Research Data Curation

A Framework for an Institution-Wide Services Approach

EDUCAUSE WORKING GROUP PAPER

MAY 2018

EDUCAUSE

Table of Contents

Universities and colleges should consider an institution-wide approach to developing services for managing and curating research data. This paper identifies service areas and includes a framework for institutions to document current research data curation services and responsibilities.

Introduction

Data curation is the process by which data are put into a state to be managed so that they can be understood and used by parties across disciplines and organizations. The passage of time should not prohibit the use of the data. This requires that appropriate measures be undertaken to ensure data infrastructure, searchability, availability, and preservation. While data curation might encompass research data and institutional data, this paper discusses only the curation of research data.

The goal of research data curation is to allow discovery and use of the data. Since new uses for the data could evolve over time, the data must remain in a state that allows them to be usable in current and future technology environments.

One of the challenges of developing and maintaining a curation process is in making choices about where to invest limited resources and predicting likely needs for the data. This will then inform the continuum of complexity around the data, imposing factors on the data curation process based on the available resources and anticipated needs.

The stakeholders in this collaborative process include researchers, scholars, librarians, IT staff, institutional administration, funders, and policymakers. As such, the authors of this paper recommend taking an institution-wide approach to developing services for managing and curating research data.

The planning matrix that accompanies this paper (see Appendix A) is designed to help institutions document current services and responsibilities undertaken in research data curation, highlight existing gaps in services, and identify steps and resources needed to move to the next level of maturity as appropriate. The framework should allow institutions of varying sizes, characteristics, and capacities to assess research data curation services in terms of their own needs and missions. Not every institution will aspire to the same level of service or maturity. The framework itself lists the various tasks involved in the research data curation process and prompts users to identify who in their institution fills the roles needed to complete each task. For instance, a user might go to the section on

Data Discovery and focus on the task of data location. At this point, the user will need to identify who actually does this task, as well as who consults with and trains researchers, who has the knowledge to make referrals for help with this task, and who provides the tools needed to accomplish the task. The process of identifying the individuals responsible for each task may indicate that a well-functioning approach to research data curation is already in place or instead may highlight the need for a better, coordinated team approach across the institution. Parties responsible for each task include third-party services, archives, college/school, dedicated management unit, department, central IT, departmental IT, library,[1] office of research, researcher, and research team. The framework should allow libraries, IT, and archives to make the case for resources to implement and support research data curation.

Data Curation Service Areas

Institutions may offer services in various permutations, in not only the types of services but also the depths to which they can go and the people to whom the services are offered. Faculty and graduate students are the most likely users and targets of data curation services, while some institutions are also interested in assisting undergraduate students and other researchers. This difference might be a capacity issue, in which more staff are needed to support increased requests, or it could be a policy decision to focus on faculty and researchers, who will remain at the institution long term or might need to comply with funding-agency policies in addition to institutional policies.

Services include the provision of the infrastructure necessary to perform data curation through licenses for preservation, analysis, and access tools. Providing centrally funded storage space for the curated data is another consideration. Training and/or consulting researchers in the stages of the process is necessary to allow researchers to take advantage of the services that are offered by various units across the institution or that are the purview of one unit, such as the library.

Data Management Planning

Research projects that conclude with reusable data usually begin with strong data planning, generally in the form of a data management plan. Data management planning is the process of creating a formal plan for managing the data before the project begins. It may include plans for data collection, metadata creation, data

sharing, and preservation. Data management and planning is one of nine core areas supported by the US Department of Health and Human Services' Office of Research Integrity (ORI) in its responsible conduct of research initiative.[2]

Major funders, such as the National Institutes of Health and the NSF, require grant submissions to include a data management plan.[3] Many institutions offer support in creating data management plans (DMPs),[4] and templates such as the DMPTool are available to aid in plan creation. While the data planning and management process is ultimately the responsibility of the research team, university research officers, librarians, and IT professionals may play a role in tracking funder requirements, consulting on data management plans, and training and advising research teams on the collection of the data and metadata and on sharing and preservation.

