Teaching Data Creators How to Develop an OAIS-Compliant Digital Curation System: Colearning and Breakdowns in Support of Requirements Analysis

Lorraine L. Richards

ABSTRACT

This article describes a project that a team of researchers from Drexel University's College of Computing and Informatics jointly undertook with the Federal Aviation Administration's W. J. Hughes Technical Center in Atlantic City, New Jersey, to develop requirements and a prototype for a data curation repository. The repository is to be OAIS-compliant and capable of allowing FAA scientific researchers across various geographical locations to share and reuse data. An action research methodology was used, which allowed the project team to engage in a series of colearning experiences that led to a negotiated and evolving understanding of requirements. The process of colearning played a key role in allowing a concrete goal and plan to emerge from communication breakdowns.

KEY WORDS

Digital curation, Data, Education, Colearning

Researchers define digital curation in various ways.[1] Ross Harvey offered a clear portrayal that captures many of the aspects of the other definitions:

> [Digital curation is] concerned with actively managing data for as long as it continues to be of scholarly, scientific, research, administrative, and/or personal interest, with the aims of supporting reproducibility, reuse of, and adding value to that data, managing it from its point of creation until it is determined not to be useful, and ensuring its longterm accessibility, preservation, authenticity, and integrity.[2]

These tasks and goals are remarkably like those of archivists, who, in a digital world, must also begin planning for and managing data as early in their life cycle as possible to ensure that long-term preservation occurs in spite of evolving formats and rapid technological obsolescence.[3] Thus, more and more archivists are engaging in digital curation activities (whether they use that term or not), both within archival settings and as consultants supporting the research needs of scientists. Often, these scientific endeavors take place within the boundaries of academic institutions, where the nature of collection and preservation services is referred to as "research data management."[4] However, many types of organizations engage in scientific research and would benefit greatly from the consultative aid of professionals trained in digital curation and archives. A case in point is that of U.S. federal agencies dedicated to scientific research. These agencies need specialists who can help them curate their scientific data, as they are often unaware of even the basic prerequisites for managing their data in a trustworthy manner. Professionals trained to act as consultants in archives and digital curation can offer that help.

U.S. federal agencies' scientific data management activities often support the immediate research needs of agencies' scientists, but they do not support federally mandated large-scale data sharing and reuse requirements or enable long-term preservation of agencies' data. As a result, agencies are scrambling to learn how to curate their scientific data sets to meet federal mandates without sacrificing current mission-oriented research activities. This article examines an action research project on a data curation initiative at the Federal Aviation Administration's (FAA) William J. Hughes Technical Center (WJHTC) during 2013 and 2014. The WJHTC contracted with a Drexel University project team in the College of Computing and Informatics (CCI) to develop requirements and build capacity for an Open Archival Information System (OAIS)–compliant[5] digital curation system. Specifically, this paper presents findings related to teaching personnel with no archival or records management training in a nonarchival organization how to develop a research data curation system.

Problem Statement

The WJHTC uses voluminous information resources in the course of large-scale aviation research, development, testing, and evaluation.[6] While staff at the WJHTC have not previously engaged in data curation as a routine activity, they now need to develop a trustworthy repository for the center's scientific research data to meet government mandates and to facilitate data sharing for present and future mission-critical research projects.

Data sharing provides great benefits for an agency like the FAA and its individual test centers. Sharing situational awareness[7] information allows air traffic controllers, pilots, and airlines to act on an accurate and shared understanding of what is happening on the ground and within the airspace. This not only improves efficiency, but also supports safer aviation practices. Improved throughput (in the form of more timely departures and arrivals), reduced emissions, improved analysis of data (which supports better quality research), increased capacity for collaborative assessment processes, and increased interoperability of aviation service providers worldwide can all result from a successful data-sharing initiative.[8] In addition, data sharing and reuse support scientific research by allowing increased replicability and improved validity-checking among a variety of researchers.

In addition to the direct benefits of data sharing and reuse, personnel at the WJHTC are well aware of the pressures of compliance when it comes to data sharing. In February 2013, the Office of Science and Technology Policy (OSTP) issued a directive to each federal agency with over $100 million in annual conduct of research and development expenditures.[9] It required such agencies to develop plans to support increased public access to the results of research funded by the federal government, including publications in scholarly journals and digital data created during the research. The directive also required these agencies to develop an "approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research."[10] In May 2013, the president issued an Executive Order requiring agencies to collect or create information in a way that "supports downstream information processing and dissemination activities," to use open licenses and review information for "privacy, confidentiality, security, or other restrictions to release,"[11] and to report their progress on the implementation of a Cross-Agency Priority (CAP) Goal (which tracks the implementation of the Federal Open Data Policy) to the chief performance officer (CPO). In July 2013, the OSTP issued an Executive Memorandum on the science and technology priorities for the fiscal year 2015 budget.[12] The memorandum gave priority to activities that will significantly increase public access to research results, support tools,

and infrastructure that will allow U.S. science and engineering to maintain its global preeminence, as well as to activities and investments that will use data to "advance agency missions and further scientific discovery and innovation while providing appropriate privacy protections for personal data."[13]

The nature of scientific data as record within federal agencies sometimes seems murky, because retention schedules do not formally include many of the data sets created in the course of scientific activity. Although this may legally suggest that preservation of such data is unnecessary, from the perspective of archival theory, these data are most certainly archival. Archival data are those with "enduring value," according to the Society of American Archivists glossary of terms.[14] The idea of enduring value typically refers to records of a person or organization. The U.S. government defines public records as including

> all books, papers, maps, photographs, machine-readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of the data in them (44 U.S.C. 3301).

