer institutional experience. Postdoctoral fellows were often used to institute new programs or initiatives which, when managed well, had significant positive impacts; nevertheless, fellows in these positions were occasionally perceived by colleagues as threats to other library

staff. When paired with other adversarial management tactics, fellowships can be misused to force particular agendas for institutional change.

Some librarians found fellows' close disciplinary relationships with faculty members challenging to their own institutional authority. When not addressed, these situations could lead to conflict among fellows and librarian colleagues. Fellows sometimes felt isolated by simply working in areas outside their colleagues' usual day-to-day activities. As one supervisor noted, "[the fellow] was working on a research project but she was in a unit where what she was working on wasn't necessarily the things that other people kind of right next to her in her physical space were working on a regular basis, and so I think from time to time she may have felt a little bit, sort of, isolated in that sense." Nevertheless, another supervisor observed that this outsider perspective was useful and productive for the host institution, arguing that it is important for fellows to remind host organizations that they are there to think about the bigger picture—"to learn, to produce, and to be willing to push back [since] many libraries don't have this perspective."

At times, fellows' enthusiasm for new approaches was perceived as counterproductive by library colleagues and collaborators. Fellows were not always aware of the institutional history, constraints, or politics experienced by longer-tenured colleagues. This could sometimes lead to friction as fellows working with short time horizons advocated for identifying and addressing issues immediately while librarians preferred to carefully plan and document new strategies for the long term.

Perhaps unsurprisingly, time and funding constraints were commonly named as significant obstacles for CLIR/DLF fellows. Within the two-year time constraint of a fellowship, it is often very difficult to obtain funds not already set aside for a particular project, and the process of planning, applying for, and obtaining larger grant funding is often impossible during a fellow's tenure. In one case, grant funding was depleted before a data repository could be created, putting at risk the research data intended for storage in the repository. Such constraints are often difficult for a fellow to address within the context of a fellowship.

Infrastructure needs should therefore be addressed prior to the start of a fellowship whenever possible. A number of fellows and project team members noted that projects met unexpected infrastructure difficulties, such as competition for resources among internal IT groups, a lack of access to servers, or repositories that were not set up to store specialized data. These types of barriers often had the potential to seriously delay or even halt data curation work. As RDM repository infrastructure, norms, and workflows and procedures continue to be developed and tailored to local institutional requirements, these problems could diminish for future fellows.

Finally, several fellows reported a need for additional software development and technical skills or a greater level of support in these areas. Institutions planning for fellows should consider devoting resources to fellows' skills development or partnering fellows with software development specialists. CLIR and DLF could also secure additional resources for broadening the array of technical training opportunities available to participants in the program.

Recommendations

A successful CLIR/DLF fellowship requires a foundation of planning and administrative support at the host institution both from the upper-level leadership and within the division or department where the fellow will conduct their day-to-day work. Host institutions and administrators should consider in advance the financial as well as human resources required for a successful fellowship. These include such items as infrastructure and computing needs, appropriately staffed project teams, and the resources to partner with software development specialists. Difficulties and delays in these areas can hinder fellows working on relatively short time horizons and can potentially derail projects entirely.

Whenever possible, project grants should be obtained before the fellowship, unless one of the explicit goals of the fellowship is grant development. Unfunded projects, or projects requiring the first year of the fellowship to be devoted to securing grant funding, often fail to fully utilize fellows' disciplinary skills and limit their ability to implement data management and curation initiatives.

Host institutions and supervisors should work with newly appointed fellows early in their tenure to identify and develop explicitly stated goals that are realistically attainable during the relatively short fellowship term. "Mission-creep" and implicit expectations often conspire to limit fellows' effectiveness in addressing the core goals of their fellowships. Moreover, host institutions and administrators should also work with their staffs to prepare their expectations for the incoming fellow, especially if the fellow will undertake work in areas that are not always among a "traditional" librarian's responsibilities (for example, conducting original research or working on faculty research teams). Setting these expectations in advance helps significantly in minimizing friction between fellows and their colleagues. Additionally, since many fellows reported spending significant amounts of time conducting outreach with faculty and other constituencies, setting expectations for how this work will be supported by colleagues and administrators during the fellowship and sustained beyond the fellowship is important for maximizing the long-term impact of fellows' contributions.

Succession planning for projects that will continue after a fellowship is completed should be addressed as early as possible during a fellow's term. Fellows often reported

difficulty in ensuring that projects would be maintained, continued, or completed once their fellowship ended. When the long-term continuation of a fellowship project was not possible, some fellows emphasized the importance of making efforts to develop collaborations both within and outside the host institution, so that their project work could be portable to a new institution if necessary. In either case, given the short time of the fellowship transition, planning is critical to maximizing the impact of fellows' work.

To support host institutions in creating appropriate and effective CLIR/DLF fellowship positions, CLIR should continue to develop and refine its criteria and guidelines for host institutions and create a set of best practices for administrators and supervisors to use in preparing for and managing the fellow's work. In particular, CLIR should discourage using fellowship positions instrumentally to pursue organizational change goals. Positions placed in the midst of broader organizational change efforts often produce work environments that do not support fellows and their work, severely limiting the fellowship's potential for positive impact on both the fellow and the host institution.