Roles and Skills

- **Understanding funder requirements:** At the onset of a research project, data management requires a clear understanding of funder requirements and how they fit into the research life cycle.[5] Many institutions and libraries have created guides to help researchers navigate the maze of policies published by funders.
- **Understanding data management at a high level:** Service providers, such as librarians and IT professionals, need to understand what data will be generated by a project, how the data will be collected and stored, and plans for reporting and sharing data to effectively consult with a research team in creating a viable data management plan. Many institutions and libraries have created guides and templates to create good data management plans.
- **Ability to consult on data management plans:** Once librarians and IT professionals have gained an overall understanding of data management, they need to put their consultative skills into practice working with individual researchers or research teams to complete the data management plan. The process is often iterative and requires strong communication skills (Appendix B identifies many institutions offering services in this space). It may also require close engagement with the institutional review board (IRB).

Data Discovery and Usability

Data discovery is the process of finding and accessing already created data sets for use in research. In many cases, researchers use not only the data sets they create but also those created and curated by others. Finding data sets in external repositories can be time consuming, and, once the data sets are found, researchers

can struggle with actual access to them (which may have costs associated), understanding and applying the terms of use (which may be restrictive), and interpreting the metadata describing the data set so it can be incorporated into their research appropriately. Use of data sets from external services can complicate sharing or publishing data further downstream if there are license terms that prohibit such activities.

Many researchers are familiar with the strategies needed for their own data discovery. The social sciences have historically been active in this area, and other disciplines may have specific expertise. Support for data discovery is often provided by libraries, research institutes, and research offices. Responsibilities to support researchers include finding, accessing, manipulating, and analyzing qualitative and quantitative data sets. In addition, some libraries or other units purchase memberships or licenses to services that allow researchers to access data sets. License terms may need to be negotiated for specific types of access or uses of data sets or for clarifications of what restrictions exist. Institutions may provide support through library acquisitions or risk-management offices because researchers are rarely allowed to accept license terms on their own behalf. In addition, librarians and others have a role in assisting the researcher to provide the appropriate citations and in determining what may be shared or published.

Roles and Skills

- **Knowledge of external services for data sets:** There are many potential sources for data sets across all disciplines. Re3data.org, a registry of data repositories, currently lists more than 1,500. Librarians—particularly subject specialists and data specialists—should be able to support researchers in locating data sets that meet their needs by identifying appropriate services to search.
- **Understanding access and use terms:** While many external services provide data sets without restrictions on access or use, many require terms, data use agreements, or membership. Librarians and IT professionals should be familiar with the memberships an institution holds (in ICPSR, for example) and be able to read and interpret the terms placed on access and use. Terms might prohibit publishing a data set even if merged with another and, thus, limit the type of sharing or publication that could happen at the end of a research project. In some cases, there may be very specific terms of access for privacy or other reasons that require additional scrutiny of the security of a workstation; this would require additional support of an IT professional and possibly a research data enclave (e.g., NORC).

- **Data acquisition:** Beyond memberships in external services such as ICPSR, some libraries have begun to acquire individual data sets on request by researchers. This often requires extended negotiation with the owner of the data set to allow for some type of campus access. Librarians—particularly those who manage the acquisitions of electronic resources—are well positioned to provide assistance in this area and to negotiate favorable terms for use of the data sets. See Data Purchase Program from the University of Illinois at Urbana-Champaign for an example of such a program.
- **Understanding data set description for use:** To determine whether a data set is useful for a research project, the metadata describing it must be extensive enough to allow a researcher to understand how it was created, cleaned, and collated. Provision of data dictionaries, codebooks, and protocols used among other elements can help a researcher understand whether the data set is appropriate for use. Librarians and IT professionals can assist in this effort by ensuring the metadata are available and usable for the researcher. This process can also be used to help researchers understand the importance of good documentation when they are share or publish their own data sets!