This definition does not specifically mention data generated in the course of scientific research, which may explain why such data frequently do not appear on public agency retention schedules. However, according to archival theory, a record is "data or information in a fixed form that is created or received in the course of individual or institutional activity and set aside (preserved) as evidence of that activity for future reference."[15] Data are either received by scientists as inputs into a work process or are produced as outputs of a work process. Anna Gold noted, "To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself."[16] Given that data, an integral part of scientific work, are records, a question remains: do data exhibit enduring value? This is difficult to deny, both from the point of view of social history and from the point of view of maintaining scientific validity. A cornerstone of scientific validity is the notion of reproducibility, which allows the findings of an experiment to be re-created and validated by other scientists. To reproduce an experiment requires using the data created in its original performance. To ensure continued confidence in scientific results, the preservation of data used to deliver these results is essential.

Data sharing presents some risks for researchers, however, and these risks highlight the importance of successful data curation to support data sharing and reuse. Individual data creators recognize that sharing their data sets makes it more likely that errors will be found and linked back to them, potentially

harming their reputations within an organization.[17] Carol Tenopir et al. pointed out other barriers to sharing as well, such as "privacy, concerns about future publishing opportunities, and the desire to retain exclusive rights to data that had taken many years to produce."[18] Likewise, scientists who share their data sets are concerned about the potential misuse of their data by researchers without adequate contextual knowledge to use it appropriately.[19] Gail Steinhart noted, ". . . researchers are often reluctant to share due to concerns over intellectual property, attribution, improper reuse, and lack of time, resources, and know-how to get the job done."[20] On the other hand, potential data reusers have requirements for engaging in data reuse. Specifically, without trust in preexisting data sets, scientists are unwilling to reuse data. If data do not contain appropriate contextual information about provenance and chain of custody, they will not be reused.[21]

Data curation activities enable data discovery and retrieval, maintain their quality, add value, and provide for reuse over time. Data curation activities also include authentication, management, preservation, retrieval, and representation.[22] Data curation is the key to increasing trust in data sets and the maintenance of necessary contextual information linking to the data sets. It also supports discoverability, thus helping to make data readily usable by scientists. As stated well by Michele Kimpton and Carol Minton Morris, "Advancing knowledge in all fields of research now requires the curation, collection, management, access, and long-term preservation of digital datasets that go far beyond burying a flat file on a hard drive."[23]

In May 2013, the Drexel CCI team met with WJHTC personnel to kick off a project the WJHTC had contracted with them. The WJHTC desired to develop data curation and sharing capabilities to ease the transition of data from other FAA test sites to the WJHTC. To do this, they realized that they needed to develop better controls over incoming and outgoing data, and to monitor the access and use of the data. The goals of the research contract between Drexel University and the WJHTC were to develop and to enhance existing data sets and sources, to mutually produce a "plan of action" for data scenarios the WJHTC judged important, to research and to select an appropriate information architecture, to develop a prototype system based on that architecture, and to develop requirements for a digital repository to serve the research needs of the WJHTC, Drexel University, and future users. The preservation capabilities of the repository are to be based upon the Open Archival Information System (OAIS) Reference Model.[24] These goals support the WJHTC's long-term goal of developing the knowledge and materials needed to issue a request for proposal or additional statement of work for a contractor to implement, build, and maintain an OAIS-compliant digital curation repository.

To meet these goals, the project team engaged in research specific to the data sources and scientific business processes at the center, focusing on the following questions:

- What is the scope and volume of data at the WJHTC, and how will that affect the development of a data-sharing repository?
- What is the nature and scope of the scientific research that occurs within the WJHTC?
- What domain ontologies and metadata taxonomies currently exist, and are they sufficient for the purposes of a data curation repository?
- What is the current capacity for curating and sharing data within the WJHTC organizational structure?
- What architecture will best suit the desired level and types of data sharing and reuse?
- What standards and policies will best suit the data curation and sharing requirements of the WJHTC?

These questions address key components of curation and archives activities. The data inventory is a basic tool of archival work, and questions about technological capabilities and requirements provide answers both to the curation-related need to serve current access and the preservation-related need to support long-term, trustworthy maintenance of the information.

Although the Drexel team was concerned both with developing concrete solutions for the WJHTC *and* furthering current research on data curation and sharing in nonarchival environments, WJHTC personnel were primarily interested in the tactical outcomes of the project. This influenced the team's choice of methodology.

Literature Review

A number of institutions have undertaken data curation projects.[25] However, the majority of published reports deal with research data residing within academic institutions, not within governmental agencies. Notable exceptions have been reported by Reagan Moore and colleagues, who developed the iRODS (integrated Rule-Oriented Data System) data grid, along with its precursor SRB (Storage Resource Broker), a grid system used by a variety of national governments and academic institutions that enables the distributed curation and preservation of huge data sets.[26] The extant academic literature discusses details of implementation, policy issues, and technical configuration.

While authors of current literature agree that collaborative measures are necessary,[27] they do not discuss the details of communication techniques and methods, as does this article. Dharma Akmon et al. did provide great detail regarding scientists' data practices, but did not specifically deal with the nature

of the relationship between data curators and scientists; nor did they deal with the colearning discussed in this article.[28] The majority of articles focus on relationships between scientists or the relation between scientists and their data. When inserting a third-party archival/curation consultant into the environment, however, it becomes necessary to examine the relationship between the data creators (i.e., scientists) and the consultants who help creators learn about curation requirements.

More and more academic libraries are providing curation/research data management services for their on-campus researchers, enough so that Sheila Corrall has questioned whether the increasing number of library-executed data management services represents a paradigm shift for libraries.[29] The nature of the management services varies, but according to Corrall could include such activities as applying metadata, enabling discoverability and citability, developing policies and platforms, planning data management, and providing curation toolkits. Reference and consultation services can include

> identifying datasets to meet student or faculty needs, providing access to data resources and advising researchers on current standards for organization of data in specific subject areas, in addition to help with the specific tasks of developing data management plans and more general awareness raising through creation of special websites to describe services available.[30]

These activities assume, however, that data management–trained librarians are readily available to researchers, an assumption that rarely holds true for government agency scientists, who frequently work in siloed environments and have not received the same outreach as likely have their colleagues in the academic sector. Furthermore, the articles mention virtually nothing about preservation-specific requirements for research data management, implying a strong need for an archival voice on the subject.

