10-24-2014

"We're Working On It:" Transferring the Sloan Digital Sky Survey from Laboratory to Library

Ashley E. Sands
University of California, Los Angeles

Christine L. Borgman
University of California, Los Angeles

Sharon Traweek
University of California, Los Angeles

Laura A. Wynholds
Utah State University

Recommended Citation

Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2014). "We're Working on It:" Transferring the Sloan Digital Sky Survey from Laboratory to Library. 10th International Digital Curation Conference, San Francisco, CA, USA.

IJDC | *General Article*

We're Working On It: Transferring the Sloan Digital Sky Survey from Laboratory to Library

Ashley E. Sands
UCLA, Information Studies

Christine L. Borgman
UCLA, Information Studies

Sharon Traweek
UCLA, History and Gender Studies

Laura A. Wynholds
UCLA, Information Studies

Abstract

This article reports on the transfer of a massive scientific dataset from a national laboratory to a university library, and from one kind of workforce to another. We use the transfer of the Sloan Digital Sky Survey (SDSS) archive to examine the emergence of a new workforce for scientific research data management. Many individuals with diverse educational backgrounds and domain experience are involved in SDSS data management: domain scientists, computer scientists, software and systems engineers, programmers, and librarians. These types of positions have been described using terms such as research technologist, data scientist, e-science professional, data curator, and more. The findings reported here are based on semi-structured interviews, ethnographic participant observation, and archival studies from 2011-2013.

The library staff conducting the data storage and archiving of the SDSS archive faced two performance problems. The preservation specialist and the system administrator worked together closely to discover and implement solutions to the slow data transfer and verification processes. The team overcame these slow-downs by problem solving, working in a team, and writing code. The library team lacked the astronomy domain knowledge necessary to meet some of their preservation and curation goals.

The case study reveals the variety of expertise, experience, and individuals essential to the SDSS data management process. A variety of backgrounds and educational histories emerge in the data managers studied. Teamwork is necessary to bring disparate expertise together, especially between those with technical and domain education. The findings have implications for data management education, policy and relevant stakeholders.

This article is part of continuing research on Knowledge Infrastructures.

Received 21 April 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to Ashley E. Sands, GSE&IS Building, Box 951520, Los Angeles, CA 90095-1520. Email: ashleysa@ucla.edu

An earlier version of this paper was presented at the 9th 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/

Introduction

This article reports on the transfer of a massive scientific dataset from a national laboratory to a university library, and from one kind of workforce to another. The Sloan Digital Sky Survey (SDSS)[1] is significant for its scope, quality, public access, and extent of uses and users. The survey covered over a quarter of the night sky with high quality optical and spectroscopic imaging. The first phase of the SDSS project (SDSS-I) ran from 2000-2005, the second (SDSS-II) from 2005-2008, and subsequent SDSS projects continue today. The SDSS data are made freely available online to astronomers and the general public through data releases. The final data release of the SDSS-I/II project collaboration occurred in June 2009 (Abazajian et al., 2009).

While participants often referred simply to 'the data,' the SDSS dataset is a complex aggregation of materials representing multiple elements of an international project. The SDSS Long-Term Scientific Data Archive (hereafter referred to as the SDSS archive) is comprised of four related datasets:

1. The Data Archive Server (DAS), which contains the processed flat image files;
2. The Catalog Archive Server (CAS), which contains multiple releases of the image and spectroscopy SQL database;
3. The Software, which includes the code generated for the data collection, data processing, database creation, user interfaces, and the SDSS website; and
4. The Raw Data, which are the unprocessed data as received from the scientific instruments.

In total, the SDSS archive forms a collection between 100-200 terabytes.

The SDSS archive is a valuable case study in the transfer of control from a scientific environment to a library for stewardship. The transfer of the SDSS archive requires a complex array of workforce expertise. For the purposes of this article, we use the term data transfer to refer to the array of social and technical activities involved in moving and stewarding a scientific dataset. We use this data transfer process to examine the emergence of a new workforce for academic scientific research data management and to illustrate the variety of tasks and persons involved in managing large datasets. The case study also demonstrates the importance of teamwork across the workforce. This article is part of a larger set of studies of data practices in astronomy, environmental sciences, and other fields conducted by the Knowledge Infrastructures (KI)[2] team at UCLA.