There was near-universal agreement among fellows, project collaborators, and supervisors interviewed for this study that successful fellows exhibit skills in communication, collaboration and project management, as well as flexibility and a facility for quickly learning new tools and approaches to problems. These observations suggest that CLIR should continue to emphasize these "soft skills" in its development and training programs for fellows. In particular, CLIR should expose fellows to current approaches to managing organizational change, and to the organizational and bureaucratic aspects of working within universities, such as governance and financial structures. While fellows have usually spent a great deal of time as students or teaching assistants at universities, they are often relative novices in understanding the administrative functions of their institutions. Developing this understanding is critical when they are required to immediately work within these structures as part of their fellowship responsibilities.

Finally, host institutions should make every effort to embed fellows in active research teams and to support the fellows in building partnerships with researchers. Contrary to our expectations following the 2011–2012 study, few of the fellows in these cohorts worked directly with research data and almost none worked with data as it was being collected. Overreliance on fellows to shepherd programmatic or administrative changes risks underutilizing the fellows' disciplinary expertise, which may be better applied to developing the trust and close relationships with faculty that data curation services require. Considerable work is still needed to develop a system of data curation that supports the preservation and access of research data while allowing researchers to fulfill their ethical and professional obligations, and an improved understanding of research processes and their variation among disciplines is critical to this work.

Supporting fellows in becoming trusted partners in the research community could be a path to bridging the communication gap between libraries, IT, and the research community.

APPENDIX A:
PROJECTS INVOLVING CLIR/DLF FELLOWS IN DATA CURATION, 2012–2014

List of Fellows' Projects
Aiding & assisting with the development of UC Davis-specific research data services
Biomedical Library partnership to develop an electronic lab notebook tool
Building South Bend (Historic Urban Environments Lab)*
Carlos Montezuma's Wassaja Newsletter: Digitization, Access, and Context*
College of Engineering Needs Assessment and Outreach Pilot Project
Creating effective system to manage data from Arctic social sciences at the National Snow & Ice Data Center*
DASH Digital Arts Sciences + Humanities
Data Curation Profiles
Data Education Working Group
Data Forward Marketing Channel
Development of a university-wide research data management (RDM) policy and service infrastructure
E-Science Research Peer Networking and Mentoring Group (ERPN-MG)
EarthCube*
Educational Data Curation Framework (ECDF)
Environmental data scan for social sciences faculty
Evaluating a Cooperative Approach to Managing Digital Archaeological Resources (ECAMDAR)*
General research regarding data curation & management
GIS consulting
GIS Day
Groundwork for the creation of restricted data services in the Penn State Libraries and throughout campus
Investigated the research data management needs & desires of UC Davis community (year 1 of fellowship)
Making Data Count
North Atlantic Biocultural Organization (NABO) and Global Human Ecodynamics Alliance (GHEA) Cyberinfrastructure project
Outreach to and work with liaisons regarding data management
Research Data Alliance
Research Data Working Group (RDWG)
Researching to understand the data publication landscape
SEAD DataNet*
Serving as "research informationist" and teacher/curriculum developer
Supporting proof of concept projects through collaborative grant writing

*Projects containing significant work with primary research data

References

Allard, Suzie, Christopher Lee, Nancy Y. McGovern, and Alice Bishop. 2016. *The Open Data Imperative: How the Cultural Heritage Community Can Address the Federal Mandate.* Washington, DC: Council on Library and Information Resources. Available at https://www.clir.org/pubs/reports/pub171/

Brembs, Bjorn, Katherine Button, and Marcus Munafo. 2013. Deep Impact: Unintended Consequences of Journal Rank. *Frontiers in Human Neuroscience* 7: 1–12.

Brischoux, François, and Frédéric Angelier. 2015 Academia's Never-ending Selection for Productivity. *Scientometrics* 103(1): 333–336.

Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. Data Journals: A Survey. *Journal of the Association for Information Science and Technology* 66(9): 1747–1762.

Cokol, Murat, Ivan Iossifov, Raul Rodriguez-Esteban, and Andrey Rzhetsky. 2007. How Many Scientific Papers Should Be Retracted? *EMBO Reports* 8(5): 422–423.

Donoho, David L., Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. 2009. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering* 11(1): 8–18.

Eisner, D. A. 2018. Reproducibility of Science: Fraud, Impact Factors and Carelessness. *Journal of Molecular and Cellular Cardiology* 114: 364–368.

Fanelli, Daniele. 2010. Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PloS One* 5(4): e10271.

Fang, Ferric C., and Arturo Casadevall. 2011. Retracted Science and the Retraction Index. *Infection and Immunity* 79(10): 3855–3859.

Fang, Ferric C., R. Grant Steen, and Arturo Casadevall. 2012. Misconduct Accounts for the Majority of Retracted Scientific Publications. *Proceedings of the National Academy of Sciences of the United States of America* 109(42): 17028–17033.

Fischer, Joern, Euan G. Ritchie, and Jan Hanspach. 2012a. Academia's Obsession with Quantity. *Trends in Ecology & Evolution* 27(9):473–474.

_____. 2012b. An Academia Beyond Quantity: A Reply to Loyola et al. and Halme et al. *Trends in Ecology & Evolution* 27(11):587–588.