In environments where little to no outreach has occurred, digital curation professionals need to be prepared to educate their clients about the curation steps that must occur to manage their data in a trustworthy manner. They must also convince them that such management and curation provide value to them individually and to their organizations. Likewise, they must tailor communication techniques because, although the expertise of the curation professionals may be taken as a given, these professionals do not enjoy the same reputation as educators as librarians within a library might. Rather, they act within the boundaries of their clients' organizations and need to develop the same trust and respect as anyone entering the environment.

Methodology

As a result of the earliest conversations between the Drexel team and WJHTC workers, Drexel team members felt that the most appropriate methodology would need to take into account that the final goal was oriented toward *change*—developing a potentially new information architecture, changing work processes, changing knowledge regarding data and the sharing of data, and increasing understanding of organizational, technical, and cultural "best practices" to support data curation and sharing. Also, because the Drexel team's involvement in the data curation project was temporally limited, WJHTC workers would need to be empowered to eventually take on the data curation processes themselves. In addition, the working relationship needed to be truly collaborative. Although the Drexel team entered the project in the capacity of consultants due to their expertise in digital curation, they are not research scientists. They therefore had to develop a great deal of disciplinary and environmental understanding of the nature of the scientists' work and the culture and norms of the organization itself. Likewise, the research scientists had no previous experience with data curation, although they provided expertise in the work processes and scientific methods used in WJHTC laboratories. Because of the needs for organizational change and to ensure that all key stakeholders acted in both learning and teaching capacities, the project team chose to undertake an action research methodology.

Abraham Shani and William Pasmore defined *action research* to be

> an emergent inquiry process in which applied behavioural science knowledge is integrated with existing organizational knowledge and applied to solve real organizational problems. It is simultaneously concerned with bringing about change in organizations, in developing self-help competencies in organizational members and adding to scientific knowledge. Finally, it is an evolving process that is undertaken in a spirit of collaboration and co-inquiry.[31]

Action research requires authentic *participatory* research and action.[32] Robin McTaggart argued that this means that all participants in an action research project must take ownership "in the production of knowledge and improvement of practice."[33] Furthermore, all participants must play roles in setting the agenda of the research, participate in data collection and analysis, and have some control over the outcomes and the process of research.[34] Within this project, this occurred through regularly scheduled meetings between WJHTC team members and Drexel team members, in which the Drexel team shared deliverables for joint review. In addition, WJHTC team members allowed the Drexel project team to take part in the data generation and testing activities.

According to Kurt Lewin, an action research project typically starts when participants agree on a goal that requires the production of knowledge to

improve a workplace practice.[35] The original goal, however, is always more of a "general idea" than a goal, insofar as the desired "improvement" is usually stated as a somewhat vague achievement. For example, although "we want to develop an OAIS-compliant data curation repository for scientific data sharing and reuse" sounds very explicit on the surface, the actual language reflects concepts that all of the participants may not fully understand or not understand in the same way; "OAIS-compliant," "data curation," "repository," "data sharing," and "reuse" are all terms that require an active negotiation of meanings among all participants. Because each individual enters into the group with his or her own language usage, only through the mutual construction and reconstruction of his or her language can the group begin to develop a common understanding. The evolving, mutual construction of a group identity and language implies that the final goals are emergent, much like the research inquiry itself.[36]

Individuals occupying three abstract roles engage together in this type of project. First are the researchers, who may be academics or consulting experts. In this case, the Drexel project team enacted the role of researchers. Second, the people in the workplace have expert knowledge of the environment and work processes. Finally, the overall "action research group" is composed of all the members of the other two groups. Each individual enters into a research agreement with less than full knowledge of the ways in which the other members use language, and thus, the entire process of action research relies on being able to move from the "idea" of improvement to a clear "goal." The goal requires the development of a plan that outlines the beginning state, the desired end state, and the process steps that must be followed to reach the end state. Like the research inquiry itself, the plan and goals emerge from a "self-reflecting spiraling" of steps, typically expressed as being comprised of planning, acting, observing, and reflecting. These four steps are followed iteratively, with a series of semirepetitive steps moving the group closer to a mutually acceptable outcome, as well as a more granular understanding of the problem and its components. During this spiraling process, the goals and inquiry are continually (re)-created and increasingly become amenable to more concrete enunciation and evaluation.

Rather than assuming that any particular set of outcomes or processes reflects the "right" way to approach the change initiative desired by an organization, an action research project relies upon a qualitative, interpretive approach to discover appropriate methods, questions, and answers through collaborative communication among all the research study's participants. Action research allows the members of the research group to elicit information from each other to understand how they view the problem and its goals. Both project team members and workplace personnel share their understandings to discover the appropriate conceptual categories with which to frame the findings and results,

where "appropriate" is a jointly agreed-upon evaluative judgment. Instead of beginning with a firm theoretical research question, the action research project will typically begin with a desired operational goal that requires organizational change to achieve. The research questions typically derive from that goal, often evolving as the project participants achieve greater mutual understanding.

The research group (of Drexel and WJHTC personnel) followed a process that integrated system development planning and repository development project planning. In particular, the Drexel project team engaged in the following activities:

- Identifying and meeting with key stakeholders;
- Investigating the legal environment affecting data sharing at the laboratory and organization level;
- Conducting a data inventory;
- Conducting a system inventory and infrastructure analysis;
- Assessing technological capabilities;
- Assessing digital curation capabilities;
- Assessing data-sharing requirements at a high level (i.e., without much granularity);
- Comparing known requirements to technological and digital curation capabilities, to fine tune the overall project plan;
- Conducting scientific workflow analysis and linking workflows to data inputs and outputs and to systems;
- Assessing current metadata practices and requirements;
- Assessing privacy and security requirements;
- Selecting a pilot set of data to populate the final system prototype;
- Providing recommendations for information and system architecture; and
- Providing a prototype demo system to illustrate the potential capabilities of a viable curation/preservation system.