Data Description

Providing descriptive information is essential to ensure that data are findable in the future, as well as available to and reusable by other interested researchers. Several layers of description can be applied to data sets, from granular, discipline-specific metadata to metadata for discovery systems and aggregations to descriptive documents accompanying the data set. Applying metadata provides information and context about the data that are not always apparent from the data alone. Metadata standards may be available to guide the information needed; discipline-specific standards exist in some areas, whereas in other areas schemas may need to be developed or more generic descriptive information provided. General metadata standards, such as Dublin Core, provide guidelines for basic citation elements. Important data-specific elements include data collection methods, processing and analysis methodologies, research questions, and context for the content. In addition to describing the data, it is crucial to describe the terms of use, including who can use the data, how the data could be used, and privacy and intellectual property issues.

Many libraries provide guidance on descriptive information that should accompany data sets. Check out your local options. If data will be deposited in a specific repository, that system may also have requirements for metadata and other descriptive information.

Roles and Skills

- **Familiarity with existing metadata standards:** A standard metadata schema and elements should be used whenever possible and vary by discipline and repository requirements. Additional elements may also be used to adequately describe the data. Many disciplines include codebooks as typical parts of their data collection process. Codebooks should be interpretable by someone outside the research team or have a key and be included with the data set. When assigning metadata, use existing standards and/or patterns for assigning new elements according to the various disciplines.
- **Familiarity with data model standards and impact for interoperability:** In order to organize the data to increase interoperability, data model standards[6] should be followed, which are often set either by a discipline or a repository. When adding new descriptive elements to understand the data, follow a data model to aid interoperability. Data models determine the structure of various components of a data set and impact the user interface and interoperability of the data sets. For example, determining whether a codebook is part of the data set or is a work on its own affects how the data set is consumed or how the data set relates to a project as a whole. Is the research project described as a primary object with the data set and publications as components?
- **User domain expertise:** Assigning metadata for data sets often requires domain expertise and participation of the researcher and/or research team. This expertise is often found in research institutes, in discipline-specific repositories, and with subject experts in libraries. Discipline-specific metadata often must be described by an expert or may be assigned by a non-subject-specific expert if derived from the README or other contextual information provided. To help with discovery, generic metadata may be added to expose data to large discovery systems, aggregations, and search engines. Information for the data set creators that ties to a system such as ORCID for disambiguation among creators is also encouraged.
- **Defining terms of use:** Knowing how data can be used for what and by whom is essential to the reuse of the data. Copyright of research data is a gray area, and you may want to seek advice from counsel. Standards for describing rights information such as using Creative Commons licenses[7] and RightsStatements.org should be used whenever possible. If there are any restrictions due to personally identifiable information (PII) or other privacy issues, they should be explicitly addressed in the description information accompanying the data set. Some data sets, particularly those derived from another source or analyzed with another tool, may also have specific citation requirements.

- **Understanding aggregate search:** Library staff are often trained to provide metadata that will help with finding data in large aggregations of disparate types of information, including general repositories, citation services, and search engines.

Data Analysis

When data have been transformed and described, qualitative, quantitative, visualization, and any number of other types of analysis can be performed on the data. Depending on the type of analysis, specific skills will be required. First, knowledge in the use of the selected software product to be used—such as SPSS, SAS, ArcGIS, BioConductor, FlowCytometryTools, and others—is essential. Each tool has a learning curve to be proficient at a basic level, and advanced techniques require even deeper skill sets. Emerging methods, such as machine learning, allow for analysis of unstructured data.

The data may also have context or complexities that require discipline-specific knowledge. Whether it be bioinformatics, GIS, or astrophysical files, each discipline's data have nuances that may require assistance from a data-services librarian, an expert from a department, or other specialist area. Involving those experts in the curation process will lead to data that are more completely defined and accessible to others.

Roles and Skills

- **Preparing data for analysis:** In the data analysis phase, data may need to be transformed, normalized, or extracted before proceeding. Skills in this area may require programming, database, and/or file manipulation expertise, which allow for data editing, transformation, and output in various formats. The operations performed for preparation may also be stored and/or described with the data set for reproducibility and context.
- **Pairing data analysis needs and tools:** Many different analysis tools and techniques may be available to analyze and process the data depending on the discipline and researcher needs. Data curation services staff may support researchers in identifying tools to help in analyses. Some services may also provide instruction and workshops to use specific tools.
- **Selecting data to output:** After actions have been performed on the data, data are prepared for output or extraction. Many different analysis tools provide multiple file formats and accompanying settings for output. For some data curation projects, the analysis settings become the data set rather than a specific data output (e.g., flow cytometry). Other data analyses are performed

in place, often due to data set size, and descriptions of the actions are outputted rather than a copy or subset of the original data. Data curation services can help identify file formats—for long-term access and for ease of reuse in current systems—and accompanying information that may need to be extracted along with the data to help ensure reusability.