Workforce Issues in Managing Scientific Data

Management of scientific data involves many people and tasks. Among the terms used to describe this workforce are research technologists (Lyon, 2007), data scientists (van der Graaf and Waaijers, 2011), e-science professionals (Stanton et al., 2011), data curators (Higgins, 2008), and more. Some of these examples are generic roles in scientific data management and others are specific to academic settings. For the

1 Sloan Digital Sky Survey: http://www.sdss.org
2 Knowledge Infrastructures: http://knowledgeinfrastructures.gseis.ucla.edu/

 doi:10.2218/ijdc.v9i2.336

purposes of this article, we are concerned with scientific research data; with the workforce responsible for managing, stewarding, storing, archiving, curating, or preserving those data; and with the knowledge and expertise required for these activities.

Analyses of the knowledge and expertise required for data management are often based on evidence gathered from interviews with professionals in the field, evaluation of job descriptions, and internship experiences (J. Kim, Warga and Moen, 2013; Y. Kim, Addom and Stanton, 2011; Pryor and Donnelly, 2009). Several frameworks elucidate the knowledge and expertise necessary for the data management workforce (see Table 1). These frameworks, while varied by goals and structure, each attempt to illustrate the array of traits required for scientific data management.

Table 1. Desired traits in data management personnel.

Data management workforce frameworks	Source
Technical expertise, information science and subject knowledge	Englehardt, Strathmann and McCadden (2012)
Curation-centric skills and domain centric skills	Hedstrom (2012)
Knowledge, skills and abilities	Y.Kim et al. (2011)
Conduit, content and context specialists	Prior and Donnelly (2009)
Storage, archiving, preservation, curation	Choudhury (2013)
Subject knowledge, technical skills and people skills	Swan and Brown (2008)

The frameworks (Table 1) tend to consist of long lists of knowledge and expertise, without ordering their individual value. While the examples are helpful to understand the range of knowledge and expertise exhibited, they often blur the skills and individuals involved in the data workforce. It is helpful to delineate the specializations and roles professionals play in data management. For example, the amount of domain knowledge required for data management and curation appears to vary by context. The existing conceptions of the data management workforce cluster together an array of careers, education, experience and expertise.

Two of these frameworks provide contrasting ways of understanding the individuals and kinds of workforce expertise required for data management (Choudhury, 2013; Swan and Brown, 2008). The first model is person-driven and describes the kinds of workers who manage data over time. The second is data-driven and expresses the kinds of actions that take place on data over time. The complementary nature of the Swan and Brown and the Data Conservancy models help to explicate the activities and expertise involved in research data management.

Swan and Brown Skills, Role and Career Structure Model

A useful way to grasp the variety of expertise necessary for the data management workforce is to focus on persons and roles. Swan and Brown (2008), note that involvement in the Digital Curation Centre lifecycle phases (Higgins, 2008) begin with data creators or data authors (the domain scientists), followed by data scientists, then

data managers, and finally data librarians at the end of the cycle (Figure 1). While not intended to be firm distinctions, the typology differentiates between roles and work that occur at different points in time. The distinctions between data management expertise required at different phases of the lifecycle helps to differentiate the steps of the data management process.

Figure 1. Swan and Brown skills, role, and career structure model. Adapted from Swan and Brown (2008) by Jillian C. Wallis.

Data Conservancy Stack Model for Data Management

As a complement to the person-focused framework, the Data Conservancy employs a data-driven model to identify the activities data managers may need to perform. The Data Conservancy (DC)[3] is a university-based collaboration addressing the curation of research data management. The DC's Stack Model for Data Management (Figure 2) provides a useful framework for the components of data management (Choudhury, 2013). Library staff generated the model to facilitate clearer communication amongst themselves and with scientists about their data.

The model is one example of how to conceptualize the tasks associated with the data management workforce. Each of the four components represents distinct kinds of knowledge necessary for that part of data management. Data storage represents the foundation of data management, with each successive layer building upon the layers below. The DC framework begins with storage, which includes the ability to backup and restore data. It then builds to archiving, which includes a layer of protection, for example unique identifiers. Next, are data preservation actions that enable data to be shared and used. Finally, data curation encompasses additional value-adding services.

