RAISING THE VISIBILITY OF PROTECTED DATA: A PILOT DATA CATALOG PROJECT

Erin D. Foster
Heather L. Coates

Abstract

Sharing research data that is protected for legal, regulatory, or contractual reasons can be challenging and current mechanisms for doing so may act as barriers to researchers and discourage data sharing. Additionally, the infrastructure commonly used for open data repositories does not easily support responsible sharing of protected data. This chapter presents a case study of an academic university library's work to configure the existing institutional data repository to function as a data catalog. By engaging in this project, university librarians strive to enhance visibility and access to protected datasets produced at the institution and cultivate a data sharing culture.

Introduction

As the landscape of data sharing evolves, infrastructure and practice are transitioning from individual transactions handled by data owners and re-users towards a more sophisticated and managed market in which there are trusted brokers and controls. With this transition, existing data discovery systems and models are challenged to support emerging mechanisms, expectations, and requirements for data dissemination and sharing. By leveraging expertise in the areas of metadata and data description as well as experience in administering existing models for data discovery (e.g., repositories, registries), libraries are well positioned to enhance the discoverability of research data and continue to support data sharing efforts moving forward.

As it relates to data description and discovery, there are opportunities for existing systems and models to tackle challenges in data sharing, particularly in regards to data deemed “protected”. For a variety of reasons, data collected for research purposes may be subject to legal, regulatory, and contractual protections that limit sharing. Consequently, the discoverability of “protected data” is constrained. Efforts must be made to ensure appropriate security and access protocols are in place for the management and sharing of these data. Current systems infrastructure and design vary in the ability to provide strong controls for these actions. As a result, work remains to be done in providing the same visibility to protected data as is afforded to data that can be shared in the open.

In this chapter, we will describe efforts at Indiana University Purdue University Indianapolis (IUPUI) to improve the discoverability of protected research data generated at the university. To achieve this, we are developing a data catalog, which will act as a search and discovery tool that describes datasets and connects potential users to data providers. This model of data discovery exists in various forms across disciplines and has the potential to enhance the visibility of protected data and facilitate its sharing. By leveraging this model, already in use by many, we strive to enable the responsible sharing of data at the university and, as such, contribute to the development of infrastructure and policies that advance data sharing more broadly.

Data sharing landscape

Data sharing, as a practice and concept, has been a discussion - or a part of practice - across research disciplines for quite some time.[1,2] The current focus on data sharing in research can, in part, be attributed to the increase in requirements from federal funding agencies, non-profit organizations, and professional societies that formalize (and follow through) on the sharing of research data generated, gathered, or created as part of a research project.[3,4] These requirements are increasingly included by journal publishers as well.[5] The guidelines for data sharing articulated in these requirements vary in the ways in which they specify sharing in terms of *what* data should be shared, *when* it should be shared, and *how* it should be shared.

Data sharing is part of the growing conversation around "open data" and the role that broader availability of data can play in ensuring high quality, reproducible research. Open data refers to data that is "available to anyone to access, use, or share" and free from restrictions from legal, financial, or technical standpoints.[6,7] A key contingent in the open data movement are government agencies such as the National Institutes of Health, the National Science Foundation, and the National Oceanographic and Atmospheric Administration that, due to the nature of their function and reliance on taxpayer dollars, have steadily invested in infrastructure to make the

1. Joan A. Sieber, "Data Sharing in a Historical Perspective", Social Science Space (blog), 2015, https://www.socialsciencespace.com/2015/09/data-sharing-in-historical-perspective/.
2. Carol Tenopir et al., "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLoS One* 10, no. 8 (2015).
3. SPARC, "Data Sharing Requirement by Federal Agencies," http://datasharing.sparcopen.org/.
4. FAIRsharing, "Data Sharing," https://fairsharing.org/policies/.
5. Ibid.
6. Anna Scott, "What is 'Open Data' and Why Should We Care?", Open Data Institute, 2017, https://theodi.org/article/what-is-open-data-and-why-should-we-care/.
7. SPARC, "Open Data," https://sparcopen.org/open-data/.

data they collect increasingly discoverable, accessible, and usable.[8,9] Certain disciplines, such as psychology, have moved to embrace the roles of openness and transparency in their research practices as a method of addressing issues related to research integrity and reproducibility.[10,11] Additionally, questions of what might further incentivize researchers to adopt openness in their research practices continues to be investigated with groups like the Transparency and Openness Promotion (TOP) Committee collaborating with journals and publishers to integrate reproducibility and transparency standards into publication guidelines.[12]

While open data is increasingly encouraged and recognized, not all data lends itself to openness. There remains a need to provide infrastructure that recognizes and supports the protection of certain data while still allowing for discoverability and data sharing. Following an approach best described "as open as possible, as closed as necessary", we seek to meet researchers where they are, to make controlled data sharing easier, and to help them experience its benefits.[13]