Data Storage

The core purpose of data storage is to ensure its security and integrity. Many federal granting agencies require researchers to submit plans for preservation and reuse, which has raised the level of awareness among both researchers and administrators for the need to store data securely to guard against loss. Even with this raised awareness, however, researchers often store their own data and may create their own informal storage systems in the initial stages of a research project, such as during data collection. Yet even at this early stage, institutions ideally would already be involved with projects to ensure proper security, preservation, and monitoring methods to protect the integrity of the data.

For large data sets that will be actively analyzed, file system performance (throughput and metadata performance) is an issue that needs to be addressed through different tiers of storage. These storage tiers should range from a high-performance parallel file system that can be accessed by advanced analytics environments to archival storage with a high latency for access for rarely used data sets.

The expertise of IT storage professionals may be required to assist with the creation and management of the storage environment being requested. The complexity of both on-premise and cloud storage environments can be daunting and may need the advanced knowledge and toolsets that a central IT storage management group can provide. Ongoing support of the environment, as well as the care and feeding of backups and archives, can be handled by researchers, but as complexity and volume increase, they may need additional support from storage professionals.

There are increasingly varied options for storage both online and offline, and with the increasing capacity of online storage, more data can be stored than ever before. Some data come with sensitivity concerns, whereas other data can be made accessible to the public. Data created at government agencies and public colleges and universities might be subject to records retention schedules.

Researchers may have funding gaps that will make it difficult to pay for the required storage for their data. When planning for data storage, it should be for sustainable data storage. Institutions should provide consulting and funding to help researchers bridge gaps. Negotiating prepaid storage with a cloud vendor and including that in the grant direct cost might be an option to mitigate the loss of access to cloud storage resources.

Roles and Skills

- **Understanding security practices:** This is the first step to ensure that data are securely stored.[8] If data are being maintained online, security features such as password access and encryption measures should be in place for data requiring limited access, sometimes referred to as a data enclave.
- **Digital archiving and preservation:** Adequate preservation ensures the authenticity and integrity of the data through a series of steps. After acquisition, data are appraised, arranged and described, and maintained according to archival protocols. During the maintenance stage, data should be checked, backed up, and refreshed to ensure their longevity. Storage for data sets should follow archival best practices. Data created by public institutions might be subject to retention schedules. Further, data sets created for particular agencies or funded by certain grants will have specific protocols attached to them. Those responsible for data storage must be aware of these different requirements and follow multiple layers of compliance guidelines. Digital preservation comes with unique challenges to its longevity, creating a need to check data periodically to ensure integrity. This is often done through fixity checks. Many institutions create two copies of data, a preservation copy and an access copy, ensuring that the preservation copy is stored and unused in case the access copy fails or is altered.

Data Access

One of the key goals of data management and curation relates to access and sharing for others to verify published results, replicate the original researchers' work, or make new discoveries or uses, perhaps even in unanticipated ways. Ingesting data into archives through well-defined and documented processes greatly bolsters the prospect for successful data access and sharing. This goal aligns well with the view that successful preservation results in being able to share data and understand the associated context without having to contact the original data producer. Increasingly, both government and private funders are emphasizing data sharing as part of their public access policies. It is worth noting that such policies do not require deposit into institutional data repositories.

Having said this, while many communities have disciplinary or domain data repositories, an institutional data repository does provide an option for communities or researchers without such options.

While some institutions have opted to leverage their document-based repositories, data sharing requirements are often more complex or demanding beyond assigning identifiers and metadata. Terms of use might prohibit publishing a data set even if merged with another, thus limiting the type of sharing or publication that could happen at the end of a research project. Data sharing might also require an understanding of the underlying data models or the expression of those data models in the form of XML schemas. Additionally, in many cases, data alone are not accessible without the accompanying algorithms or software that were used to create the data. It is worth mentioning the importance of workflows since data are often transformed through different stages of calibration and processing.