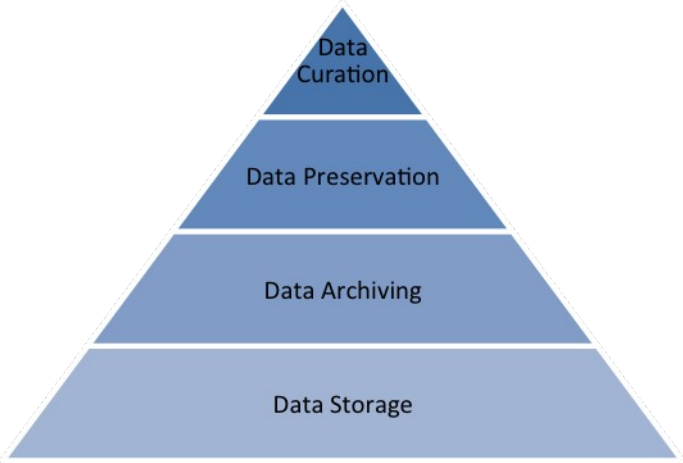

Figure 2. Data Conservancy Stack Model for Data Management. Adapted from (Choudhury, 2013) by Jillian C. Wallis.

3 Data Conservancy: http://dataconservancy.org/home

 doi:10.2218/ijdc.v9i2.336

The existing typologies begin to demonstrate that the research data management workforce involves multiple kinds of persons and activities. However, these two frameworks do not encompass the full range of persons and expertise involved in scientific research data management. The SDSS data transfer case study exposes more explicitly the persons, education and expertise involved in the research data management workforce.

The SDSS-I/II data were transferred from a national laboratory to two different university libraries. This data transfer involved four memoranda of understanding (MOU) documents. A MOU is a formalized, high-level agreement explaining a partnership between two or more institutions. The document formalizes the relationship between institutions with an agreed-upon time frame and outcomes. This article focuses on the transfer from the laboratory to one of the libraries, concentrating on the library workforce and transfer process. Future publications will examine the transfer amongst all of the institutions.

Methods

The findings reported here are based on semi-structured interviews, ethnographic participant observation, and archival studies from 2011-2013. The study population includes astronomy faculty, students, and staff from a national laboratory; and library, archive, and administrating staff from a university library. The data presented here are a subset of the UCLA KI research into the data practices of astronomers. Interview quotes are indicated with a three-digit number, which is the unique identifier of the interviewee within the larger UCLA KI dataset.

Five weeks of ethnographic participant observation were conducted at the university library and one week at the national laboratory. The fieldwork was accompanied by extensive note-taking, leading to dense descriptions of the everyday practices of the research subjects.

Semi-structured interviews were conducted with key participants in the data transfer. The interviews reported here were drawn from a set of 19 interviews with 14 key informants at the two sites involved in the SDSS data transfer described in this article. Interviews averaged about one hour in length. The interviews were semi-structured around the SDSS data transfer and the emerging workforce in data management.

We analyzed documents publicly available and those provided by key informants. Sources include public websites for the library and laboratory, formal work agreements, and the library's team collaboration spaces. Interviews were audio-recorded and transcribed. The notes and transcripts were analyzed with NVivo 9 qualitative coding software using grounded theory (Glaser and Strauss, 1967).

Findings

The SDSS data transfer described here was an experiment to move more than 130 terabytes of astronomy data from a national laboratory (hereafter referred to as the lab) to a university library (hereafter referred to as the library). The findings presented here focus on the transfer to the library. Domain scientists have completed multiple other SDSS data transfers; those transfers will be discussed in another publication.

At the transition from the SDSS-I/II to the SDSS-III project, a number of institutional collaboration members changed. The lab had been an integral member of the SDSS-I/II project, including hosting and providing online access to the scientific data. While choosing not to continue into the SDSS-III data collection process, the lab continued serving the SDSS-I/II data for public online access into the near future. The lab signed an agreement with the Astrophysical Research Consortium (ARC)[4], the formal body representing the SDSS projects, formalizing the commitment for a five-year period.

