# Library-Mediated Collaborations: Data Curation at the National Agricultural Library

Ricardo L. Punzalan and Adam Kriesberg

## Abstract

To effectively support research activities and data stewardship, library and information professionals engage in collaborative projects that involve diverse disciplinary and institutional partnerships. While this idea is stressed in existing literature, the different ways in which librarians and domain experts working in library and information organizations engage in collaboration is rarely made explicit. This paper proposes the term *library-mediated collaborations* to capture the ways in which library and information professionals perform actions that facilitate, coordinate, and even create opportunities for multiple stakeholders to leverage their resources and expertise in data curation. By *mediation*, the paper refers to the active and critical involvement of institutional actors, in this case information professionals in a national library, in ensuring the creation and execution of a project over a period of time. The paper discusses the various manifestations of library-mediated collaborations in four data curation projects currently taking place at the National Agricultural Library (NAL). A national library located within the United States Department of Agriculture (USDA), NAL has long supported the preservation of and access to agricultural information. The paper concludes by identifying important questions that information professionals may consider asking when they participate in collaborative data curation projects.

## Introduction

Libraries and information professionals play a role in collaborative data curation by participating in scientific research projects or stewarding research data for long-term access (Borgman et al., 2015; Borgman, Wallis,

LIBRARY TRENDS, Vol. 65, No. 3, 2017 ("Collaboration in Agricultural Librarianship and Information Work," edited by Sarah C. Williams and Christine D'Arpa), pp. 429–447. 

& Enyedy, 2007; Gold, 2010; Heidorn, 2011; Palmer, 1996). However, few studies examine this responsibility in depth. Current discussions around data curation in libraries tend to focus more on creating services to support the needs of patrons, with data curation seen as an extension of traditional repository services (Mayernik et al., 2012; Toups & Hughes, 2013). In this paper we present the findings of our current research project that examines the roles and contributions of information professionals in collaborative projects that advance data curation goals. In partnership with the United States Department of Agriculture's (USDA) National Agricultural Library (NAL), we examine the various ways that its Knowledge Services Division (KSD) engages in data curation collaborations in agricultural sciences.

Over the past few years KSD has taken a leadership role in a number of projects designed to facilitate the access, preservation, and sharing of data within an agricultural research context. Within the division, one of four major units at NAL, a series of initiatives focused on the management and curation of research data are currently under way (Parr, 2016). We derive much of the insight presented in this paper by focusing our analysis on four of these ongoing initiatives. The i5k Workspace@NAL is one such project, which seeks to house 5,000 insect genomes that were assembled from dispersed laboratories that would not otherwise have the ability to host genomic data. NAL provides data services to these contributors, including the creation of easily searchable landing pages for individual species and annotation tools (Poelchau et al., 2015). Another digital project managed by NAL is the Life Cycle Assessment (LCA) Commons, a partnership involving the library, other federal agencies, and academics. This site aggregates the many datasets needed for this complex analysis and makes them more accessible by academic and industrial researchers (Lohrey, 2014; USDA, n.d.-a). The Long-Term Agroecosystem Research (LTAR) project is an access portal for data created by eighteen sites conducting research on sustainability in agricultural production. Finally, NAL is currently developing Ag Data Commons, a repository that provides access to a range of agricultural data from research supported by the USDA. This project, still in development, will continue to grow through 2017 (GODAN, 2015; Parr, 2016).

We noted that within KSD, information professionals with specialized domain expertise assume a variety of roles, level of involvement, and commitment in collaborative projects. The roles they assume and the resources they utilize vary depending on the object, product, goals, and outcomes of those collaborations. However, despite the variation in roles and responsibilities, information professionals are key to guaranteeing that projects move forward and targets are met. Hence, we propose to use *library-mediated collaborations* to represent the critical responsibility that information professionals play in creating, maintaining, and facilitating collaborative

efforts, and to capture the wide-ranging possibilities and active roles of libraries and information professionals in such endeavors.

This paper aims to identify examples of library-mediated collaborations in agricultural data curation by focusing on the four KSD projects mentioned above. We begin by briefly describing current perspectives on the role of libraries and information professionals in collaborative data curation in key literature. We then proceed to describe the data gathering methods employed for this research before presenting and discussing KSD's collaborative efforts in each of the four projects. We conclude by identifying important questions that information professionals may consider asking when they collaborate in data curation projects.

## Library-Mediated Collaborations

Understanding how collaboration works—its mechanisms for achieving meaningful outcomes, changing organizational structures and professional standards and practices, and creating new fields of study and disciplines—has been the subject of research in diverse areas that include organizational studies, cultural heritage, and social studies of science (Hardy, Philips, & Lawrence, 2003; Hedstrom & King, 2007; Lawrence, Hardy, & Philips, 2002; Olson et al., 2008; Schrum, Genuth, & Chompalov, 2015; Trant, 2009; Wood & Gray, 1991). Michael Schrage (1990, p. 40) defines *collaboration* as a "process of shared creation: two or more [groups] . . . interacting to create a shared understanding that none had previously possessed or could have come to on their own." Bringing the concept into the library, archives, and museum (LAM) realm, Diane Zorich, Gunter Waibel, and Ricky Erway (2008, p. 10) characterize collaboration as "a process in which two or more groups work together toward a common goal by sharing expertise, information, and resources." In their *Beyond the Silos of the LAMs* report, Zorich et al. note several activities that fall under *collaboration* in the cultural heritage field, highlighting a continuum of activities that begins at the point of contact, to cooperation, to coordination, to collaboration, to convergence. Thus collaboration in cultural heritage organizations involves a broad continuum of interrelated activities that produce a variety of outcomes for a variety of stakeholders.