Roles and Skills

- **Knowledge of available data repositories and associated requirements for deposit, including publication:** While an increasing number of institutions offer local platforms for data, it is important to note that many communities and disciplines already have data archives in place. Examples include ICPSR and the data centers funded by NASA, NOAA, and others (such as the National Snow and Ice Data Center, NSIDC). Data management consultants at colleges and universities should be able to refer researchers to relevant data archives for both deposit and access purposes. Requirements may include licensing and specific metadata.
- **Assignment and use of persistent identifiers:** While publishers are increasingly requiring the use of identifiers such as DOI, data management consultants can provide an additional point of reference regarding the value of and choices regarding persistent identifiers. Examples include alternatives to DOIs (e.g., ARKs) or identifiers for different purposes, such as for individuals (e.g., ORCID).

Points to Consider

An institution's service delivery model is based on a number of factors, including the type of services being offered and to whom, which unit or units are required for the provision of service, and the number of staff and other resources available. Service delivery models are typically of three types—ad hoc, formal, or

institution-wide programs—that in each instance may include a fee-for-service approach to offset increased costs for providing these new services.

With an ad hoc approach, members from one or several units provide assistance on specific data curation efforts but without any formal policies or specific coordination in place. For instance, IT may provide storage space for a specific research team, while libraries assist with data description for another.

In a formal service approach, units across the institution provide defined services, and though there may be broader coordination between the units, it may be limited in nature. These units have dedicated staff time for data curation services and formalized descriptions of what services they offer and to whom, and they might market their services openly. For example, libraries, IT, or research offices may offer training on specific tools chosen by the institution.

An institution-wide programmatic approach is one in which the institution has identified local data curation needs and created a data curation policy, as well as an administrative and technical infrastructure to support those needs.

Policies in Place to Govern Research Data Curation Efforts

While organizations such as the Digital Curation Centre (DCC) and ICPSR have developed community resources (with ICPSR also curating data) for many years, most colleges and universities in the United States are relatively new at providing data curation services. Additionally, there is no single US entity that can provide guidelines or even guidance for the diverse network of higher education institutions. Rather, many organizations have produced papers, studies, training videos, and kits to help guide US institutions.[9]

In this context, it is particularly important to consider existing institutional policies that relate to data. For example, institutions may have relevant policies for data retention, data protection or privacy, tech transfer, metadata, and others. Typically, the office of research administration is a good starting point for such existing policies. Given the specific nature of health sciences data, schools of medicine may be another useful resource to consider.

Individual institutions with developed data curation models can likewise offer insight to peers developing data curation services. Several individual institutions make their data curation policies publicly available, typically to inform researchers who are depositing data about how their data sets will be managed.[10] These resources can also help guide institutions looking to develop their own policies.

Funding

Funding models for data curation programs can be as varied as the programs themselves. There is no one-size-fits-all solution, and as such careful consideration and planning should be conducted to find the best fit for an institution.

Data curation funding could potentially be sourced centrally, from a single entity on campus. This may be the library, IT, or the provost's office. Regardless of the source, however, careful planning and consideration is required to ensure that adequate funding is provided for all necessary activities.

Alternatively, budget responsibilities may be derived from multiple entities, based on any number of allocation strategies. For instance, IT may be responsible for technical infrastructure, the library for accessing data and creating metadata, and the provost or individual research elements for providing software and/or tools for searching, viewing, and acquiring data from a curation archive.

Finally, an institution may want to fund data curation efforts via chargeback models (possibly including grant funds), whereby usage of the curation services is paid for by the service users. Rates for various types of activities would have to be agreed upon and published. A mechanism for measuring usage, calculating fees based on usage, and transferring funds would have to be put in place, managed, and audited. One consideration in this model is whether the rates charged are meant to be a sustaining fee, or whether the fees are expected to additionally cover growth of the infrastructure and service.