The library had previously collaborated with the SDSS on a small scale. At the end of 2008, the library and ARC also signed a MOU formalizing the library's agreement to "perform digital archiving and preservation services" for the SDSS-I/II data for the five-year period. This MOU between ARC and the library is the formalized representation of the data transfer described in this article.

The process involved transferring the SDSS archive from the laboratory to the library. There, the library performed digital archiving and preservation services. The tasks outlined in the MOU included maintaining archival-quality copies of the SDSS archive, providing value-added services to the data to "support long-term understanding of data", and serving as a secondary mirror for data, as needed. In summary, the primary task of the library was to create lasting copies of the SDSS archive to ensure the longevity of the dataset. The library therefore focused more on data preservation, rather than data access, which was addressed to differing extents by other collaboration entities.

The SDSS Data Transfer Process

The library team that received the SDSS archive included three individuals: the administrative team lead, the preservation specialist, and the system administrator. The team lead did not work with the data directly, and instead provided administrative support for the transfer. A couple of other library team members were involved infrequently during the data transfer process in either administrative capacities or to help with specific performance problems.

The library agreed to perform digital archiving and preservation services on the SDSS data. They wanted to take on the archiving and curation of the dataset, in part, to gain experience in managing large scientific datasets. A team member explained that experience is important to data management, because:

> 'This won't be the last time that this kind of data needs to be archived and we need to learn how to do it.' (Library Team Member 140).

While not highly experienced in the transfer and management of scientific datasets of this size, the library team recognized the significance of undertaking the transfer, archiving, and preservation of large quantities of scientific data. Data management is a complex task, and all datasets have different considerations. Indeed, while the library team could not have known what problems they would encounter with the SDSS archive, they knew challenges arise when working with any large scientific dataset.

4 Astrophysical Research Consortium: http://arc.apo.nmsu.edu/

 doi:10.2218/ijdc.v9i2.336

Performance Problems

Despite careful planning and face-to-face meetings, neither the lab nor library foresaw the extent of the time commitment that the full SDSS archive transfer and verification would require. The team encountered two technical performance problems:

1. A slow and unreliable data transfer over the network, and
2. A slow and unreliable verification and validation of the data, once received.

The transfer and integrity checks required 10% to 15% of the library's preservation specialist's time for the full five year MOU period. During the process, a team member explained:

> 'We'll be done by the fifth year but it would be nice to be done before that, so we're actually holding the data for some period of time by then.' (Library Team Member 140).

Despite the performance problems, the library team eventually received and verified the SDSS data for long-term management.

The library team viewed the process as transferring and validating all of the SDSS data from the laboratory to the library's servers and software. A team member explained:

> 'So, this phase of our association with [the lab] is, "how do you get the stuff [to the library, from the lab], in a validated, effective and efficient way? What do we do from our perspective to prepare it for the preservation and what is necessary to move it into the archives?"' (Library Team Member 139).

The library team would have preferred to receive the data physically on a server. However, the lab was unable to ship a server to them. Instead, it was the library's job to acquire the data over the network, which took far longer. The library had to discover or create tools to overcome the unreliable, slow process of transferring the data over the network. The necessity to procure the data over the network was the first challenge that the library team managed.

The library's second major roadblock was the process of verifying the transferred data, which included millions of pieces of information. The library server storage system was not ideal for managing millions of small files. As the system administrator explained:

> 'The [tape storage server] system as we bought it, really, was not designed to accommodate this particular dataset.' (Library Team Member 141).

Millions of pieces of data slowed the server, preventing speedy data verification or checksums. The preservation specialist explained the problem:

> 'The biggest issue has been basically shuttling the content around, so staging content back especially when there are millions of files because it just has so many files in its queue… It just really causes severe performance problems.' (Library Team Member 140).

The library team worked on overcoming the server and verification problem, and the system administrator found it necessary to take a weeklong course to learn more about the storage system. After speaking with technical assistance at the server company, the system administrator admitted:

> 'If we had known that at the beginning, I don't know if we would've… [bought a] different system.' (Library Team Member 141).

This weakness in the storage system for the SDSS dataset slowed attempts to verify the integrity of all of the data that had transferred. The library team members identified the technical and performance problems, sought solutions, and can now move into future projects with the expertise from the experience transferring the SDSS data.