The subject of collaboration occupies a significant area of discussion in the study of large-scale science where data curation has become a primary concern (Borgman et al., 2015). There are available, extensive literature emphasizing the collaborative nature of scientific research practice in specific fields and subfields of science (for example, physics, zoology, or astronomy) (Borgman, Wallis, & Mayernik, 2012; Edwards, Mayernik, Bacheller, Bowker, & Borgman, 2011). Significant attention is also given on collaboration in content cocreation using common online crowdsourcing platforms (Rotman, Procita, Hansen, Parr, & Preece, 2012). These are all important threads in the growing literature that underscores the col-

laborative nature of scientific research and discovery; they also highlight the important role of library and information professionals in research data management and preservation (Corti, Van den Eynden, Bishop, & Woollard, 2014).

In this paper we focus on the perspectives of library and information professionals on their role in collaborative efforts around data curation. In the data curation literature consulted for this project, we paid particular attention to what scholars in the field identify as the role of information professionals and libraries in collaborative data curation. In brief, authors collectively acknowledge that effective data stewardship requires library and information professionals to actively participate in collaborative projects that often involve diverse disciplinary and institutional players (NRC, 2015). Furthermore, they identify the library as an ideal site for hosting collaborative endeavors. Some sources call upon librarians to actively rethink their roles in the research process and gain the necessary resources and skills in order to meaningfully engage in collaboration (Weber, Palmer, & Chao, 2012).

Scholars of data curation note the emerging role of information professionals in contributing more actively in the research process. Librarians are called upon to participate in, and not only to provide support for, sponsored research projects (Garritano & Carlson, 2009). According to Anna Gold (2007),

> [The] key [to] libraries or librarians playing more "upstream" roles in data science is their ability to position themselves as partners in research. By collaborating closely, and early, in the research process, librarians may become involved in creating data curation prototypes, or otherwise supporting the use of documentation, practices, or standards that will assure the longevity of the data downstream. (n.p.)

Data curation is a complex endeavor that often involves multiple institutional actors and disciplinary expertise. There is an increasing demand for librarians and information professionals to collaborate with faculty and researchers on a range of projects, such as data analysis, integration, and visualization. Coordinating the acquisition, representation, and preservation of the research data is another area that requires collaboration. According to Weber, Palmer, and Chao (2012):

> Data curation is not, however, an activity that will be isolated in libraries or in any one type of institution or organization. It is a collaborative enterprise that requires the application of a range of data expertise, beginning with research planning and extending through phases of long-term stewardship and the reuse of data for new purposes. Information professionals that specialize in curating research data must be active in many kinds of organizations where data are generated and used, as well as traditional venues like libraries, archives, and data centers. Moreover, knowledge, skills, and principles from information science and archival science, as well as other cognate areas, are critical

> to the development of data curation expertise needed for a research data workforce. (p. 306)

The quote above offers another perspective on the role of information professionals in data curation: that of coordinating interinstitutional and interdisciplinary data curation work. Furthermore, Hedstrom (2012, p. 2) notes that data curation "is a shared responsibility organized around a life cycle model where data producers play a role in data curation while data are collected and actively used and archives or repositories assume responsibility for curation of data once it has become inactive but needs to be maintained for future use by a designated community." Within this context librarians are expected to coordinate with various actors involved in the creation, dissemination, and use of data. In many libraries this involves the creation of guidelines in the ingest and acquisition of research data, hosting data in institutional repositories, and providing data support services, to name just a few (Toups & Hughes, 2013). In an environment where the library is but one of the many other units providing data-management support to researchers, information professionals must learn how to work with their peers who contribute in data curation efforts outside of the library or archival setting (Wright, Whitmire, Zilinski, & Minor, 2014).

It is evident that among the expected roles of information professionals in data curation is the facilitation of a range of collaborative activities (Latham & Poe, 2012; Ray, 2014). These may involve diverse expertise and multiple institutions at various stages of the life cycle of data (Karasti, Baker, & Halkola, 2006); it also requires the sharing of expertise around policy development and infrastructure creation (Macdonald & Martinez-Uribe, 2010).

Our use of library-mediated collaborations aims to capture the ways in which libraries perform collaborative actions to become and remain involved in data curation projects. By *mediation*, we mean the sustained involvement of an information organization in the management, curation, preservation, and project-coordination activities around research data. We argue that library-mediated collaborations, as a concept, reflect how information professionals create, shape, and participate in collaborative relationships involving the curation and management of agricultural research data.