Various campus organizations will need to be involved, depending on which funding model is selected. These may include the budget office, accounting/finance, the provost, and sponsored research administration. Divisions and departments that are users of the service would be impacted, as well as potentially IT and the library.

Staffing

By its very nature, data curation is a collaborative activity. Partnerships generally include research offices, information/computing technology services, and libraries. For instance, there may be an institution-wide steering committee for research data with library membership, faculty partnerships to support research activities, or a research technology task force. A strong partnership model presents the opportunity to leverage existing skills and talents across the

institution to support data curation and allows each partner to focus on the services and activities that existing and new staff can best provide in support of data curation.

Staffing for data curation depends on the service delivery model and its maturity. Institutional service models generally fall into two categories: 1) institute, departmental, or library initiatives or 2) cross-campus partnerships, often including such entities as the university's research office, information/computing technology team, or research committee. Data curation initiatives that originate from institution-wide task forces and mandates often receive more support for staffing curation services, though sometimes an individual department will step up to the opportunity offered by these mandates with additional funding. As data curation services grow and stabilize, staffing and the need for skills and expertise generally grow too.

Data curation services can provide some insights into how existing positions can change focus and/or assume new responsibilities to meet new research needs. At the beginning, staff may take on additional duties or redefine positions to provide new services, learning new skills through training and applying old skills in new ways. At this stage, there is rarely a full-time staff member dedicated to data curation. Instead, there is typically a mix of partial FTEs, sometimes supplemented by volunteers, interns, and graduate assistants, who provide data curation services.

Some organizations are able to hire additional or dedicated staff, especially as data curation services expand. Position titles for new positions reflect a range of duties and research interests. Titles include director for curation services, data curator, data services coordinator, and research data management coordinator. Researchers, IT, and libraries must all work together, refocusing positions, retraining staff, and delineating responsibilities for data throughout the data life cycle. As a group, they will need to provide a coordinated set of consultative, training, and technical services that build upon and expand current skill sets.

Communication

Communication is essential to developing an institution-wide awareness of data curation activities. There are three primary audiences for communicating data curation activities and services:

- Institutional leadership (e.g., vice president for research, provost, director of technology transfer, director for sponsored projects, deans)

- Data curation services partners (e.g., libraries, information technology units, research office)
- Data curation customers (e.g., researchers, faculty, students)

A mixed strategy of marketing materials should be used to communicate with these audiences, including websites, feature news stories, email, fliers, internal discussion lists, posters, and social media.

Institutional Leadership

Having leadership support is essential to sustaining data curation services, particularly for any successful institution-wide data curation effort. Articulating the values and benefits of data curation services, such as raising the profile and reach of the university or college and sharing regular assessment and usage, should be part of the communication to the administrative offices.

Data Curation Service Partners

In order to develop institution-wide data curation services, each partner and its role should be identified, and communication channels between the units should be established. In addition, partners should work together to create a communication plan for how the groups will market their services, as well as how training and outreach activities should be incorporated. The service partner communication plan should also address how units coordinate on shared or related services, referrals across units, and responding to requests and issues (which may be managed via a service management system and/or well-documented processes).

Data Curation Customers

As a data curation services program is implemented, the communication processes will go through several phases. As service partners are brought together, getting buy-in from the proponents and potential customers at your institution will be critical to your success.

Communication strategies to effectively reach researchers include the following:

- Integrate data curation communications into existing channels, such as research support communication channels, in departmental communications and events, and via library liaison communications.
- Communications should emphasize the value and benefits of the services for researchers. These include but are not limited to:

- Increasing data reuse, with contextual support for understanding the data, often to unanticipated fields and uses
- Increasing findability of data
- Increasing the impact factor of the research
- Expanding the reach of the research beyond the academy and globally
- Providing means of data integrity and reproducibility of research
- Supporting compliance with funder requirements for data availability

Conclusion

Robust data curation services are nascent in many institutions, research communities, and organizations. At the same time, many funding agencies require researchers to make their data publicly available. The need for research data curation services is essential to the continued success of the higher education research community, and implementing research data curation institution-wide can help make sure that a broad array of curation services is available and scalable, while working with sometimes limited resources. However, an institution-wide approach hinges on collaborative partnerships and varied skills. This paper identifies key areas that most institutions will want to address, but there is no one-size-fits-all model. Each institution will need to develop an approach to data curation based on its own resources, ongoing research projects, and strategic vision.