The project team presented detailed deliverables to WJHTC employees as they performed each set of activities. For example, when the scientific workflows were finalized, the project team met with the WJHTC scientists and liaison to validate that the workflows were correct and had been modeled at the correct degree of granularity for the purposes of the project.

Prior to entry into the WJHTC, the Drexel project team anticipated first performing a data inventory process similar to that offered by the Digital Curation Centre (DCC) and JISC Data Asset Framework (DAF).[37] Although the DAF was developed specifically to support data creation in higher education institutions rather than in government agencies, it provides a broad and highly flexible set of activities. With some content modification, these activities can serve the needs of a public-sector agency seeking to manage large-scale research

data sets. At the broadest level, the DAF presents four major steps involved in auditing research data: 1) planning a data audit; 2) identifying and classifying assets; 3) assessing management of data assets; and 4) reporting and recommendations.[38] However, the DAF *Implementation Guide* also explicitly notes that in environments where little or no data curation has occurred in the past, steps 2 and 3 may need to be reversed. The Drexel project team found this to be the case at the FAA WJHTC, where no previous data inventory had ever been performed and the data environment was highly complex.

The Drexel project team also realized during early conversations that the quantity of data at the WJHTC overall was simply too huge to handle within the course of a single curation project phase. For example, when the team received the data associated with a single experiment within the Human Factors Laboratory, the data required about two and a half terabytes of hard-drive space. As a result, the project team focused on educating WJHTC personnel about digital curation and on learning about the scientific environment to prioritize data management needs before diving into the data inventory itself. The Drexel team and WJHTC personnel jointly agreed that the project must "begin small," and only after exhibiting some project successes could other laboratories be engaged. As a result, the research group focused first on two laboratories that showed an interest in data sharing—the Human Factors Laboratory (HFL) and the Target Generation Facility (TGF). The HFL relies upon simulation data, externally and internally provided preexisting data sets, and sensor, observational, and survey/interview data created during the process of experimentation and simulation. It engages in all facets of human factors research to study and to improve safety and operations within the aerospace environment. The TGF uses external and internal simulation data to create detailed simulations that are then provided to its customers inside and outside the WJHTC. The TGF thus acts as an intermediary between data creation and collection and data use by other scientific laboratories; it creates the simulations that other laboratories use in their scientific studies. For example, if the HFL wants to conduct tests on how pilots will react under certain flight conditions, it contacts the TGF and conveys information about what types of conditions it needs, for instance, weather, topographical information, flight speed, type of aircraft, and so on. The TGF then finds and creates a simulation mapping actual flight paths (for which it has numerous data sets already) to those conditions and presents the simulation to the HFL, which then conducts its tests within the simulation environment using actual pilots.

The project team initially created a detailed survey instrument to gather information at the WJHTC. However, when the seventeen-page survey was first presented to the scientists, program managers, and other personnel with whom the team was collaborating, the respondents reported it was too lengthy to

be practical in their busy, research-intensive environment. In addition, they expressed both confusion and concern about the questions. As a result, the project team members went through a series of semistructured and unstructured personal interviews, stakeholder meetings, work shadowing, ethnographic observations, FAA training workshops, and a process of enculturation and learning of their own that allowed them to understand the FAA's aviation "geek speak" before the scientists felt comfortable sharing detailed workflow and data information. In addition, the project team engaged in internal training on the National Airspace Systems (NAS) and basic air traffic controller activities and terms. They were also introduced to the flight and air traffic control tower simulators. During the selection and reduction of test data for a scientific simulation, a Drexel team member spent a week at the WJHTC to take part in the activities that led to the final creation of simulation data for the upcoming simulation. In addition, the team inventoried the data for a single experimental study and evaluated sample research data sets. Finally, they surveyed FAA and WJHTC websites to ensure that they captured all relevant source systems that inform the laboratories' scientific experiments. Throughout these processes, the team kept WJHTC scientists and program managers involved in the project informed and received corrective feedback on the ongoing research.

Findings

This article highlights the communication mechanisms that helped the Drexel project team and WJHTC personnel increase the WJHTC's capacity to engage in digital curation of its scientific data sets and to prepare WJHTC personnel to undertake a more formal, pilot project to test the benefits of data sharing through a "real-world" curation and analysis project. At the end of the contract period, eight FAA scientists and data managers visited Drexel to discuss the overall findings and recommendations from the project. They mutually agreed to engage in further work to develop a pilot system that would allow the Human Factors Laboratory and the Aviation Safety Information Analysis and Sharing section of the Systems group (a data analysis and database management group) to build a data visualization application that uses data stored in an iRODS federated, OAIS-compliant pilot repository. Before this final agreement could occur, however, the project team and WJHTC personnel had to engage in a long process of communication and colearning.

As mentioned earlier, iRODS is a computing grid system that allows federated management of data. It gives each "member" of the federation the ability to manage his or her own data sets and upload them into the centralized grid repository. Other members will then have access to the data sets and can themselves create logical names for the data elements that fit their own specific

needs, while the iRODS system maintains a record of the original data set and links it to any new data sets created by reusers. This allows site-specific ownership and management of data while still giving external sites the opportunity to use the data sets as they prefer—a highly desirable capability in an environment where scientists are accustomed to managing and controlling their own data. Because the WJHTC hopes to develop further capacity, it needs a system that allows multiple sites to engage in a centralized repository without requiring centralized management of original data sets.