## Methods

Collaboration is necessary for the creation of data curation infrastructures and the delivery of data curation services to various stakeholders. However, the roles of libraries and information professionals in collaboration-driven data curation have not been fully studied. Our study asks the following question: How do information professionals at NAL mediate data curation collaborations? This project benefits from a multiyear cooperative agree-

ment between the authors and NAL to support and study the development of data curation initiatives at the library. As the site of the library's data curation activities, KSD has been the focus of our work at the library to date. During the initial phase of the agreement our interactions with KSD have been broad, allowing us to engage with and understand each of the main projects currently ongoing across the division.

This paper presents observations gathered from our early work with KSD and existing literature on data curation, agricultural information, and collaboration. We employ a participant observation technique to conduct research while working with KSD by engaging in the activities of the division while at the same time studying it (Becker, 1958; Moeran, 1997; Tedlock, 1991). Our larger project aims to develop a digital preservation strategy and infrastructure for the library. This allowed us to establish regular and sustained presence in the repository. Embedding in the library's activities expanded our understanding of the various types of digital materials currently managed by KSD projects. Our regular interaction with members of NAL staff and partners included attendance at meetings and brainstorming sessions. We were given access to key project documentations and reports. We also participated in "brown bags" and symposia hosted by the library. Beyond these organized meetings, our interactions also consisted of informal conversations and discussions with library staff. Over the course of these interactions, beginning in July 2015, we identified different paradigms that characterize collaborative activities running through each project. We considered the role of NAL as a mediator of collaboration in its projects focusing on data curation, management, and access.

## Overview of Collaborative Projects

We focused on four KSD initiatives in order to understand the range of active projects across the division. Through our work we have identified key themes of collaboration that unites these KSD projects. Furthermore, we view collaboration as an extension of the institution housing this work, a federal library. When USDA was established in 1862, it was tasked "to acquire and preserve in his department all information concerning agriculture which he can obtain by means of books and correspondence and by practical and scientific experiments" (United States Congress, 1862). This responsibility expanded as the collection of information grew into NAL's holdings, among the largest agricultural collections in the world today. This long history of work involving agricultural information affords NAL staff members credibility, allowing the KSD team to seek out and establish collaborative relationships to complete projects and contribute to the library's broad mission of collecting, preserving, and providing access to agricultural information.

Currently, KSD maintains four main projects in support of its efforts

to increase the data curation and management capacity at NAL. In these diverse initiatives the division engages different points along the research data lifecycle and develops expertise in dealing with a variety of data types. Table 1 introduces and compares these four projects, identifying the fields and types of data involved, as well as highlighting the collaborative aspects of the projects. The remainder of this section will introduce each project, focusing on the ways in which KSD staff collaborate and foster further interactions among their users.

*i5k Workspace@NAL*

The i5k Workspace@NAL is an insect genomics project that accepts data from arthropod genome projects with no other means of hosting their data. It grew out of the i5k Initiative, a long-term project with the goal of sequencing 5,000 arthropod genomes (i5k, 2017). While some work in this area takes place in large research labs, other projects come out of smaller sites with limited ability to publish and collaborate around data. The i5k Workspace utilizes open source tools already in use in the genomics community to facilitate international collaboration. This is a collaborative project; the design of its data access system explicitly facilitates data sharing and collaboration of raw and analyzed data that would otherwise be largely inaccessible to users.

Table 1. Summary of NAL Collaborative Projects.

| | *i5k* | *LTAR* | *LCA* | *ADC* |
|---|---|---|---|---|
| Scientific field | Insect genomics | Agro-ecology | Life-cycle assessment | All agricultural science |
| Types of data | Gene sequences | Meterological instruments, hydrology, ecological data | Unit processes for use in model construction | Inclusive of research data from agricultural sciences |
| Collaborative nature of project | Facilitates collaboration for users through sharing of gene annotations. Project grew out of relationship between USDA researchers and international community of insect genomics researchers | NAL serves as data curation organization for network of eighteen research sites maintained by USDA and land-grant universities | Following the success of the initial project, NAL partnering with two other federal agencies (DoE and EPA) to develop federal LCA Commons, uniting access to public LCA data | Infrastructure-building project that promises to lead to additional collaboration between NAL and agricultural researchers, both inside and outside of government |

Due to its position within a library that counts agricultural data as one of its focus areas, the i5k Workspace team has built a system that facilitates the interaction and development of new knowledge by researchers who would not otherwise be able to share their data. One of the system's key features is the ability for registered users to contribute annotations to published datasets from the site (i5k Workspace, 2017). Using WebApollo software, users can visualize and edit sequences, as well as see how other researchers have annotated the data (Genome Architect, 2017).

In addition to the project's motivations and system design, the team itself has collaborated broadly during the course of the project. Through a cooperative agreement with National Taiwan University, bioinformatics students have come to work on the i5k Workspace project at NAL as part of an exchange program. These students have made meaningful contributions to the system, including the development of an Official Gene Set (OGS) pipeline for combining computational and user-generated genome annotations and pushing them to the repository of the National Center for Biotechnology Information (NCBI). National Taiwan University provides shorter-term student labor for the project, while NAL remains committed to supporting the project for the longer term. In this collaborative relationship among organizations, each participant contributes what it can in service of a more successful final product.