Halme, Panu, Atte Komonen, and Otso Huitu. 2012. Solutions to Replace Quantity with Quality in Science. *Trends in Ecology & Evolution* 27(11): 586; author reply 587-8, accessed October 10, 2016.

Ioannidis, John P.A. 2014. How to Make More Published Research True. *PLoS Medicine* 11(10): e1001747.

Jahnke, Lori, Andrew D. Asher, Spencer D. C. Keralis, and Charles Henry. 2012. *The Problem of Data*. Washington, DC: Council on Library and Information Resources.

Lane, A., O. Luminet, G. Nave, and M. Mikolajczak. 2016. Is there a Publication Bias in Behavioural Intranasal Oxytocin Research on Humans? Opening the File Drawer of One Laboratory. *Journal of Neuroendocrinology* 28(4).

Nosek, Brian A., Jesse Graham, Nicole M. Lindner, Selin Kesebir, Carlee Beth Hawkins, Cheryl Hahn, Kathleen Schmidt, Matt Motyl, Jennifer Joy-Gaba, Rebecca Frazier, and Elizabeth R. Tenney. 2010. Cumulative and Career-Stage Citation Impact of Social-Personality Psychology Programs and their Members. *Personality & Social Psychology Bulletin* 36(10): 1283–1300. Available at http://www-bcf.usc.edu/~jessegra/papers/NGLKHHSMJFT2010.pdf; accessed October 20, 2016.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science* 7(6): 615–631.

Ostriker, Jeremiah P., Paul W. Holland, Charlotte V. Kuh, and James A. Voytuk. 2009. *A Guide to the Methodology of the National Research Council Assessment of Doctorate Programs*. Washington, D.C.: National Academies Press.

Sarewitz, Daniel. 2016. The Pressure to Publish Pushes Down Quality. *Nature* 533(7602): 147.

Schooler, Jonathan. 2011. Unpublished Results Hide the Decline Effect. *Nature* 470(7335): 437.

Sills, Jennifer. 2016. Measures of Success. *Science* 352(6281): 28–30.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22(11): 1359–1366.

Smaldino, Paul E. and Richard McElreath. 2016. The Natural Selection of Bad Science. *Royal Society Open Science* 3(9).

Steen, R. Grant. 2011. Retractions in the Scientific Literature: Is the Incidence of Research Fraud Increasing? *Journal of Medical Ethics* 37(4):249–253.

Stodden, Victoria. 2011. Trust Your Science? Open Your Data and Code. *Amstat News* (July 1). Available at http://magazine.amstat.org/blog/2011/07/01/trust-your-science/.

van Dongen, Stefan. 2011. Associations Between Asymmetry and Human Attractiveness: Possible Direct Effects of Asymmetry and Signatures of Publication Bias. *Annals of Human Biology* 38(3):317–323.

Voytek, Bradley. 2016. The Virtuous Cycle of a Data Ecosystem. *PLoS Computational Biology* 12(8): e1005037.

Young, Neal S., John P. A. Ioannidis, and Omar Al-Ubaydli. 2008. Why Current Publication Practices May Distort Science. *PLoS Medicine* 5(10): e201.

Data Jobs: A Place for Science and Social Science PhDs in the Libraries?

Jodi Reeves Eyre

Introduction

This white paper is a snapshot of trends in library jobs related to data management, curation, and services (*data jobs*) current in the first four years of the program. The paper focuses on how responsibilities and qualifications listed in data job postings from 2013–2017 relate to qualifications and experience held by recent PhDs in the Sciences and Social Sciences.

For this study, Jodi Reeves Eyre, then of Eyre & Israel, LLC, and now program officer for CLIR, conducted an analysis of more than 161 job postings from 2013–2017 to identify trends in desired qualifications, requirements, and the numbers of data jobs advertised, with focus on the suitability of PhDs in the Sciences and Social Sciences to fill these positions as advertised.

The analysis shows the following findings:

- Libraries seek candidates who can provide services related to data management, curation, instruction, and visualization, and related scholarly communication activities[1].
- Ideally, candidates have a knowledge of metadata, qualitative and/or quantitative research experience, and some library service experience.
- Hard barriers that keep recent PhDs from these positions include requirements for a library degree, library experience, or both.
- Soft barriers include the emphasis on service/consultation and lack of emphasis on active research opportunities in job postings.
- There were no appreciable increases or decreases in the total number of data-related position postings analyzed between 2013 and 2017 when compared to trends in the number of overall postings.

Further research should be done into whether recent PhDs are interested in entering service/consultation fields and whether recent graduates have the skills required for this type of work. CLIR's early experiences with the Fellowships in Data Curation for the Sciences and Social Sciences suggest that data service and consultation work can be

[1] These terms mean different things to different people depending on their professional or disciplinary contexts. They may also refer to similar sets of activities. This is illustrated by the common co-occurrence of terms, as discussed in the Results section of this paper. CLIR's *The Open Data Imperative* has useful definitions of data curation, digital curation, and data management (Allard et al. 2016, 9).

attractive to some recent PhDs, and that these candidates bring and can quickly develop skills that make them valuable library colleagues.