Existing barriers

Protected data must be shared in a responsible way. The mechanisms for sharing protected data are significantly more costly than open data sharing due to the additional security, access, and

8. Office of Management and Budget, "M-13-13, Open Data Policy: Managing Information as an Asset," ed. Executive Office of the President (2009).
9. Barack Obama. "Executive Order - Making Open and Machine Readable the Default for Government Information." (2013).
10. Amy Novotney, "Reproducing Results," *Monitor on Psychology* 45, no. 8 (2014).
11. Mark Appelbaum et al., "Journal Article Reporting Standards for Quantitative Research in Psychology: The Apa Publications and Communications Board Task Force Report," *Am Psychol* 73, no. 1 (2018).
12. Brian A. Nosek et al., "Scientific Standards. Promoting an Open Research Culture," *Science* 348, no. 6242 (2015).
13. European Union, "The Transition to an Open Science System: Council Conclusions," (2016).

trust controls associated with them. In particular, the approval processes associated with United States regulations like the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Family Educational Rights and Privacy Act of 1974 (FERPA) that specify how data can be shared for research purposes, require review and authorization by the data owner. This authorization must be given by an individual who is vested with managing access and controls for the data. Within an institution, this person sometimes holds the title “data manager” or “data steward”. These individuals have expertise in the legal regulations, possess detailed knowledge of the data, and are granted the power to authorize release on behalf of the institution. These policies and processes function to protect against irresponsible or unethical release of data which could cause harm to participating individuals. Such safeguards are very much necessary, though the associated procedures may not function as smoothly or quickly as researchers would like.

One example common to academic institutions is human subjects research. The Common Rule (45 CFR part 46, subpart A) specifies protections and rights for all human subjects participating in research. In particular, it specifies the right to confidentiality and privacy. An example of research where the Common Rule would apply is substance abuse research involving illegal activity. The Common Rule, in most cases, provides assurance that participants in these studies are not subject to criminal prosecution as a result of what they might disclose. The Common Rule also requires that participants provide informed consent to the procedures associated with the study. The consent documents specify what the participants are expected to do and potential risks and benefits. It also provides the opportunity for participants to consent to the reuse and/or sharing of their data. For some researchers and institutions, it has been common practice to include clauses in informed consent statements that specify that data will not be shared beyond

the project team or for reanalysis. These statements are unnecessary, limit future use of that data, and restrict participant choice.[14]

In the case of sharing electronic patient health information (ePHI), there are several options for researchers to reuse this data. The first involves de-identifying the data - achieved through removal of specific pieces of identifiable information from datasets - and allows for reuse of the data for purposes beyond those specified in the original consent. In the case of ePHI, the HIPAA Privacy Rule specifies eighteen identifiers that qualify as protected health information. In order to assure de-identification, all 18 identifiable elements must be removed or an expert must determine that the data have been statistically de-identified.[15] Once the dataset is de-identified, the dataset, now referred to as a limited dataset, is no longer subject to HIPAA regulation. The second method for reusing ePHI pertains to reuse of data that contain identifiers. This typically requires approval from authorized officials (e.g., HIPAA Privacy Officer) within an institution, by signing a data use agreement (DUA). The DUA specifies how the data may be used and the applicable data security requirements for the period of time specified in the agreement.

Models for sharing and data discovery

To address these challenges, and to supplement existing processes for enabling protected data sharing (e.g., HIPAA Data Use Agreement), we looked to domain and community-based repositories, information exchanges, and catalogs for examples of infrastructure used to support

14. Michelle N. Meyer, "Practical Tips for Ethical Data Sharing," *Advances in Methods and Practices in Psychological Science* 1, no. 1 (2018).
15. Health and Human Services, "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (Hipaa) Privacy Rule," https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

the sharing of protected data. Since the 1980s, sharing or exchange of data via controlled mechanisms has been a goal of healthcare organizations (i.e., health information exchanges (HIE)).[16] and practiced in the biomedical sciences.[17] Since much of the protected data generated at our university is health and biomedical data, we focused on existing models in this area. Probably the reference data repositories best known in biomedicine for facilitating controlled data sharing are those maintained by the National Center for Biotechnology Information (NCBI).[18] These include databases such as GenBank and dbGap, which provide nucleotide sequences and genotype/phenotype interaction datasets respectively.[19, 20] Other examples include disease specific data sharing portals such as the Alzheimer's Disease Neuroimaging Institute (ADNI) and the Accelerating Medicines Partnership - Alzheimer's Disease (AMP-AD).[21, 22] Clinical data warehouses, like the Indiana Network for Patient Care as part of the Indiana Health Information Exchange (IHIE), aggregate patient and provider data and provide access to these data in compliance with legal requirements.[23, 24] Beyond biomedicine, social science repositories