*Long-Term Agroecosystem Research (LTAR)*

In the Long-Term Agroecosystem Research (LTAR) project, KSD functions as a centralized hub for managing research data. The project team provides data curation and management services for a network of eighteen interdisciplinary research sites across the United States. The LTAR network consists of eighteen different sites around the country that have been working for decades "to ensure sustained crop and livestock production and ecosystem services from agroecosystems, and to forecast and verify the effects of environmental trends, public policies, and emerging technologies" (USDA, Agricultural Research Service n.d., n.p.). Since each of these sites has its own goals and research beyond the LTAR collaboration, the leadership team enlisted NAL to provide data management and curation services for LTAR data.

Since its start in fall 2014, the LTAR team at KSD has worked to build data access systems for the LTAR network. The current data portal uses a map view of the United States to highlight the locations of each research site, as well as various layers composed of data from the Agricultural Research Service (ARS) and other sources, including the U.S. Geological Survey (USGS). In addition, the portal provides access to real-time meteorological data from LTAR sites, with the ability to view visualizations of variables, such as temperature, air pressure, and wind speed, over time via a web application. Users may also download data for offline analysis.

Future plans for the data portal include the addition of new data layers, such as soil erosion and ammonia levels, as well as work to increase metadata compliance and participation among LTAR sites. To achieve these goals NAL must continue its work developing collaborations among research sites while not fundamentally changing the research taking place at LTAR locations. Data curation and management activities need to exist in harmony with the scientific imperatives of the ongoing data collection, analysis, and publication activities, adding value to this work rather than additional overhead. This places the project team in a difficult position, albeit one with significant potential to build meaning and value through connecting longitudinal data about agriculture from across the country.

Functioning as the data management hub for the highly dispersed LTAR project has affected the work at KSD in various ways. The organizational location of NAL within ARS was essential to forging the collaboration and establishing deeper relationships with LTAR sites, many of which are operated by ARS. During the planning of the research network, the library was understood to be a neutral location, which the various sites could work with on issues of data curation and preservation. Each site is unique; in some cases ARS has been conducting research at these locations for a century or more. Their commitment to and level of maturity around data curation and management varied widely before joining the LTAR project. Two related challenges facing LTAR are finding ways to work with sites that use and manage data in different ways, and to get each to accept the NAL requirements with regards to data format and standards. These issues need to be regularly revisited to ensure that, as new data types are added to discovery tools, they fit within existing data access systems.

*Life Cycle Assessment (LCA) Commons*

The Life Cycle Assessment (LCA) Commons project at NAL has been working since 2010 to publish agricultural data for use in modeling analyses. Initially, the project's goal was to create a product using USDA research data that could be inserted into LCA models used by researchers around the world. With only two full-time staff members assigned to the project, the LCA Commons team has engaged with collaborators in the international LCA community, U.S. universities, and other federal government agencies.

One of the primary challenges faced by the LCA Commons project is the lack of standards regarding data management and sharing in LCA research. One of the steps researchers perform while employing LCA analysis techniques is to harmonize disparate datasets into a model, but, during the initial stages of the project, the team realized that its challenges concerning data management were greater than anticipated during the planning phase. When attempting to normalize data from a university researcher, the team realized that its assumptions about the ease with which

a future user could take data from different models and recombine them in a new LCA analysis were misguided.

Throughout the course of the project, LCA Commons has forged a series of partnerships with researchers and other organizations, within and outside of the federal government. This project has introduced team members to other LCA research taking place across the federal government. In particular, the Department of Energy (DoE), the Environmental Protection Agency (EPA), and the Department of Defense have LCA groups that have shared resources with NAL about data management. The LCA team at NAL has also partnered with engineers and agricultural researchers at universities across the country to curate and publish LCA models and unit processes for use by other researchers.

Project team members understand that being a part of NAL allows LCA Commons to focus on data management and information services rather than building models. The opportunity to interact with different research groups using LCA in a range of contexts has given the team new insights about interoperability and data issues that isolated researchers would not necessarily recognize. For example, the data harmonization and cleaning steps undertaken to build an LCA model are not as standardized as initially believed. LCA techniques can be applied to a range of domains, and each researcher has individual methods for building models, the details of which are not often shared in published papers. Only through the LCA Commons team's shift in perspective from that of the researcher to the information professional did it come to understand the nature of the challenge inherent in increasing access to LCA data.

Currently, the project team is working to develop a closer collaboration with DoE and EPA through a new initiative called Federal LCA Commons. In KSD quarterly update meetings, as well as through the project website (http://drupal.lcacommons.gov/catalog), the team has been reporting on its progress in drafting a memorandum of understanding to formalize this relationship and lay the groundwork for future development of a centralized data portal. Signing the memo, a major goal for the project, represents significant effort coordinating across agencies, but it would unify access to LCA data from the federal government. Project leadership sees this as vital to the future of the project and the growth of LCA research within the federal government. Due to the complex nature of LCA analysis, few federal researchers employ it. Formalizing the relationship among agencies that conduct LCA research would thus increase coordination and access to data suitable for use in these analyses. This type of collaborative relationship growing out of a pilot database project at NAL demonstrates the value of libraries in collaborating around scientific data by facilitating and maintaining a platform that allows for data access and sharing.