Methodology

Data Collection

The majority of data for 2013 through most of 2016 came from the ARL Position Description Bank, a service that the Association of Research Libraries (ARL) provided for its members. ARL staff shared data on thousands of jobs with CLIR. The ARL Position Description Bank was curated by ARL staff with human resources officers at ARL member institutions submitting the job listings for distribution within the ARL community. CLIR received the data from ARL on October 20, 2016.[2] The data set included 2,773 job postings. Reeves Eyre identified duplicates and unposted listings during the initial review of the ARL data but determined that these listings did not significantly affect the analysis needed for this study. These duplicates and unposted listings were left in the original ARL data set for quantitative analysis.

Additional job postings were collected for October 2016 to December 2017 from the Digital Library Federation (DLF) Job Board, a job posting service for DLF member institutions.[3][4] The final DLF data set included 275 job postings with no duplicates. Combined with the ARL data, the entire sample included a total of 3,048 job postings. Not all postings contained complete job descriptions. Some described the positions, but did not list all requirements or compensation, while others included basic information and a now defunct link to the complete posting.

Creating the Data Jobs Data Set

Job postings that contained the words *data, digital curation*, or *curator* in the position title were selected for quantitative and qualitative analysis. Some job postings advertised two positions in one posting. These postings were split in two with each position analyzed as if it were a separate posting. This resulted in a subset of 130 data job postings.

For comparison, 31 additional job postings that did not contain the words data, digital curation, or digital curator in the position title were included in quantitative and qualitative analysis. For this subset, one posting meeting the criteria out of every 100 consecutive postings for each year was arbitrarily chosen. This guaranteed at least one

[2] Thank you to Sue Baughman and Gary Roebuck of ARL for their help in making the ARL Position Description Bank data accessible.
[3] The DLF is a program of CLIR.
[4] Thank you to James DeVos, Arizona State University, for his help in collecting these job postings.

selection from each year under study and a selection representing approximately one percent of the original 3,048 jobs postings.

The resulting subset of 161 job postings was analyzed to identify general trends regarding job responsibilities and required qualifications using the quantitative and qualitative methodology outlined below.

Analysis Methodology

A mixed methods approach was used to determine any commonalities among the requirements and preferences employers sought in the data-related job postings, in comparison with the sampling of other jobs from the same time period, as well as whether these requirements and preferences shifted over the four-year time span covered by the postings. Quantitative data collected included the date submitted (ARL postings) or posted (DLF postings), degree required or preferred (if any mentioned), tenure-track status (if mentioned), years of experience required or preferred (if mentioned), and whether the position was a liaison role for one or more academic departments (if mentioned). Job descriptions were analyzed using the qualitative analysis software package, Dedoose. The descriptions were coded, a process in which excerpts from the position descriptions were selected and assigned a set of tags. These tags, or codes, were then used to identify trends in the postings.

Coding was based initially on the terms used in the postings themselves (data management, service, and teaching are some examples). As coding proceeded, however, some terms were grouped under more general codes, such as *specific tools* as a code for any reference to a specific software package, digital tool, or programming language, etc. that was mentioned. Teaching, workshops, and training were grouped together under *instruction*.

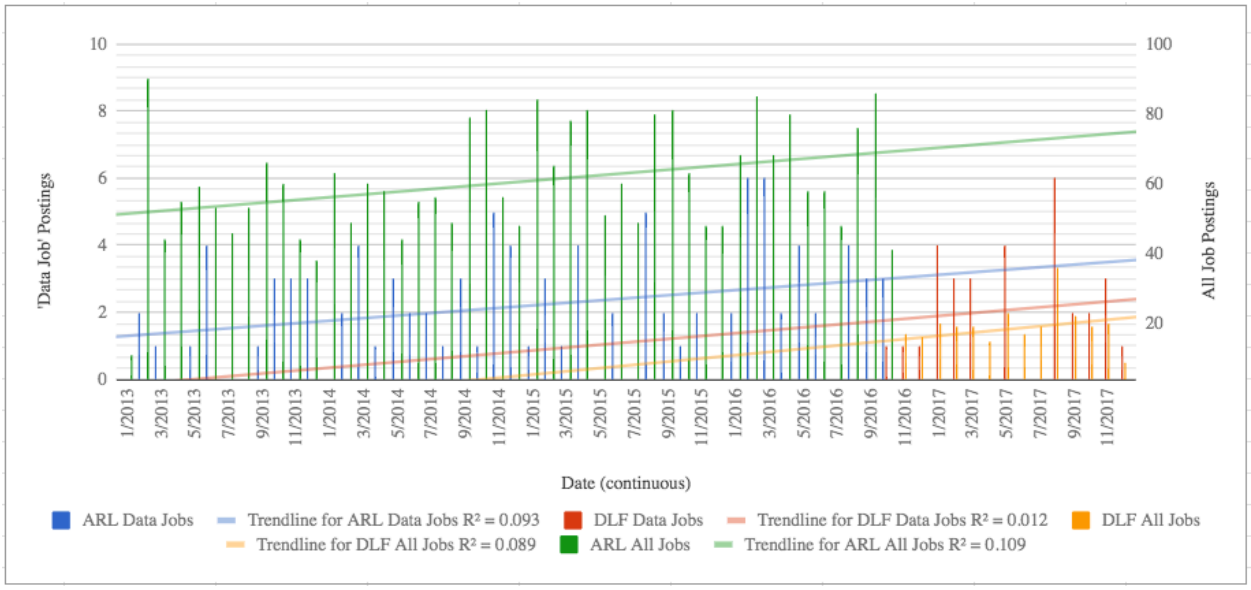
Results

Quantitative Trends in the Numbers of Data Jobs Advertised

The frequency of data job postings matching the selection criteria was compared to the overall frequency of job postings between 2013 and 2017. There was no appreciable increase or decrease in the number of data postings analyzed over this period when compared to trends in the number of overall postings (figure 1).