The Drexel project team provided a variety of deliverables: a data inventory, a system- and data-source inventory, use cases/scenarios, an architecture assessment and recommendation, a legal/environmental scan, an analysis of retention schedules, a review of current WJHTC data dictionaries with an introduction to an approach for developing a taxonomy, a scientific workflow, a data flow and map, an iRODS data grid presentation, and a prototype curation system, a presentation, and a demo. They also recommended that the WJHTC use a federated architecture to enable local management of data while simultaneously supporting data sharing across a variety of geographic domains. Although the deliverables were specified in the contract, during the various presentations of deliverables, it became clear that the WJHTC personnel were *really* waiting for one primary deliverable: the prototype system. They viewed this system as providing concrete "proof of concept" for a curation system and wanted to use it to show their colleagues within the WJHTC the analytical possibilities provided by a data curation system. From the beginning to the end of the project, the prototype system evolved from being one of many deliverables to the *pièce de résistance* that would both prove to others the value of the project and provide evidence that a future phase, involving the development of a pilot curation system storing actual research data, would be worth the resources needed to construct it.

As mentioned earlier in this article, before the Drexel project team was provided access to data sets and allowed to shadow scientists in the course of their research work, the project team needed to build trust and to provide evidence that they had the knowledge necessary to interpret shared information accurately. In addition, WJHTC personnel had to be shown that they were equal partners during every step of the project; they did not want consultants to come in and "give them a solution." Rather, the project team had to provide detailed information to the scientists regarding why each component activity was necessary, how it fit into the overall goals of the project, and how the goals of the project would benefit the WJHTC. For example, in a very early meeting, the project team expressed a desire to model the scientific workflows. One scientist appeared confused about why workflows needed to be modeled just to build requirements for a data repository. He queried, "Is this project about

workflows or is it about data?" When the team explained that one could not assess the value or meaning of the data without understanding the workflows in which they are used and the value of those workflows to organizational goals, he evinced some surprise. He brought up the question two more times over the course of the first three months and did not make himself available for further interviews or discussion of data until he fully understood how and why the value of the data depends upon the value of the research that leads to the data generation. Furthermore, he (and other scientists) insisted that the Drexel project team evince a satisfactory level of understanding of the WJHTC environment and domain knowledge before he would provide access to or answers about processes and data. After the project team had undergone training in his laboratory's processes, source systems, and typical research topics and were able to "translate" their curation-specific language into discourse consistent with the scientist's own research terminology and scientific understanding, he appeared to have an "Aha" moment and thereafter was willing to offer more of his very busy time.

Both of these examples illustrate the role of action research in a project of this nature. To make recommendations and determine requirements for a curation system, first steps involved a data inventory and scientific workflow analysis to track data flows and creation. When the notion of a data inventory was first introduced, however, scientists within the WJHTC were perplexed, both about what a data inventory involved and, furthermore, why it would be helpful in determining curation requirements. Several iterations of discussion needed to take place to show the relationship between the data inventory and the scientific workflows. This required the Drexel project team to be able to discuss the role of the workflows in the overall process of scientific knowledge creation and to be able to indicate that understanding the sources of data available to the scientists would enable them to map the data, in all of its iterations, to the scientific workflows themselves—a task that would allow scientists to begin to see where automated processing might be possible and helpful, and where manual processing would still be necessary. In addition, by learning how the data are tested, the Drexel team could better understand why some manual processing will always be necessary when test data are created. A mutual language was created through the resolution of concerns on both sides of the conversations.

The question of how the curation system would benefit the scientists arose again and again throughout the project as WJHTC personnel tried to reconcile the detailed requirements of a curation and preservation system with their primary interest in being able to access large amounts of data for analytical purposes. That is, WJHTC personnel were not particularly interested in data curation and preservation per se. Rather, they were interested in having a system to store data and to allow access to them in a manner that would facilitate their

complex scientific workflows and analyses. This true goal was not immediately obvious to the Drexel project team in the early months, but became so after a series of "breakdowns" in communication during which WJHTC personnel continued to ask what the benefits of curating the data would be. The Drexel team continued to answer the question by referring to compliance concerns, quality of data, previous research showing that scientists would not reuse data without confidence in its authenticity, reliability, usability, and so on and were somewhat confused as to why the question kept arising again and again. Still, the question of "what good is this?" continued to arise, even months into the project. Finally, a WJHTC participant asked, "We know we need good quality data and we need to be able to find it, but how does this system help us with our analyses?" This question led to a discussion that revealed that when the WJHTC personnel had asked for a digital curation repository, they were *not* asking for a curation repository as much as for a trustworthy data store that would directly support the analysis and visualization of data. In other words, they wanted a lot more than what they had contracted for. The Drexel team then needed to talk specifically about the technological requirements that would allow a curation system to feed authentic, reliable, and easily discoverable information to their analytical systems. The project team had to explicitly show that while the system could not, in and of itself, perform analysis, it was nonetheless a necessary precondition for analysis that could be trusted. The WJHTC requested highly detailed backend information about the iRODS prototype system being built and an iRODS curation system in general: what its technical specifications were, with what systems it could interface, how it could be configured, and so on. The project team set up a meeting between the original iRODS developers at the University of North Carolina, Chapel Hill, and WJHTC personnel on the project. After all of their technical questions were answered, the scientists showed a great deal of enthusiasm for engaging in a pilot project. By discussing the specific types of analyses that the scientists desired to perform and the specific functionality that the iRODS system could provide, the WJHTC scientists were able to fit their high-level ideas about data curation to their detailed understanding of the analytical components of research and the technological requirements for supporting their analyses.