*Ag Data Commons*

Ag Data Commons is KSD's general-purpose data repository and catalog. A relatively new project, it highlights the unique ability of a federal library to build a platform for data management, preservation, and access in response to a policy directive. This project positions the library to play a larger role on issues around research data in the future, both in USDA and across the government, by building an infrastructure for agriculture data curation and long-term preservation that also integrates with the government's expectations for public access to data. As with the other KSD projects, Ag Data Commons' success is a result of being located in a library where infrastructure projects are supported. This type of repository and catalog could not easily arise out of a single research lab or department in which the goals for generating new knowledge are laid out in short and medium timescales. The location of the Ag Data Commons project within NAL and the larger ARS allows the repository to incorporate disciplinary knowledge from the agricultural sciences with information expertise and recognition of the policy and legal obligations of federal government agencies with regard to providing access to scientific data.

In 2013 the Office of Science and Technology Policy (OSTP) released a memo to all federal science agencies on "Increasing Access to the Results of Federally Funded Scientific Research" (Holdren, 2013). The goals of this Obama administration initiative are to push federal agencies with more than $100 million in annual research expenditures to develop and implement a plan for providing public access to scientific publications and data developed with federal support. Part of USDA's plan includes the creation of a repository for agricultural research data generated with federal funds (USDA, 2014). This project, called Ag Data Commons, is developed and maintained within KSD.

Ag Data Commons is currently live and operating as a beta release (http://data.nal.usda.gov). The repository is built on the DKAN platform (http://www.nucivic.com/dkan), a software package combining Drupal with at data catalog and repository inspired by the Comprehensive Knowledge Archive Network (CKAN). DKAN is used by multiple open-government data repositories and facilitates pushing data and metadata to the federal government's open data portal (https://www.data.gov). It is built to comply with the Project Open Data standard, further enhancing USDA's compliance with the Obama administration's OSTP memo.

Creating Ag Data Commons was not the direct result of a collaborative effort involving KSD and another organization, but this project stands to impact the ability for NAL as a whole to engage with other units of USDA, as well as other federal government agencies, around research data. Building a repository to serve as a home for ARS research data positions the Ag Data Commons team to play an important role in the evolving conversa-

tion regarding data curation at USDA. Project team members have participated in meetings about big data within USDA and have briefed NAL's director on how Ag Data Commons fulfills OSTP's mandate, while at the same time enabling more data curation initiatives within the agency. For example, the National Institute of Food and Agriculture (NIFA), USDA's funding agency, began a pilot program in the 2015 funding cycle requiring grant applicants to include data management plans along with their larger proposals. With Ag Data Commons operational, KSD has the ability to provide leadership around metadata standards, data dictionaries, file types, and preservation expectations for USDA research. Ag Data Commons is infrastructure to support the next generation of information services from ARS and USDA.

Each of these four projects demonstrates the value that libraries can bring to data curation projects through their ability to effectively collaborate with government agencies, the private sector, universities, and individual researchers. By embracing its role as facilitator and developing expertise in data curation skills, NAL has expanded its mission of providing access to agricultural information to include the provision of data curation services. While in the past, NAL's collections included research data, this often took the form of tables included in published research reports (see, e.g., USDA, Agricultural Research Administration, 1951, p. 43). The creation of KSD and the library's focus on agricultural data curation take a broader view of data access than the inclusion of data in reports. By engaging with researchers and users of agricultural research data, NAL is working to ensure that these data are increasingly available and useable in digital formats.

## Discussion

Research libraries can play an important role in data curation activities, but successful projects still require clear vision and articulation of how a given library's participation will positively impact a project. The four NAL-KSD projects described in this paper demonstrate the diversity of roles that libraries can play in data curation contexts. The common thread through these projects is a commitment to collaboration. The library has assumed different roles in partnerships with universities and other agencies within the federal government (and in some cases the private sector) and built systems that enable further collaboration among users of agricultural data. These efforts also serve as exemplars for similar collaborative efforts among data science researchers, curators, and managers. Our experiences at NAL have highlighted data curation collaborations facilitated or enabled by a federal research library. Out of these examples, we saw three broad areas in which information professionals play a role in library-mediated collaborations. These are as follows.

*Organizational Collaboration*

Libraries can engage in collaboration on an organizational level. This is where a project involves a library and at least one other organization. Every one of KSD's projects involves this type of collaboration in some way. The LCA Commons project hosts data created by researchers from academia as well as DoE, collaborating to make it understandable and ready for reuse by other LCA researchers. The i5k Workspace project grew within KSD thanks to preexisting collaborations among ARS researchers and the insect genomics community. The project has continued to produce new collaboration opportunities, including a partnership with National Taiwan University to bring students to NAL as graduate student interns. The geospatial LTAR team at NAL continually engages with the LTAR network research leadership, as well as representatives from each of its eighteen sites, to increase consistency among the converging data streams, which leads to data access systems that are more user-friendly. Consistent communication is necessary to ensure that the data being aggregated on NAL's LTAR data access system can be represented clearly.