Workforce Expertise

While the library team did not have much expertise handling hundreds of terabytes of scientific research data, the team did possess the skills to adjust to the experience. Despite the performance problems in transferring and verifying the data, the library team was able to evaluate the scenario and adapt to these challenges. The team overcame the performance problems because of their combined workforce expertise. The expertise included the ability to problem-solve and seek out solutions, the ability to work in a team, and the ability to write code. However, one kind of expertise missing from the library team was astronomy domain knowledge.

Problem solving

Despite receiving the SDSS archive over the slow network, and using a server that posed challenges for performing checksums to ensure verification of the data, the two library team members working directly with the data reacted to complications by active experimentation. The preservation specialist tried a number of mechanisms to reduce the time taken to transfer the data while also reducing the unwieldy nature of the process. The system administrator had insights into the particular storage system and the performance of the network. Their problem solving activities complemented one another. The system administrator focused on increasing their knowledge about the hardware, while the preservation specialist focused on discovering and enabling tools to facilitate the data transition and verification. Both team members' knowledge and expertise informed the trial and error process of managing the slow and unwieldy nature of the transfer and verification.

The library team members possessed library expertise and conducted research into the computer science aspects of the data transfer. They sought information from journals, technical reports and presentations. Combined with the preservation specialist's 25 years of computer science experience at the university in systems programming and digital library services, they employed trial and error, searching the web, and attending meetings to discover all the additional tools and techniques necessary to proceed with transferring and verifying of the SDSS archive.

Working in a team

The success of the data transfer relied on the individuals in the library working together as a team. For example, a number of occasions arose when the preservation specialist relied on the resources of others. The most common source for assistance with the data transfer and validation was within the library team itself. The preservation

 doi:10.2218/ijdc.v9i2.336

specialist, who performed most of the data handling tasks, attributes project success to the system administrator's ability to "think outside the box" (Library Team Member 140). Other library team members were called upon at times to assist with problem solving or administrative aspects of the SDSS data transfer.

Writing code

The library team wrote new code to address the performance problems that slowed the data transfer and checksum operations. Sometimes this involved creating a program to make the data transfer more smooth. For example, the preservation specialist discovered software to help package the millions of small files into 20-gigabyte parcels. Other times the coding was at a smaller level to modify an existing program or solve a simple problem. For example, the team came to understand the necessity for file lists on the laboratory-side for easier confirmation on the library-side. Sometimes the code was necessary to enhance the automation of the process and reduce the number of necessary person-hours devoted to the transfer. The preservation specialist explained that writing code was necessary to decrease the time spent "just watching screens to see how much data is being transmitted, watching to make sure the jobs hadn't died" (Library Team Member 140). The team members leveraged their existing expertise to write software customized to this particular data transfer.

Astronomy domain knowledge

The library team members drew most heavily on their library expertise and less on astronomy domain knowledge. The library team saved up their questions and contacted the SDSS astronomers only a few times a year. A team member explained that early on they had visited the laboratory and the astronomers had "been accommodating. But at the same time, they're on to different projects, they don't really have funding here anymore" (Library Team Member 140).

None of the library staff have degrees or coursework in the domain of astronomy. The lack of domain knowledge led to difficulties conducting some of the curatorial activities. For example, the CAS data are in a Microsoft SQL database format. For long-term, preservation purposes, the goal was to move the data into a database-neutral format. The library staff recognized that they do not have the expertise or ability to move the CAS into a database-neutral format without assistance from astronomers. They know enough about the data to understand that trying to move the CAS into a database neutral form would "definitely need to have more input from the scientists" (Library Team Member 140).

A second example of the importance of astronomy domain knowledge in SDSS data management occurred prior to the main data transfer. Two library team members took the time to learn about astronomy data to perform curation on a subset of SDSS data. However, they realized an astronomer would need to perform the curation work instead of them to ensure scientific accuracy. While the library team was pleased with their technical handling of the data and the ability to interoperate with international standards, they knew that their curation work would not support scientific use. A library team member explained that if the prototype had needed to work for scientific purposes, they would have needed the assistance of someone with greater domain knowledge, perhaps an astronomy PhD student (Library Team Member 143). While the curatorial work was successful in terms of the library staff's computer science expertise, they acknowledged it was imperfect from the astronomy domain perspective. The preservation specialist explained, "I don't think there is any replacement for understanding from scientists what they need" (Library Team Member 140). The

curation and value-adding aspects of the data transfer require more astronomy domain knowledge than the library team currently possesses.