16. Joshua R. Vest and Larry D. Gamm, "Health Information Exchange: Persistent Challenges and New Strategies," *J Am Med Inform Assoc* 17, no. 3 (2010).
17. Kent Smith, "A Brief History of Ncbi's Formation and Growth.," in *The Ncbi Handbook* (Bethesa MD (US): National Center for Biotechnology Information (NCBI), 2013).
18. NCBI Resource Coordinators, "Database Resources of the National Center for Biotechnology Information," *Nucleic Acids Res* 46, no. D1 (2018).
19. Dennis A. Benson et al., "Genbank," 2018.
20. Kimberly A. Tryka et al., "Ncbi's Database of Genotypes and Phenotypes: Dbgap," ibid.42, no. Database issue (2014).
21. Susanne G. Mueller et al., "Ways toward an Early Diagnosis in Alzheimer's Disease: The Alzheimer's Disease Neuroimaging Initiative (Adni)," *Alzheimers Dement* 1, no. 1 (2005).
22. Richard J. Hodes and Neil Buckholtz, "Accelerating Medicines Partnership: Alzheimer's Disease (Amp-Ad) Knowledge Portal Aids Alzheimer's Drug Discovery through Open Data Sharing," *Expert Opin Ther Targets* 20, no. 4 (2016).
23. Clement J. McDonald et al., "The Indiana Network for Patient Care: A Working Local Health Information Infrastructure. An Example of a Working Infrastructure Collaboration That Links Data from Five Health Systems and Hundreds of Millions of Entries," *Health Aff (Millwood)* 24, no. 5 (2005).
24. HealthIT.gov, "Health Information Exchange," https://www.healthit.gov/topic/health-it-basics/health-information-exchange.

like the Interuniversity Consortium of Political and Social Research[25] as well as national systems, such as the United Kingdom Data Service[26] and the Australian National Data Service[27], provide models for data indexing and sharing of sensitive datasets. These systems differ in their scope, purpose, intended audience, and access mechanisms, but provide infrastructure and governance examples that demonstrate how to establish policies, procedures, and processes for controlled data sharing. Additionally, frameworks such as the "Five Safes" are helpful for thinking about the aspects of carefully controlled pathways for sharing data, especially the balance between managerial and statistical control mechanisms.[28]

A key way that the NCBI, ADNI, and ICPSR platforms differ from a data catalog is that they contain the data files. A catalog typically refers to a collection of metadata that describe objects (in this case, datasets) that are stored elsewhere. A data catalog could also be described as a "registry" or "index" of datasets. As discussed earlier, there are existing - and increasing - examples of United States government institutions that maintain data catalogs as discovery tools for internal and external stakeholders.[29] A 1984 paper from the National Aeronautics and Space Administration (NASA) describes their climate data catalog as intended to alleviate "a major problem confronting climate researchers and other potential users of NASA-climate related data [which is] determining what data exist, or are planned which are appropriate to support their research efforts".[30] This issue of discoverability of datasets is a primary reason for these efforts

25. Interuniversity Consortium of Political and Social Research (ICPSR), https://www.icpsr.umich.edu/icpsrweb/.
26. United Kingdom (UK) Data Service, https://www.ukdataservice.ac.uk/.
27. Australian National Data Service (ANDS), https://www.ands.org.au/.
28. Felix Ritchie, "The 'Five Safes': A Framework for Planning, Designing and Evaluating Data Access Solutions," (Zenodo2017).
29. Data.gov, "Open Government," https://www.data.gov/open-gov/.
30. Mary G. Reph, "Nasa Climate Data Catalog," (1984).

at our institution - as it has been for other libraries and organizations that inspired and informed our work.[31,32]

The fundamental component of a data catalog's functionality is tied to its metadata. Often metadata are heavily dependent on the technical infrastructure used to operationalize the data catalog, which can lead to certain challenges and limitations. Metadata standards continue to evolve in regards to describing datasets.[33,34] These efforts are supplemented by current work that aims to identify metadata elements that operationalize sharing of protected data in particular.[35] Emerging methods for "tagging" sensitive datasets further contribute to our goals of building a data catalog that would direct to sensitive and non-sensitive datasets generated at IUPUI.[36]

We strove to keep these existing initiatives and models in mind when developing the catalog, using them, from the onset, to address challenges that researchers on our campus may experience in their sharing and reuse of protected data. The design of the Data Catalog was shaped by our priorities to increase visibility and make data sharing as easy as possible for the researcher. To that end, the Data Catalog as a service encompasses the Data Catalog platform as well as a set of