*Collaboration in System Design*

Another paradigm for library collaboration in data curation occurs when the access system facilitates collaboration among data users through its design. This characteristic is primarily displayed through the i5k Workspace project. The open source software used in this project allows users of insect genomics data to run their own analyses and share and compare their work with others on the site. One of the most important tools in the i5k system is Web Apollo, a web-based genomic annotation package that allows geographically disbursed users to observe and describe genomic data over the web (Lee et al., 2013). While not designed specifically for insect research, the i5k team was able to adapt this open source software for use in its workspace, along with other tools like the Basic Local Alignment Search Tool (BLAST), an analysis tool developed by the National Library of Medicine. Through combining these different web-based genomics tools, the i5k team built a platform that incorporates the expectations and standards around data sharing and collaborative research, which are well-developed in genomics and reflected in the software in use in the community.

*Facilitating Collaboration*

Finally, libraries can function as a collaboration venue, bringing different groups together and mediating issues around data curation and access. Within KSD, the LTAR and i5k projects demonstrate the value that libraries can bring to a collaborative project by acting as a focal point around which a scientific community can share and provide access to data. Re-

search libraries like NAL have the institutional stature to lead conversations about information access and data management, given their history of working to increase access to scientific and technical literature. NAL's involvement in building i5k Workspace grew out of the larger i5k Initiative in which some USDA researchers were involved. When this group realized that the data from smaller labs were at risk of being lost or inaccessible, NAL recognized that it was in a position to support this project and secure the data within a more stable organization. Similarly, the role played by KSD in the LTAR project is to provide guidance to researchers, acquiring and facilitating access for disparate data through development of a centralized access system.

## Key Questions in Library-Mediated Collaborations

In this paper we have presented examples from NAL as exemplars of ways in which libraries can collaborate in and structure research data curation activities. While these examples demonstrate some of the themes raised in the literature on data curation, particularly in illustrating the role of libraries and information organizations in data curation practice, we will now explore four major questions that will develop issues that libraries might consider at the early stages of data curation collaborative projects. We offer these in service to further operationalizing this literature and refining its observations. This is not an exhaustive list, but rather is meant to prompt planning discussions to include these issues that manifest in KSD's work.

### *What Stages of the Data Life Cycle Is the Library Becoming Involved in for a Given Data Curation Project?*

Data life-cycle models (of which numerous examples exist) are a useful visual aid for understanding and conceptualizing the process of creating, managing, and disseminating research data. KSD's work involves engagement at different stages of the life cycle, from the coordination among research sites of the LTAR team to the repository services offered by Ag Data Commons. It is important for a library to consider the various points where it has the opportunity to engage. If a library is given the opportunity to build an innovative access system but has limited influence over the quality of data or the standards used to collect and describe them, expectations will not be the same if librarians are engaged further upstream in the process.

### *What Is the Position of the Library in Relation to Users of Research Data?*

Libraries wishing to engage in data curation need to define the scope of those activities. Considering who the users are of a given access system is one way to fulfill this need. As explained in the OAIS reference model, digital repositories must identify a "designated community" for whom their

data are tailored and demonstrate their knowledge of the requirements of this community (CCSDS, 2012, n.p.). The four KSD projects presented in this paper are oriented toward very different designated communities, but each project understands its relationships to users and the types of access and data reuse it can enable.

*What Is the Level of Resource Commitment for the Library in This Collaboration?*
While some financial costs, such as those of hardware and software, are easily tallied, others affect possible data curation collaborations. If the library will be contributing a specific set of skills (for example, metadata expertise, design, technical skills, and so on) to a project, it must know its current capabilities. KSD has variety of ways it staffs its projects, including using federal employees, contractors, interns, and graduate student fellows. While drawing from different pools of resources for different lengths of time has allowed the division to advance projects, a longer-term risk remains in not having sufficient permanent staff members to maintain and expand operations. Human resource issues continue to impact the division; potential projects must include an honest calculation of the work required from all organizations involved in a given project. Some potential collaborations have been scaled back or declined due to a lack of available staff in KSD. Over the long term, digital repositories can raise sustainability concerns for host institutions (Erway, 2012); in collaborative relationships these issues should be considered as early as possible so as not to arise as projects move from the building to maintenance phase.

*What Are the Outcomes and Benefits of Collaboration for the Library, Data Producers, and Users?*
Libraries should consider possible impacts and benefits for all stakeholders in a collaboration, including data producers, users, collaborators, and library staff. The opportunity to collaborate on a data curation project is also a time when library leadership can reflect on the institutional mission and scope and consider how the potential project contributes to the broader goals of the library. Within KSD these conversations take into account NAL's mission to provide increased access to agricultural information in all forms (USDA, 1990). The four KSD projects introduced in this paper fall within this broad mandate, but other considerations must be taken into account in order to assess outcomes. For example: Are users satisfied with the system and data? Has the library expanded its influence into another community that aligns with its mission?

## Conclusion

We set out to investigate ways in which information professionals facilitate, contribute to, and mediate collaboration in data curation projects. From our observation of the KSD projects, library professionals play a crucial

role in developing project goals and milestones. In some instances information professionals are the instigators of collaboration, because KSD has the resources and expertise to bring various stakeholders together. They do this by hosting research data for a multisite collaboration effort, specifically in the LTAR project; building access systems that involve various actors outside the library in all four projects; coordinating with various designated communities in LCA Commons and i5k Workspace; and leading the efforts to setting standards over metadata in LCA Commons, i5k Workspace, and Ag Data Commons.