A significant challenge for the development and maintenance of an institution-wide data curation service is in making choices about where to invest limited resources and understanding likely uses for the data. The strategic framework included in this paper can help institutions of varying sizes, characteristics, and capacities assess institutional research data curation at an enterprise level to help libraries, research offices, IT, and archives make the case for resources to implement and support institution-wide research data curation.

That said, additional areas will require future exploration. Systems are not static. New research will inevitably provide those individuals and organizations responsible for data curation with new tools and updated best practices as technology evolves. Further, researchers will continue to create new data on top of the data already deposited within curation environments, which will require additional resources to maintain. Those responsible will need to develop methods for predicting increases in submissions to their repositories. Finally, this paper

does not explore in depth the need for assessment, evaluation, and metrics regarding data curation services. Subsequent working groups may focus on these areas and others that will emerge with new exploration in the area of data curation.

Authors

Special thanks go to the following EDUCAUSE Research Data Curation Working Group authors of this report:

Sayeed Choudhury
Associate Dean for Research Data Management
The Johns Hopkins University

Esmé Cowles
Digital Infrastructure Developer
Princeton University

Nahali (Holly) R. Croft
Assistant Professor, Digital Archivist
Georgia College & State University

Karen Estlund
Associate Dean for Technology and Digital Strategies, Libraries
The Pennsylvania State University

Michael Fary, Co-chair
Senior Consultant for Strategy and Governance
University of Chicago

Grace Faustino
IT Project Manager III
University of New Mexico

Thomas Hauser
Director, Research Computing
University of Colorado Boulder

Anne Linton
Director, Himmelfarb Health Sciences Library
The George Washington University

Clifford Lynch
Executive Director
Coalition for Networked Information

Karen Menard
Assistant Vice President, Institutional Analysis and Research
University of Guelph

David Minor
Director, Research Data Curation Program, Library
University of California San Diego

Gregory E. Monaco
Research Associate Professor Emeritus
Kansas State University & Research Advisors Group

Daniel Noonan, Co-chair
Associate Professor, Digital Preservation Librarian
The Ohio State University

Sarah Shreeves
Vice Dean, University Libraries
The University of Arizona

David Ulate
Executive Director of Institutional Research and Planning
Foothill-DeAnza Community College District

Natalie Waters
Head Librarian, Schulich Library of Physical Sciences, Life Sciences, and Engineering
McGill University

Notes

1. A recent Association of Research Libraries SPEC Kit on *Data Curation* provides a good overview of how many research libraries are providing data curation services and of what kind.
2. To learn more, see the ORI *Guidelines for Responsible Data Management in Scientific Research*.
3. For more information, see "NIH Data Sharing Policy" and the NSF's "Dissemination and Sharing of Research Results." See also the 2013 OSTP memo "Expanding Public Access to the

Results of Federally Funded Research," which "directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally funded research freely available to the public—generally within one year of publication." Finally, the Association of Research Libraries maintains a list of current public access mandates by federal funders and provides some analysis of the various agency policies—see "Access to Federally Funded Research."

4. See the 2013 EDUCAUSE Working Group paper *Developing an Institutional Research Data Management Plan Service* for guidelines on creating a DMP service.
5. For examples of data curation life cycles, see the DCC Curation Lifecycle Model and the DataONE Data Life Cycle.
6. Examples of interoperable data models include PCDM and IIIF.
7. For instance, Creative Commons provided a CC0 license that is often associated with published data sets.
8. See the EDUCAUSE Information Security Guide: Best Practices and Solutions for Higher Education for some helpful information regarding data security. For instance, the guide includes a toolkit on Data Classification as well as Top Information Security Concerns for Campus Executives & Data Stewards.
9. These organizations include, but are not limited to, the Council on Library and Information Resources (CLIR), OCLC, the Digital Library Federation (DLF), the Association of Research Libraries (ARL), and the National Network for Libraries of Medicine (NNLM).
10. Research institutions such as Johns Hopkins University have established policies related to data management and sharing; the university also has a Digital Research & Curation Center. Further, the Digital Curation Centre (DCC) maintains a spreadsheet of the quality of each element of member institutions' data curation policies and provides links to each (see "Overview of UK Institution RDM Policies").