The period of time covered by the data could account for the lack of significant trends. In 2013, the US government issued a mandate requiring federal agencies with annual research and development expenditures of more than $100 million to create plans for

Fig. 1. *Number of jobs and data jobs submitted to ARL or posted to DLF. Data jobs follow the scale along the right vertical axis. All jobs follow the scale along the left vertical axis.*



public access to agency data (Holdren 2013). Including jobs listed prior to the mandate in the analysis may have yielded an increase in the frequency of postings for data jobs starting in 2013. Limiting analysis to postings with data or digital curation in the job title may also have affected the ability to detect time-based trends. For example, analyzing a larger group of postings including all jobs mentioning any responsibilities related to data may have yielded different results. Table 1 shows seven postings from the 32 non-data job comparison sample that do mention data-oriented responsibilities or qualifications (table 1).

Table 1. Non-data *job postings that include data-oriented responsibilities or qualifications*

Job Posting	Year Posted/Created	Data-oriented Keywords
Digital Archivist	2013	Data Management, Metadata
Metadata Librarian	2014	Metadata
Digital Services Specialist	2014	Metadata
Social Sciences and Documents Librarian	2015	Data Management, Data Curation, Other Data Services/Support
Authorities/Identities Metadata and Cataloging Librarian	2016	Metadata
Black Studies Librarian	2016	Data Services, Data-intensive Research Support
Digital Scholarship Repository Specialist	2017	Metadata, Data Curation, Linked Data

Responsibilities

Serving Research Needs: The analysis showed that data jobs posted during this period were service- or consultancy-oriented as opposed to partnership-oriented. This was determined through qualitative and quantitative analysis of all the data job postings and by analyzing a subset of the postings. The subset of 13 data job postings selected for in-depth analysis represented approximately 10% of the total data job postings analyzed. For this subset, one out of every 10 consecutive postings was selected for analysis, guaranteeing at least one selection from each year under study.

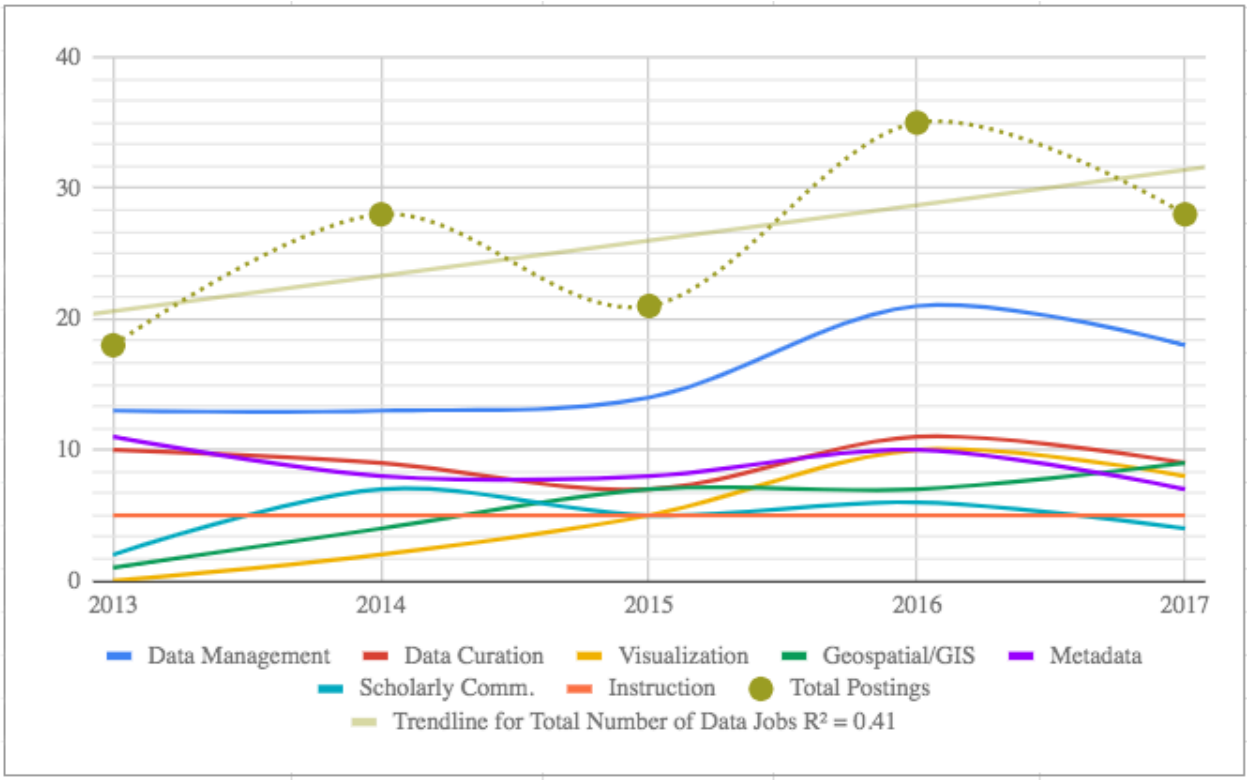
Five of the 13 postings selected for in-depth analysis were for candidates that would be based in the library and expected to work within internal library research support groups or centers; five were for liaison roles where candidates would be assigned to work with specific academic departments; and two were for candidates who would collaborate with or serve external research centers. None of the postings were for positions embedded or engaged with a specific faculty or departmental research project. This was a departure from the model for CLIR postdoctoral fellowship positions; hosts for these positions are encouraged to arrange dual appointments between the library and an academic department. Among the 13 positions selected for detailed analysis, there was only one dual appointment identified: a physics and research data librarian position at the University of Toronto that would be shared by the Department of Physics and the Science Libraries.