The concept of library-mediated collaborations focuses attention on the roles that information professionals play in data curation projects. However, more work would increase our understanding of the place of libraries and information professionals in collaborative endeavors. While this study has demonstrated the different collaborative relationships in place at NAL and considered how these may serve as more generalizable insights for other libraries, the dynamics of library-mediated collaborations would benefit from further verification and elaboration beyond the four projects we studied. We hope that the questions explored in our study will inspire colleagues involved in data curation to assess their collaborative readiness, capabilities, and capacities.

## Acknowledgments

The authors wish to acknowledge the USDA-NAL for supporting this research through a cooperative agreement. Additionally, they would like to thank Cynthia Parr for her suggestions and comments on this manuscript.

## References

Becker, H. S. (1958). Problems of inference and proof in participant observation. *American Sociological Review, 23*(6), 652–660.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries, 16*(3–4), 207–227.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journalon Digital Libraries, 7*(1–2), 17–30.

Borgman, C. L., Wallis, J. C., & Mayernik, M. S. (2012). Who's got the data? Interdependencies in science and technology collaborations. *Computer Supported Cooperative Work, 21*(6), 485–523.

Consultative Committee for Space Data Systems (CCSDS). (2012, June). *Reference model for an open archival information system (OAIS).* Recommended practice, issue 2. CCSDS 650.0-M-2. Washington, DC: CCSDS Secretariat. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice.* London: Sage.

Edwards, P., Mayernik, M. S., Bacheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science, 41*(5), 667–690.

Erway, R. (2012, April). *Lasting impact: Sustainability of disciplinary repositories.* Dublin, OH: OCLC Research. Retrieved from http:// www.oclc.org/research/publications/library/2012/2012-03.pdf

Garritano, J. R., & Carlson, J. R. (2009, Spring). A subject librarian's guide to collaborating

on e-science projects. *Issues in Science and Technology Librarianship, 57*. Retrieved from http:// www.istl.org/09-spring/refereed2.html
Genome Architect. (2017). Apollo. Retrieved from http://genomearchitect.github.io/about
Global Open Data for Agriculture and Nutrition (GODAN). (2015, July 15). Major funding from US Government for open data to tackle hunger and global poverty. Retrieved from http:// www.godan.info/major-funding-from-us-government-for-open-data-to-tackle-hunger-and-global-poverty
Gold, A. (2007). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine, 13*(9–10). Retrieved from http:// www.dlib.org/dlib/september07/gold/09gold-pt2.html
Gold, A. (2010). Data curation and libraries: Short-term developments, long-term prospects. Retrieved from http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1027&context=lib_dean
Hardy C., Philips, N., & Lawrence, T. B. (2003). Resources, knowledge and influence: The organizational effects of interorganizational collaboration. *Journal of Management Studies, 40*(2), 321–347.
Hedstrom, M. (2012, December). *Digital data curation—examining needs for digital data curators*. Paper presented at the International Conference on Trusted Digital Repositories and Trusted Professionals, Florence, Italy. Retrieved from http://nbn.depositolegale.it/urn:nbn:it:frd-9271
Hedstrom, M., & King, J. L. (2007). Epistemic infrastructure in the rise of the knowledge economy. In B. Kahin & D. Foray (Eds.), *Advancing knowledge and the knowledge economy* (pp. 113–134). Cambridge, MA: MIT Press.
Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration, 51*(7–8), 662–672.
Holdren, J. P. (2013, February 22). Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research. Executive Office of the President: Office of Science and Technology Policy, Washington, DC. Retrieved from http:// www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
i5k. (2017). About the i5k Initiative. Retrieved from http://i5k.github.io/about
i5k Workspace @ NAL. (2017). Rules for Web Apollo annotation with the i5k pilot project. Retrieved from https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project
Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work, 15*(4), 321–358.
Latham, B., & Poe, J. W. (2012). The library as partner in university data curation: A case study in collaboration. *Journal of Web Librarianship, 6*(4), 288–304.
Lawrence, T. B., Hardy, C., & Philips, N. (2002). Institutional effects of interorganizational collaboration: The emergence of proto-institutions. *Academy of Management Journal, 45*(1), 281–290.
Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., Stein, L., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2013). Web Apollo: A web-based genomic annotation editing platform. *Genome Biology, 14*(8), R93.
Lohrey, M. (2014, October 6). *LCA metadata: Challenges and opportunities*. Retrieved from http://vtechworks.lib.vt.edu//handle/10919/51268
Macdonald, S., & Martinez-Uribe, L. (2010). Collaboration to data curation: Harnessing institutional expertise. *New Review of Academic Librarianship, 16*(1), 4–16.
Mayernik, M. S., Choudhury, G. S., DiLauro, T., Metsger, E., Pralle, B., Rippin, M., & Duerr, R. (2012, September/October). The data conservancy instance: Infrastructure and organizational services for research data curation. *D-Lib Magazine, 18*(9–10). Retrieved from http:// www.dlib.org/dlib/september12/mayernik/09mayernik.html
Moeran, B. (1997). From participant observation to observant participation. In S. Ybema, D. Yanow, H. Wels, & F. Kamsteeg (Eds.), *Organizational ethnography: Studying the complexity of everyday life* (pp. 139–155). London: Sage.
National Research Council. (2015). *Preparing the Workforce for Digital Curation*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/18590
Olson, J. S., Hofer, E. C., Bos, N., Zimmerman, A., Olson, G. M., Cooney, D., & Faniel, I.