At the beginning of the project, both parties believed that they shared a set of goals because they cowrote and cosigned the contract. However, as mentioned earlier, during the requirements analysis that followed contracting, the Drexel team realized that although both parties had agreed to jointly develop capacity and requirements for an OAIS-compliant repository, they had quite different expectations about what this meant in terms of functionality, collaborative activities, and final product—in spite of having gone through numerous status checkpoints at which all engaged *agreed* that they shared a mutual

understanding of project goals. The apparent early agreement between the project team and WJHTC workers about the goals of the project was continually revised and re-created as a result of those moments when WJHTC returned again to the question, "But what value will this system provide us?" With each apparent breakdown in communication about the value of the curation system and the need for particular deliverables (e.g., data inventory, scientific workflows, data testing), further discussion and clarification of the expectations of each party occurred. This clarification allowed both the project team and the WJHTC to develop a better mutual and more granular understanding of the detailed plans needed to develop a usable digital curation system. In short, the initial goal could be described as a bit of a fantasy; only through collaborative communication around the two groups' unique and separate knowledge could an increasingly concrete set of steps and requirements emerge. As the project team and WJHTC personnel asked questions and discovered gaps in their mutual understanding, these gaps, or breakdowns, in understanding allowed all involved to suddenly realize that what they had thought was a mutual understanding was not. (For example, when a data inventory was presented, the immediate response from the WJHTC team was "This is all very well and good, but what does it actually do for us? What is the value of this?" A single explanation of the role of a particular step was typically not sufficient. Several iterations of explanation about each step in the process were necessary to embed a firm understanding of why each needed to occur and why each deliverable was ultimately valuable for the project as a whole.) As a result of the breakdowns, the participants had to continue negotiating meanings. Not until a "final" goal acceptable to both parties' understandings of what the potential data curation system could provide was identified did both parties agree that the project was indeed "complete." This "completion" point was achieved when the project team was able to provide much, but not all, of the originally conceived deliverables in the contract.[39] More important, it occurred when WJHTC personnel believed that the project team fully understood the scientists' requirements and were able to communicate those requirements through the specification and presentation of a prototype system that the scientists believed to be technologically capable of meeting the WJHTC data analysis and visualization desires. Colearning allowed the project team and WJHTC personnel to converge on a final goal and set of requirements that they could both agree would provide the necessary value—a system capable of supporting analysis for WJHTC workers and of providing trustworthy curation and preservation for the Drexel project team.

The notion of *colearning* is as much about trust building as it is about education in a consultative curation environment. When a curation and archives expert enters a noncuration environment, he or she must take a series of steps to assimilate into the work environment. While it is already a commonplace

that some domain-specific knowledge is necessary to curate research data, it is also the case that a consultant must integrate into an environment as if he or she were a member of the organization, while still maintaining his or her ultimate "otherness" throughout the project. The more the consultant can generate an atmosphere in which everyone feels comfortable expressing his or her *lack* of knowledge of particulars, the greater the ability to create the feeling of equality that action research projects require. All individuals can be aware of what they don't know while recognizing that they still have knowledge integral to the project as a whole. Because consultants are often initially believed to be the experts that will single-handedly solve a problem, this is not always a given. By engaging in colearning, they illustrate their respect for the organizational personnel's knowledge and remove the divide between expert and nonexpert. All participants are consciously aware of each other as experts.

Conclusion

It would be easy to assume that an archival/curation consultant's goal is to provide expertise and training to staff in the nonarchival organization that has hired him or her to develop a curation system. What this project has shown, however, is that providing expertise, although necessary, is no more important than building shared commitment, trust, and a willingness to engage in colearning with clients to create a shared language to express value. Also, the project has shown that for staff in a nonarchival organization, the value of preservation is not immediately obvious; knowledge of this value must be developed slowly and with a continual focus on the ultimate use to which the system will be put. This is, of course, quite consistent with the final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF). The final report[40] remarked that when making a case for preservation, one must always tie it to access because preservation is a *derived good*. That is, people do not want preservation just for its own sake; they are willing to engage in preservation to ensure they can access authentic and reliable information in the future. This argument has one further logical step that the BRTF's final report does not explicitly state, however: access itself is *also* often a derived good—accessing information provides the user with knowledge in the service of his or her final productive goals. In the case of the WJHTC, the ability to reuse previously created data sets had to support the scientists' ability to engage in high-quality analysis of scientific questions. And, to support this ability, the scientists needed to directly understand how a curation system would fit into their entire system of servers, networks, and applications to add value to those "means of scientific research production."

Information systems researcher Susan Gasson remarked that the analysis of system requirements involves a series of negotiated understandings among all participants engaged in the design of boundary-spanning information systems.[41] By definition, such systems serve multiple stakeholders from a variety of knowledge domains, representatives of which must be involved in defining both scope and requirements for the new information system. However, the various stakeholders only partially understand the overall requirements and the whole set of business processes affected. Gasson explored the dichotomy between an ideal "symmetry of ignorance," which recognizes that knowledge about the business processes and system requirements is distributed and incomplete from the point of view of any single participant. She also suggested that communication breakdowns in design processes may be *productive* during the design phase. During such breakdowns, redefining project deliverable focus can bring about a productive form of collective breakdown, which leads to greater individual involvement in group decision-making. Her paper suggested that the traditional IT requirements strategy of coming to an early lock-in of the planned form and functions that inform system design may therefore be detrimental to project success. This is quite consistent with the findings of the Drexel-WJHTC data-sharing and curation project.

The Drexel project team and WJHTC workers underwent a process of colearning that involved mutual negotiation of meaning to develop a mutual understanding of curation and sharing requirements. Initially, the Drexel project team attempted to highlight the benefits of a data curation system in terms of the improved data quality, easier data discovery, and the ability to satisfy federal mandates. Although WJHTC personnel recognized these benefits would provide some value, it was not the value they felt to be most essential to them. Through a continued and increasingly granular negotiation of the meaning of "digital curation" and the expected functionality of a digital curation system, the required final goal of the project evolved from one in which the two parties had different concepts of what comprised a "completed project," to one in which both parties not only felt they had a mutual understanding, but were able to express that understanding through the concrete requirements that emerged during their colearning.