While data job candidates were not typically expected to lead or collaborate on specific research projects, the larger data set shows that prior research experience (18 of 130) and grant writing experience (9) were occasionally required or desired. Candidates with the skill set and experience that would enable them to lead or collaborate on a research project outside of a service role may have been attractive for those positions, even in cases where these leadership skills were not explicitly mentioned. Additional analysis of responsibilities and required or requested experience showed that strong research, data management, analysis, and visualization skills were valued in potential candidates for the 130 data job postings. These skills are notably similar to those valued in among research collaborators working outside library organizations.

Data Responsibilities and Experience: Most of the data-oriented positions analyzed for this study focused on providing data services and consultation. Data management services were the most dominant in the sample (mentioned in 79 of the 130 data jobs), followed by data curation support and services (44). One third of the postings also emphasized responsibilities and experience related to metadata (46). Instruction (34), data visualization (25), scholarly communication services (24), and geospatial services

(28) were also in high demand, relative to other types of responsibilities and experience mentioned. Following these areas of emphasis, project management was the next most frequently mentioned responsibility, but it appeared in only 16 job postings. There was no discernable increase or decrease over time in terms of the types of data-related responsibility and experience included in the postings (figure 2).

Fig. 2. *Number of data job postings that mention major data responsibilities and experience over time*



Most postings required a range of experience and responsibilities related to data services. Each of the seven most frequently mentioned areas of responsibility and experience co-occurred with the others, with the highest instances of co-occurrence for data management and data curation, data management and metadata, and data curation and metadata (table 2). This may indicate that these three areas of expertise are seen as integral to one another; for example, a professional with solid data management and metadata expertise would have good qualifications for providing data curation services.

Table 2. *Number of data job postings where co-occurrence of primary data responsibilities and experience is present*

	Data Mgmt.	Scholarly Comm.	Geospatial	Instruction	Visualization	Data Curation	Metadata
Data Mgmt.		20	9	24	15	37	34
Scholarly Comm.			10	10	6	11	6
Geospatial				10	15	4	11
Instruction					10	13	14
Visualization						2	6
Data Curation							26
Metadata							

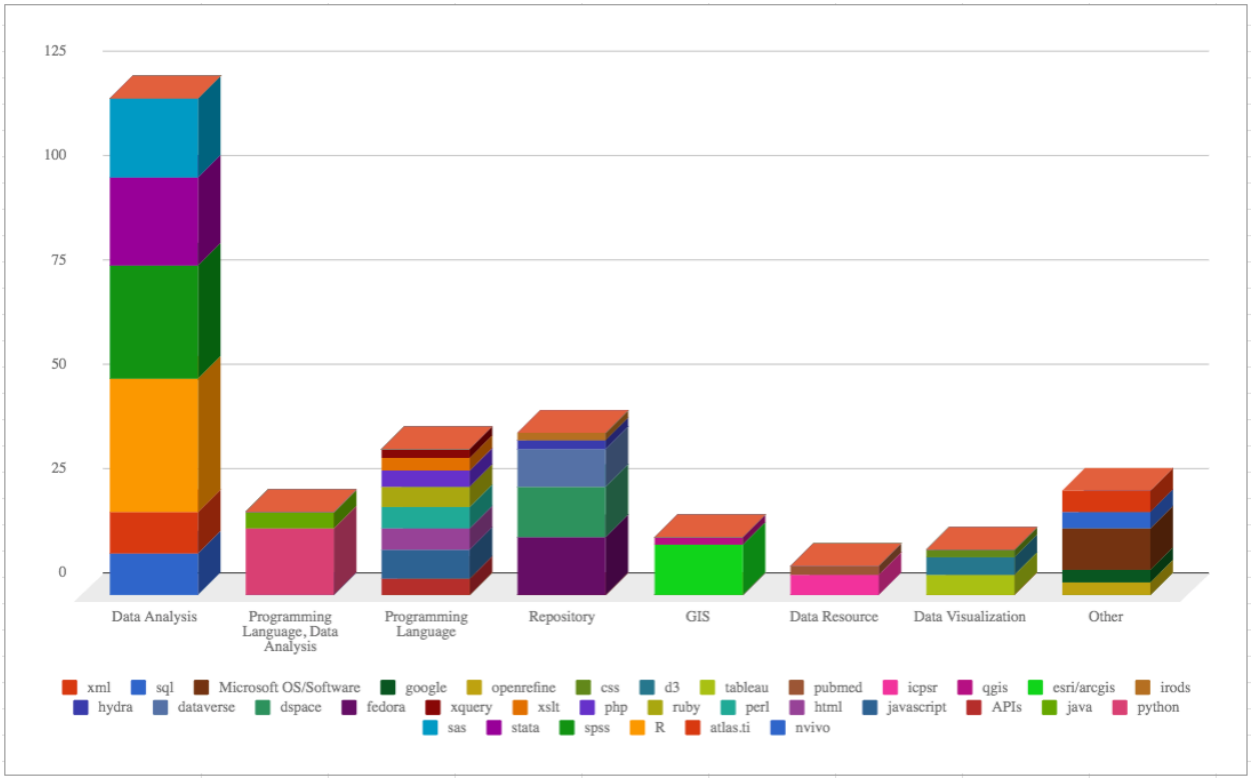
The required level of familiarity and experience with these seven areas varied among the postings. A posting for a digital data repository specialist required a candidate with an "[u]nderstanding of the research process, current issues in scholarly communication, and the role of libraries and librarians in research data curation." A research data management librarian needed to demonstrate "active engagement in the scholarly discussions and experimentation taking place to address questions about long-term research data curation and preservation for reuse." Neither asked for a specific number of years' experience with data curation, unlike the research data librarian posting that required "four to five years of data preservation, curation management including data sets, content description and representation, metadata standards, and relevant workflows."