(2008). A theory of remote scientific collaboration. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.), *Scientific collaboration on the internet* (pp. 73–97). Cambridge, MA: MIT Press.
Palmer, C. L. (1996). Information work at the boundaries of science: Linking library services to research practices. *Library Trends, 45*(2), 165–191.
Parr, C. S. (2016, April). *An overview of the Ag Data Commons: A new USDA catalog and repository for agricultural research data.* Paper presented at the US Agricultural Information Network (USAIN) 2016 Biennial Conference, Gainesville, Florida.
Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.-Y., et al. (2015). The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Research, 43*(D1), D714–D719. Retrieved from https://doi.org/10.1093/nar/gku983
Ray, J. M. (Ed.). (2014). *Research data management: Practical strategies for information professionals.* West Lafayette, IN: Purdue University Press.
Rotman, D., Procita, K., Hansen, D., Parr, C. S., & Preece, J. (2012). Supporting content curation communities: The case of the Encyclopedia of Life. *Journal of the Association for Information Science and Technology, 63*(6), 1092–1107.
Schrage, M. (1990). *Shared minds: The new technologies of collaboration.* New York: Random House.
Schrum, W., Genuth, J., & Chompalov, I. (2015). *Structures of scientific collaboration.* Cambridge, MA: MIT Press.
Tedlock, B. (1991). From participant observation to the observation of participation: The emergence of narrative ethnography. *Journal of Anthropological Research, 47*(1), 69–94.
Toups, M., & Hughes, M. (2013). When data curation isn't: A redefinition for liberal arts universities. *Journal of Library Administration, 53*(4), 223–233.
Trant, T. (2009). Emerging convergence? Thoughts on museums, archives, libraries, and professional training. *Museum Management and Curatorship, 24*(4), 369–387.
United States Congress. (1862, May 15). An act to establish a department of agriculture. 37th Cong., 2nd sess., chap. 72. Retrieved from https:// www.nal.usda.gov/act-establish-department-agriculture
United States Department of Agriculture (USDA). (1990, March 23). National Agricultural Library (Departmental regulation 1020-001). Retrieved from http:// www.ocio.usda.gov/sites/default/files/docs/2012/DR1020-001_0.html
United States Department of Agriculture (USDA). (2014, November 7). *Implementation plan to increase public access to results of USDA-funded scientific research.* Retrieved from http:// www.usda.gov/documents/USDA-Public-Access-Implementation-Plan.pdf
United States Department of Agriculture (USDA). (n.d.). Federal LCA Commons, Life Cycle Assessment Commons. Retrieved from https:// www.lcacommons.gov/catalog
United States Department of Agriculture (USDA) Agricultural Research Administration. (1951). Gypsy and brown-tail moths control. Retrieved from http://naldc.nal.usda.gov/naldc/catalog.xhtml?id=CAT40003215
United States Department of Agriculture (USDA), Agricultural Research Service (ARS). (n.d.). Long-term agroecosystem research data overview. Retrieved from https://ltar.nal.usda.gov
Weber, N. M., Palmer, C. L., & Chao, T. C. (2012). Current trends and future directions in data curation research and education. *Journal of Web Librarianship, 6*(4), 305–320.
Wood, D. J., & Gray, B. (1991). Toward a comprehensive theory of collaboration. *Journal of Applied Behavioral Science, 27*(2), 139–162.
Wright, S., Whitmire, A., Zilinski, L., & Minor, D. (2014). Collaboration and tension between institutions and units providing data management support. *Bulletin of the Association for Information Science and Technology, 40*(6), 18–21.
Zorich, D. M., Waibel, G., & Erway, R. (2008, September). *Beyond the silos of the LAMs: Collaboration among libraries, archives and museums.* Dublin, OH: OCLC Research. Retrieved from http:// www.oclc.org/content/dam/research/publications/library/2008/2008-05.pdf

---

Ricardo L. Punzalan is an assistant professor in the College of Information Studies at the University of Maryland, where he teaches courses on archives and digital curation. He holds a doctorate in information from the University of Michigan (UM) School of Information. In addition to an MLIS from the University of the Philippines, he

completed two certificates of graduate studies at UM, one in science, technology, and society (STS) and the other in museum studies. His areas of research include understanding the relationship of archives and collective memory, the politics and dynamics of digitization decision-making in collaborative and interinstitutional settings, the uses and users of digitized archival images, and issues related to agricultural data curation.

Adam Kriesberg is a postdoctoral scholar in the College of Information Studies at the University of Maryland. His research focuses on access to public-sector information, digital preservation, and data curation. He works with the National Agricultural Library on research related to agricultural research data curation and management. He completed his doctorate in 2015 at the University of Michigan School of Information.