This finding is important for archivally trained experts who wish to venture into new areas of consultation. More and more preservation- and curation-aware expertise will be needed as modern science progresses and ever-larger data sets need to be managed. Traditional techniques of archival consultation (as practiced within an archives by an individual who assumes the materials will enter his or her own archives) will be coupled with techniques that recognize that some types of materials will not enter an external archives but will still need to be managed in a trustworthy manner within the organization or community

of data creation. In these cases, it will be crucial for the archival consultant to understand more about how to integrate into a nonarchival organization for the duration of a development project.

Notes

[1] Daisy Abbott, "What Is Digital Curation?," *DCC Briefing Papers: Introduction to Curation* (Edinburgh: Digital Curation Centre, 2008), Handle: 1842/3362, http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation#sthash.a1jt5jkq.dpuf; C. Dallas, "An Agency-Oriented Approach to Digital Curation Theory and Practice," in *The International Cultural Heritage Informatics Meeting Proceedings*, ed. J. Trant and D. Bearman, *Archives and Museum Informatics* (Toronto: DCC, 2007); Ross Harvey, *Digital Curation: A How-to-do-it Manual* (London: Facet. JISC, 2010); C. A. Lee and H. Tibbo, "Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills," *Archivaria* 72 (2011): 123–79; P. Lord and A. Macdonald, *e-Science Curation Report—Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision* (Twickenham, Eng.: JISC, 2003); Maureen Pennock, Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information, *Library and Archives* 1 (2007) *Journal*, http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf; Elizabeth Yakel, "Digital Curation," *OCLC Systems and Services: International Digital Library Perspectives* 23, no. 4 (2007): 335–40, doi: http://dx.doi.org/10.1108/10650750710831466.

[2] Harvey, *Digital Curation: A How-to-do-it Manual*.

[3] D. Akmon, A. Zimmerman, M. Daniels, and M. Hedstrom, "The Application of Archival Concepts to a Data-Intensive Environment: Working with Scientists to Understand Data Management and Preservation Needs," *Archival Science* 11 (2011): 329–48, doi: http://dx.doi.org/10.1007/s10502-011-9151-4

[4] K. G. Akers, F. C. Sferdean, N. H. Nicholls, and J. A. Green, "Building Support for Research Data Management: Biographies of Eight Research Universities," *International Journal of Digital Curation* 9, no. 2 (2014): 171–91, doi: http://dx.doi.org/10.2218/ijdc.v9i2.327; M. H. Cragin, W. J. MacMullen, J. Wallis, A. Zimmerman, and A. Gold, *Managing Scientific Data for Long-term Access and Use*, unpublished manuscript, https://deepblue.lib.umich.edu/bitstream/handle/2027.42/57315/14504301123_ftp.pdf;sequence=1; P. Doorn, and H. Tjalsma, "Introduction: Archiving Research Data," *Archival Science* 7 (2007): 1–20, doi: http://dx.doi.org/10.1007/s10502-007-9054-6; D. Minor, M. Critchlow, A. Hutt, D. Fleming, M. L. Bergstrom, and D. Sutton, "Research Data Curation: Lessons Learned," *International Journal of Digital Curation* 9, no. 1 (2014): 220–30, doi: http://dx.doi.org/10.2218/ijdc.v9i1.313; A. H. Poole, "How Has Your Science Data Grown? Digital Curation and the Human Factor: A Critical Literature Review," *Archival Science* 15 (2015): 101–39, doi: http://dx.doi.org/10.1007/s10502-014-9236-y.

[5] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS), Magenta Book, Issue 2* (Washington, D.C.: CCSDS Secretariat, 2012).

[6] "About the Technical Center," Federal Aviation Administration, http://www.faa.gov/about/office_org/headquarters_offices/ang/offices/tc/about/.

[7] Situational awareness can be understood as individuals' "awareness of their surroundings, the meaning of these surroundings, a prediction of what these surroundings will mean in the future, and then using this information to act," "Situational Awareness," Aviation Knowledge, http://aviationknowledge.wikidot.com/aviation:situational-awareness.

[8] "NextGen: Data Sharing Helps Airlines Reduce Delays," Federal Aviation Administration (June 2012), http://www.faa.gov/nextgen/snapshots/stories/?slide=9.

[9] Office of Science and Technology Policy (OSTP), *Increasing Access to the Results of Federally Funded Scientific Research* (Washington, D.C.: Executive Office of the President, 2013), https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[10] OSTP, *Increasing Access*, 2.

[11] Office of Management and Budget (OMB), "Open Data Policy—Managing Information as an Asset" (memorandum for the heads of executive departments and agencies) (Washington, D.C.:

Executive Office of the President, 2013), http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

[12] OSTP, "Science and Technology Priorities for the FY 2015 Budget" (memorandum for the heads of executive departments and agencies) (Washington, D.C.: Executive Office of the President, 2013), http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-16.pdf.

[13] OSTP, "Science and Technology Priorities," 4.

[14] Richard Pearce-Moses, *Glossary of Archival and Records Terminology*, s.v. "archival," Society of American Archivists, http://www2.archivists.org/glossary/terms/a/archival#.VyC6J1aDFBc.

[15] Pearce-Moses, *Glossary*, s.v. "record," SAA, http://www2.archivists.org/glossary/terms/r/record#.VyC-V1aDFBc.

[16] Ann Gold, "Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians," *D-Lib Magazine* 13 (September/October 2007), http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

[17] Anonymous FAA researcher, interview with author, 2014.

[18] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE* 6 (2011): (6 e21101) 2, doi: http://dx.doi.org/10.1371/journal.pone.0021101.

[19] Anonymous FAA researcher, 2014.

[20] G. Steinhart, "An Institutional Perspective on Data Curation Services: A View from Cornell University," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, Ind.: Purdue University, 2014), 304.