Qualifications

Knowledge and Skill Sets: To determine the range of skills and knowledge sets candidates were expected to possess for the data-related positions, the text of all 130 postings was compiled into a single corpus for text mining. *Terms* from Voyant Tools identified the raw frequency of the common terms that had been grouped under *specific tools* through the coding process in Dedoose. This list was limited to tools and software that appeared more than once. The tools and software on this list were manually

grouped into broader categories (figure 3). These categories overlapped in many places; for example, some tools for data analysis were also programming languages. The tools grouped under data analysis encompassed both qualitative and quantitative analysis.

Fig. 3. *Frequency of references to tools and software, including tools and software mentioned more than twice. Tools and software have been grouped into broad categories.*

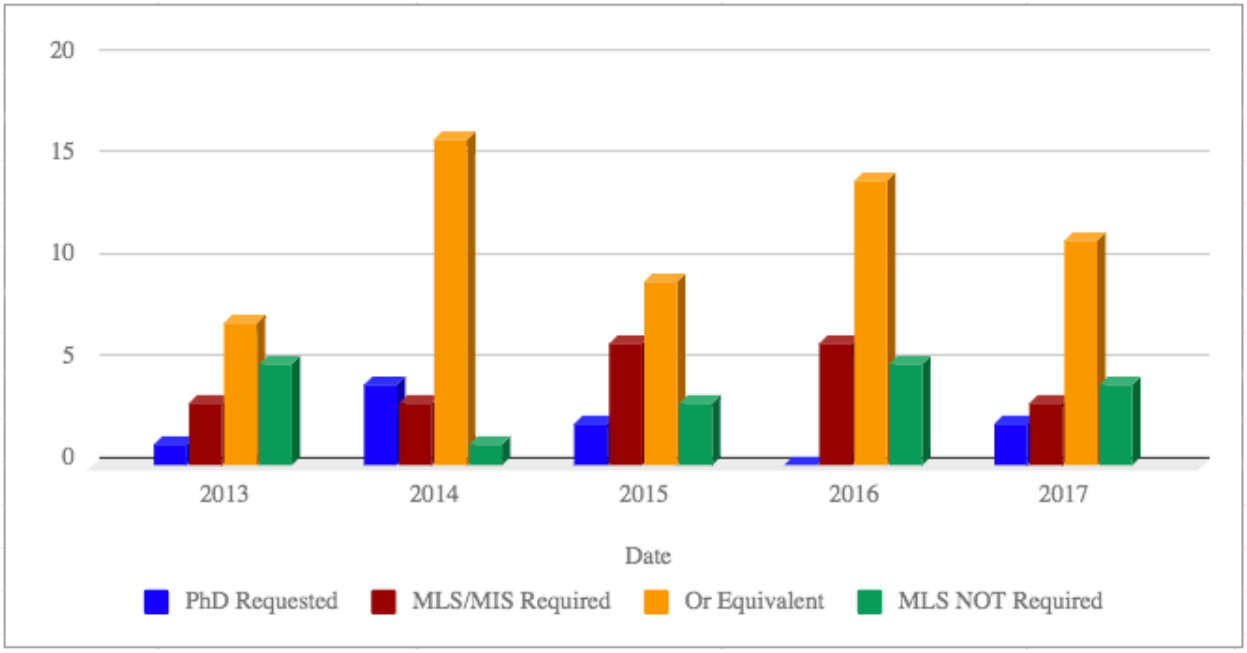


The majority of sought-after skills related directly to both quantitative and/or qualitative data analysis. This may indicate that data analysis experience is seen as an intrinsic part of providing data services, even if the primary focus of a position is on management and curation. For example, a candidate with data management experience and experience using R to analyze and visualize qualitative and quantitative data would likely be well qualified for a number of the library-based data jobs in the sample.