Discussion

Experience in tackling performance problems with the SDSS archive prepared the library team members for managing large volumes of research data in other domains. They explained that experience was the only way to gain the appropriate data management knowledge.

The library team encountered performance problems during the SDSS data transfer from the lab to the library. The preservation specialist and the system administrator worked together closely to discover and implement solutions to the slow data transfer and verification processes. The ability of these team members to solve problems, collaborate as a team, and write custom code enabled the success of the project. However, specific curatorial add-ons, for example migrating the CAS to a database-neutral format, were not completed because the library staff possessed library and computer science expertise but did not additionally possess an astronomy education. Some degree of domain knowledge is necessary for the stewardship of scientific data.

Large Number of Tasks and the Data Conservancy Model

Analysis of the SDSS data transfer brings to light an example of the range of tasks that must take place in a large-scale data transfer between different kinds of institutions. The SDSS data management workforces have diverse educations and experience. The data-driven DC Model (Figure 2) identifies the importance of settling on a specific vocabulary to describe the tasks that individuals carry out in data management projects. The model can be used as a tool to facilitate conversations between data stakeholders.

The DC model describes four data management components. In the case of the SDSS archive, the library team states they are storing the SDSS-I/II data, as they have experience storing other university research data. Indeed, they transferred and verified the SDSS data. The library team refers to their work as primarily archival. They are working to ensure the longevity of the data, including unique identifiers. However, the library staff states that while they are making gains toward their goal of working at the data preservation level, they are not yet comfortable stating that the SDSS data are currently being preserved. Also, while they have experimented with curation, they have not implemented all of the curatorial services that they had initially hoped due to their lack of astronomy expertise. As defined by the DC model, the library stores and archives the SDSS data, however they do not yet have the expertise to support preservation and curation activities. While the library team has the experience and expertise to manage the SDSS data, they do not currently have the resources to actively perform all four components of data management as depicted in the DC Model. The SDSS data transfer itself has provided the library team with the experience to enable further data management on the SDSS archive and future datasets.

Importance of Teamwork and the Swan and Brown Model

As shown by the SDSS case study, interdisciplinary teamwork is necessary among all people in a data transfer. It is important that all sides be conscious of what demands they

 doi:10.2218/ijdc.v9i2.336

are making of the other. For example, as the data were transferred to the library, astronomy staff at the lab were unable to provide step-by-step guidance to the library. The lab team members largely had moved on to new projects. While the MOU guided the data transfer operations, the agreement did not stipulate in detail the personnel time and expertise required from the lab.

The SDSS data transfer employed all four types of workforces (data creators/authors, data scientists, data managers, and data librarians) in the Swan and Brown person-based model (Figure 1). The data creators were the astronomers who conceived of and gathered the SDSS data, including those from multiple universities and laboratories. These astronomers initiated the plan for data management and approached the actors who would perform the tasks. The data scientists were the astronomers hosting and serving the data at the lab. Data managers existed both at the lab and the library and ensured that the logistical functions of the dataset were in place. The data managers coordinated between the lab and library for the actual data distribution and are astronomers, computer scientists, or those in management roles. The data librarians are those at the receiving end of the SDSS data transfer. They perform the storage, archiving, preservation and curation at the library. The Swan and Brown model maps clearly onto the SDSS data transfer process.

The Swan and Brown model identifies four kinds of persons involved in data management. However, the model makes no indication of the importance for these different kinds of data managers to cooperate with one another, or at what kind of timescale this cooperation would take place. The case study demonstrates that all four kinds of expertise in the model were necessary for the SDSS data transfer. Since teamwork is important to a successful data transfer, the earlier the data managers can begin working with the domain scientists the better.

Conclusion

Data management encompasses multiple tasks that require many types of expertise. A large number of people, with diverse educations and work experiences, are involved in the SDSS data management including domain scientists, computer scientists, software and systems engineers, programmers, and librarians, which map onto the Swan and Brown Model. Additionally, the SDSS data management involves at least four different kinds of data actions including storage, archiving, curation, and preservation, which map onto the Data Conservancy Model.