Appendix A: Data Curation Roles Planning Matrix

Institutions may offer data curation services in various permutations, in terms of not only the types of services but also the depths to which they can go and to whom the services are offered. While faculty and graduate students are the most likely users and targets of data curation services, some institutions are also interested in assisting undergraduate students and other researchers. This might be a capacity issue, where more staff are needed to support increased requests, or it might be a policy decision to focus on faculty and researchers, who will remain at the institution long term or might need to comply with funding-agency policies in addition to institutional policies. Depending on institution size and the maturity of data curation program(s), the services may be a formalized program, with the roles and responsibilities centralized into a handful of units; wholly distributed among all campus units; or a set of informal, ad hoc services.

This matrix is a tool for identifying who is responsible for the various data curation roles described in this report. It can be used as an initial assessment and gap-analysis tool, as well as documentation of services at one's institution. It is described below and can be accessed online as a spreadsheet or as an Excel file.

The matrix identifies each of the major set of roles and skills for the six service areas discussed in this paper (data management planning, data discovery and usability, data description, data analysis, data storage, and data access), and asks you to consider who at the institution is actually responsible for various aspects of those roles and skills:

- Who does the work?
- Who consults?
- Who trains?
- Who coordinates?
- Who refers?
- Who provides tools?

For each of those questions, you can choose a response from a pull-down menu that includes:

- Third Party
- Archives
- College/School

- Dedicated Data Management Unit
- Department
- IT—Central
- IT—Departmental
- Library
- Office of Research
- Researcher
- Team (if this option is chosen, there is another cell set aside to explain or describe the team; please choose "team" if two or more parties are responsible)
- None
- N/A

This list of selections is from a lookup table that can be modified as necessary by the institution. See table A1 for an example of how the matrix may be used, in this case, for data management planning.

Table A1. Matrix Example

Data Curation Tasks	Who does the work?	Who consults?	Who trains?	Who coordinates?	Who refers?	Who provides tools?	If "team," specify
Data Management Planning							
Funder requirements	Researcher	Office of Research	Office of Research	Office of Research	Library	IT—Central	N/A
Data management practices	Researcher	Team	Team	Office of Research	Team	IT—Central	Team: Library & Office of Research

Appendix B: Data Management Planning Services

Sample guides for data management planning services can be found here.

- Carnegie Mellon University Libraries, Research Data Management
- Cornell University, Research Data Management Service Group
- Johns Hopkins Libraries, Data Management Services
- MIT Libraries, Data Management
- NYU Health Sciences Library, Data Management
- Penn State University Libraries, Data Management Toolkit
- Purdue University Libraries, Research Data
- Rutgers University Libraries, RUcore: Rutgers University Community Repository.
- Stanford Libraries, Data Management Services
- University of Virginia Library, Research Data Services + Sciences
- University of Washington, University Libraries, Data Management

About EDUCAUSE

EDUCAUSE is a higher education technology association and the largest community of IT leaders and professionals committed to advancing higher education. Technology, IT roles and responsibilities, and higher education are dynamically changing. Formed in 1998, EDUCAUSE supports those who lead, manage, and use information technology to anticipate and adapt to these changes, advancing strategic IT decision making at every level within higher education. EDUCAUSE is a global nonprofit organization whose members include U.S. and international higher education institutions, corporations, not-for-profit organizations, and K–12 institutions. With a community of more than 99,000 individuals at member organizations located around the world, EDUCAUSE encourages diversity in perspective, opinion, and representation. For more information please visit educause.edu.

Citation for This Work

Choudhury, Sayeed, et al. *Research Data Curation: A Framework for an Institution-Wide Services Approach.* EDUCAUSE working group paper. Louisville, CO: ECAR, May 2018.