Degree Qualifications: Ninety-seven of the 130 data job postings listed some kind of required or desired degree. Most required or desired a library or information science master's degree or equivalent degree (also see Figure 4). Specifically:

- Twenty-one required or desired a library or information science master's degree, usually from an accredited program. This is a barrier for PhD holders aspiring to these positions.
- Fifty-seven required or desired a library or information science master's degree or an equivalent degree—allowing for some leeway for considering candidates with other advanced degrees and relevant experience.
- Nine data postings requested candidates with PhDs or advanced degrees beyond an MLS/IS. One of the 31 non-data jobs analyzed requested a PhD. One data job mentioned a LIS PhD.[5]

Fig. 4. *Number of postings per year that stated whether a library or information science degree was required or not, or whether an equivalent was required or desired, as well as number of postings that stated whether a PhD or terminal degree was required/desired.*



There was a substantial increase in positions accepting *equivalent* degrees or experience as opposed to only MLS/MIS degrees from 2013 to 2014. Requests for PhDs also increased at this time. This coincides with the publication of the 2013 OSTP memo mandating public access to data (Holdren 2013), but a causal relationship cannot be established.

[5] "Master's or PhD in Library Science, Information Science, Library and Information Science, or Data Science; strong background in a social science/humanities discipline with a tradition of generating and using qualitative data . . . strongly preferred – OR – Master's or PhD in a social science with a demonstrably strong background in data management, information science, and/or data science."

Extent of Experience: Thirty-seven of the data jobs that listed a required degree also listed a minimum number of years of experience (table 3).

Table 3. *Number of job postings with a minimum number of years of experience required categorized by the number of years required*

Min Years' Experience	Number of Data Job Postings
1	12
2	7
3	12
4	6

Suitability for Recent PhDs in the Sciences and Social Sciences

If only research and subject knowledge, experience, and skill requirements are taken into account, PhD holders with the following kinds of experience might reasonably be assumed to be attractive candidates to these positions:

- accessing data from various sources, including repositories and databases common to a field/s
- managing larger amounts of data
- analyzing, visualizing, and making data available using the tools mentioned
- preparing data to a standard for publication or as part of funding responsibilities for internal or external reuse
- being aware of subject-specific descriptive and metadata practices and able to quickly build that skill set
- teaching or developing workshops, dependent on the program, subject, funding, or other factors

However, clear barriers exist for PhD-holding candidates for positions where the specific master's degree is required. Hard barriers included the requirement that candidates hold a master's in library or information science, or that they have library-based experience. Most of the postings accepted equivalent degrees or experience. One area for further study is whether libraries consider research or other experience as part of this requirement or if equivalent experience needs to be based on an academic library or data repository. PhDs interested in library positions need experience in a library or a service or consultation setting to be competitive. This kind of experience can be difficult to obtain for very recent science and social science graduates who are more likely to

have worked in collaborative settings, such as in laboratories, in the field, or in teaching, but not in a setting easily analogous to an academic library.

There is another, albeit soft, barrier to PhDs applying to and obtaining data job positions in libraries: the focus on service. Many positions required service-oriented individuals or individuals with the experience or disposition to work as a service provider to faculty. Several positions mentioned involvement with research centers, or the opportunity to conduct research as part of a tenure-track library position. This may lead recent PhDs to wonder whether data-related positions in libraries may limit the development of their own research interests or diminish their value as researchers. Recent PhD graduates may actually see many of their interests exemplified in these positions: the chance to use and grow current research skill sets, improve access to data, teach, and assist in the research endeavors of others. With some additional emphasis on the value of research and independent professional development and a broad vision for building data curation capacity through the infusion of diverse skills into the provision of data services, academic libraries may find they can attract strong candidates from a variety of disciplinary backgrounds to their teams.

References

Allard, Suzie, Christopher Lee, Nancy Y. McGovern, Alice Bishop. 2016. *The Open Data Imperative: How the Cultural Heritage Community Can Address the Federal Mandate*. Council on Library and Information Resources: Washington, DC. Available at https://www.clir.org/pubs/reports/pub171/.

ARL Staff. 2016. Descriptions from job postings from the ARL Position Descriptions Bank. 2016.obs_desc_10-20-16.csv. Data set created October 20, 2016.

ARL Staff. 2016. Metadata for job postings from the ARL Position Descriptions Bank (Jobs_meta_10-20-16.csv). Data set created October 20, 2016.

Atkins, Winston, Carol Kussmann, and Katherine Kim. 2017. "Staffing for Effective Digital Preservation 2017: An NDSA Report." *Open Science Framework*. October 20. doi:10.17605/OSF.IO/3RCQK.DLF Jobs. Available at https://jobs.diglib.org. Accessed October 15, 2017 and January 8, 2017.

Holdren, John P. 2013. *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research.* Available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Sinclair, Stéfan and Geoffrey Rockwell. "Terms." *Voyant Tools.* Web. 20 December 2017. Available at https://voyant-tools.org.