The case study reveals that the emerging workforces include skills spread across multiple individuals. A variety of backgrounds and educational histories emerge in the data managers studied. Teamwork is necessary to bring disparate expertise together, especially between those with technical and domain education. These larger workforces involve a spectrum of people, from scientists, to librarians, to programmers, to management, to policy makers and others. These two models and the SDSS data transfer case study elucidate the diversity of workforce education, experience, and expertise necessary for scientific research data management.

Acknowledgements

Research reported here was supported by grants from the National Science Foundation (NSF) and the Alfred P. Sloan Foundation (Sloan): Knowledge and Data Transfer: the Formation of a New Workforce, NSF Award #1145888; and the Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective, Sloan Award #20113194.

We offer our appreciation to the individuals who granted us interviews and to the project team members at the university library and the national laboratory who granted us the opportunity to conduct participant-observation ethnographic fieldwork for this project.

We also thank the other members of the UCLA Knowledge Infrastructures team for their thoughtful reflections on earlier drafts: Peter T. Darch, Jillian C. Wallis, and Rebekah Cummings. Thank you to David S. Fearon, Jr. for conducting preliminary interviews.

References

Abazajian, K.N., Adelman-McCarthy, J K., Agüeros, M.A., Allam, S.S., Prieto, C.A., An, D., … Zucker, D.B. (2009). The seventh data release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series, 182*(2), 543–558. doi:10.1088/0067-0049/182/2/543

Choudhury, S. (2013, August). *Open access and data management are do-able through partnerships.* Keynote Lecture presented at the ASERL Summertime Summit: Liaison Roles in Open Access and Data Management – Equal Parts Inspiration and Perspiration, Atlanta, GA. Retrieved from https://smartech.gatech.edu/handle/1853/48696

Engelhardt, C., Strathmann, S., & McCadden, K. (2012). *Report and analysis of the survey of training needs.* Retrieved from Digital Curator Vocational Education Europe (DigCurV) website: http://www.digcur-education.org/eng/Resources/Report-and-analysis-on-the-training-needs-survey

Glaser, B.G., & Strauss, A.L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Hedstrom, M. (2012, December). *Digital data curation – Examining needs for digital data curators.* Paper presented at the Third International Conference on Cultural Heritage Online, Florence, Italy. http://93.63.166.138:8080/dspace/handle/2012/98

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134–140. doi:10.2218/ijdc.v3i1.48

Kim, J., Warga, E., & Moen, W. (2013). Competencies required for digital curation: An analysis of job advertisements. *International Journal of Digital Curation, 8*(1), 66–83. doi:10.2218/ijdc.v8i1.242

Kim, Y., Addom, B.K., & Stanton, J.M. (2011). Education for eScience professionals: Integrating data curation and cyberinfrastructure. *International Journal of Digital Curation, 6*(1), 125–138. doi:10.2218/ijdc.v6i1.177

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities, and relationships.* Retrieved from UKOLN website: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19

Pryor, G., & Donnelly, M. (2009). Skilling up to do data: Whose role, whose responsibility, whose career? *International Journal of Digital Curation, 4*(2), 158–170. doi:10.2218/ijdc.v4i2.105

Stanton, J.M., Kim, Y., Oakleaf, M., Lankes, R.D., Gandel, P., Cogburn, D., & Liddy, E. D. (2011). Education for eScience professionals: Job analysis, curriculum guidance, and program considerations. *Journal of Education for Library and Information Science, 52*(2), 79–94. Retrieved from http://www.questia.com/library/journal/1P3-2333983191/education-for-escience-professionals-job-analysis

Swan, A., & Brown, S. (2008). *Skills, role and career structure of data scientists and curators: Assessment of current practice and future needs.* Retrieved from the Jisc website: http://www.jisc.ac.uk/publications/reports/2008/dataskillscareersfinalreport.aspx

Van der Graaf, M., & Waaijers, L. (2011). *A surfboard for riding the wave: Towards a four country action programme on research data.* Retrieved from the Knowledge Exchange website: http://www.knowledge-exchange.info/surfboard