[21] I. M. Faniel and T. E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *Computer Supported Cooperative Work* 3, nos. 3–4 (2010): 355–75. Also, D. Roure et al., "Towards Open Science: The myExperiment Approach," *Concurrency and Computation: Practice and Experience* 22 (2010): 2335–53.

[22] "Data Curation: What Is Data Curation?," Council on Library and Information Resources (CLIR) (2014), http://www.clir.org/initiatives-partnerships/data-curation.

[23] M. Kimpton and C. M. Morris, "Managing and Archiving Research Data: Local Repository and Cloud-Based Practices," in *Research Data Management*, ed. Joyce M. Ray (West Lafayette, Ind., Purdue University Press, 2014).

[24] The Tech Center personnel were unaware of the OAIS Reference Model when it was introduced to them. For a variety of reasons, however, it was found to be a worthwhile model to use to direct development efforts. From the point of view of FAA workers, having been initially designed for NASA, its federal data practices provided a positive caché. The Drexel project team used the model to explain the processes required to enable trustworthy long-term preservation of data as well, so Tech Center personnel found that it provided an authoritative pointer during all phases of the project. The Drexel project team proposed attempting to follow OAIS recommendations throughout the project to begin teaching Tech Center personnel a way of thinking about the preservation aspects of digital curation and to begin developing curation readiness in what was otherwise an organization with little to no digital curation readiness.

[25] Akers, *Building Support;* Neil Beagrie, Julia Chruszcz, and Brian Lavoie, *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities* (np: JISC, 2008), http://www.webarchive.org.uk/wayback/archive/20140615221657/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf; Neil Beagrie, Brian Lavoie, and Matthew Woollard, *Keeping Research Data Safe 2* (np: JISC, 2010), http://www.webarchive.org.uk/wayback/archive/20140615221405/http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf; D. Scott Brandt, "Purdue University Research Repository: Collaborations in Data Management," in *Research Data Management*, 325–46; Sheila Corrall, "Roles and Responsibilities: Libraries, Librarians and Data," in *Managing Research Data*, ed. Graham Pryor (London: Facet, 2012), 105–33; Cragin, *Managing Scientific Data*; Minor, *Research Data Curation*; G. Henry, "Data Curation for the Humanities: Perspective from Rice University," in *Research Data Management*, 347–84; Kimpton, *Managing and Archiving Research Data*; Daniel Noonan and Tamar Chute, "Data Curation and the University Archives," *The American Archivist* 77 (Spring/Summer 2014): 201–40; Steinhart, *An Institutional Perspective*; B. Westra, "Developing Data Management Services for Researchers at the University of Oregon," in *Research Data Management*, 375–94.

[26] R. Moore, "Towards a Theory of Digital Preservation," *International Journal of Digital Curation* 3, no. 1 (2008): 63–75; Reagan W. Moore, Arcot Rajasekar, Michael Wan, Wayne Schroeder, Antoine de Torcy, Sheau-Yen Chen, Mike Conway, and Hao Xu "Concepts in Distributed Data Management or History of the DICE Group," http://irods.org/wp-content/uploads/2015/01/DICE-History.pdf.

[27] Brandt, *Purdue University Research Repository*; Tracey P. Lauriault, Barbara L. Craig, D. R. Fraser Taylor, and Peter L. Pulsifer," Today's Data Are Part of Tomorrow's Research: Archival Issues in the Sciences," *Archivaria* 64 (Fall 2007): 123–79.

[28] Akmon, *The Application of Archival Concepts*.

[29] Corrall, *Roles and Responsibilities*.

[30] Corrall, *Roles and Responsibilities,* 110.

[31] A. B. Shani and W. A. Pasmore, "Organization Inquiry: Towards a Model of the Action Research Process," in *Fundamentals of Organization Development*, vol. 1, ed. D. Coghlan and A. B. Shani (London: SAGE, 2010), 439.

[32] Robin McTaggart, "Guiding Principles for Participatory Action Research," in *Participatory Action Research: International Contexts and Consequences*, ed. Robin McTaggart (Albany: State University of New York, 1997), 27. See also Janet Masters, "The History of Action Research," in *Action Research Electronic Reader*, ed. I. Hughes (Sydney, Aus.: The University of Sydney, 1995).

[33] McTaggart, "Guiding Principles," 28.

[34] McTaggart, "Guiding Principles," 29.

[35] Kurt Lewin, "Frontiers in Group Dynamics: II. Channels of Group Life; Social Planning and Action Research," *Human Relations* 1 (1947): 143–53, doi: http://dx.doi.org/10.1177/001872674700100201.

[36] McTaggart, "Guiding Principles," 33.

[37] DCC, *Data Asset Framework. Four Steps to Effective Data Management*, http://www.data-audit.eu/.

[38] Digital Curation Centre (DCC) and JISC, *Data Asset Framework Implementation Guide* (2009), http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf.

[39] For example, less work on taxonomies was performed than the Drexel team had originally hoped for. However, the extensive work that would be necessary to update data dictionaries and come up with taxonomies was eventually judged to be more time intensive than the project contract would allow.

[40] Blue Ribbon Task Force on the Sustainability of Digital Preservation and Access (BRTF), *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information* (San Diego: San Diego Supercomputer Center, 2010), 24–25.

[41] Susan Gasson, "Progress and Breakdowns in Early Requirements Definition for Boundary-Spanning Information Systems" (paper read at Information Conference on Information Systems [ICIS], Montréal, Québec, 2007).

ABOUT THE AUTHOR

Lorraine L. Richards (formerly Eakin) assistant professor at the College of Computing and Informatics at Drexel University, performs research in the areas of digital curation and digital preservation. She is interested in the impact of emerging technologies on theory and practice in archives and records management and in developing curation capabilities in noncuratorial organizations. She also performs research about the costs and benefits of digital curation. Dr. Richards teaches in the areas of digital curation, digital preservation, electronic records management, and archives.