31. Kevin Read et al., "Promoting Data Reuse and Collaboration at an Academic Medical Center.," *International Journal of Digital Curation* 10, no. 1 (2015).
32. Lucila Ohno-Machado et al., "Finding Useful Data across Multiple Biomedical Data Repositories Using Datamed," *Nature Genetics* 49, no. 6 (2017).
33. Susanna A. Sansone et al., "Dats, the Data Tag Suite to Enable Discoverability of Datasets," *Sci Data* 4 (2017).
34. World Wide Web Consortium, "Data Catalog Vocabulary (Dcats)," https://www.w3.org/TR/vocab-dcat/.
35. Sam Grabus and Jane Greenberg, "Toward a Metadata Framework for Sharing Sensitive and Closed Data: An Analysis of Data Sharing Agreement Attributes" (paper presented at the MTSR: Research Conference on Metadata and Semantic Research, Tallin Estonia, 2017).
36. Latanya Sweeney, Merce Crosas, and Michael Bar-Sinai, "Sharing Sensitive Data with Confidence: The Datatags System," *Technology Science* (2015), https://techscience.org/a/2015101601.

policies, processes, and procedures to guide creators and users through the negotiation as easily as possible.

IUPUI pilot/case study

IUPUI was founded at the request of Governor Richard Lugar to establish a "great state university in Indianapolis" the capital of Indiana. Initially, the university was created in 1969 by merging schools and programs at the Indiana University Indianapolis campus and the Purdue Indianapolis Extension Center. The campus is now the premier urban research university in Indianapolis, comprised of 18 schools with over 200 degrees offered.[37] As of the fall of 2017, there were nearly 30,000 students enrolled and approximately 2,800 faculty. IUPUI researchers were awarded $428.9 million in external funding in 2015-2016. IUPUI is unique for its focus on community engagement, having the first School of Philanthropy in the world, and having won the Higher Education Excellence in Diversity Award six years in a row.

To support the research mission of the university, the Data Catalog project intends to enhance the discovery and reuse of data created by IUPUI researchers. Since much of the research data generated on our campus is protected in some way (e.g., HIPAA, FERPA, the Common Rule or 45 CFR Part 46, subpart A), an open data repository is not a viable option. Despite the restrictions associated with legal protection, many researchers would like to share their data in a controlled manner.[38] As mentioned earlier, the HIPAA Privacy Rule offers one such mechanism for sharing health data through use of a data use agreement. Associated with this mechanism are

37. Indiana University Purdue University Indianapolis (IUPUI), "History," https://www.iupui.edu/about/history.html.
38. Tenopir et al.

processes and procedures for how requests are reviewed, fulfilled, and documented. While the requirements for other types of protected data are typically less stringent, our pilot program focuses on the sharing of protected health data as a primary use case. This is due to the established policies surrounding it, which enable us to model policies for the Data Catalog after this approach in combination with those used by other domain and institutional data catalogs. The chosen focus on discovery and enabling reuse has guided our decisions regarding infrastructure, metadata, workflow, and policy.

The impetus for this service arose from ongoing conversations with the Director of the Clinical Data Management team in the Indiana University School of Medicine/Richard M. Fairbanks School of Public Health Department of Biostatistics. This team is faced with the challenge of storing research data associated with completed/closed clinical trials. While there is a requirement to store the data, these data are assets that are potentially useful for additional research and the studies contain unique observations that cannot be reproduced. However, as mentioned earlier, the informed consent provided by participants may not allow sharing of data beyond the original project and personnel. Yet, in some cases, there may be options for controlled data sharing that do not violate the original consent terms and approved IRB protocol. The IUPUI Data Catalog could enable this path for reuse of data from these studies, thus providing an opportunity to engage with researchers about the benefits of planning for sharing in advance. This is particularly important as it relates to addressing how data will be shared in the IRB protocol and consents. Engaging early in the research process is also a valuable opportunity to discuss how tracking the reuse of data effectively can support advancement in their careers.

In the course of providing data services to the IUPUI campus, we have encountered other situations in which researchers would like to share their data but cannot do so in a completely open manner. Some examples include geographical information system (GIS) drone data, archeology site data, interview transcripts on sensitive topics, as well as data gathered in partnership with community organizations and local businesses. We are also reaching out to bench scientists who are not ready to share openly and/or who have data that fall outside the scope of existing domain or subject repositories. The pilot is limited to data owned by Indiana University and for which the University and affiliated researchers have the right to share data; this means that we will not extend this pilot to data generated through industry sponsored awards (e.g., clinical trials) or contracted work for the state, county, or city.

Staffing and resources

This project is led by two data librarians at the IUPUI campus, each representing a different library. Erin Foster is the Data Services Librarian in the Ruth Lilly Medical Library, part of the Indiana University School of Medicine, and whose role is to serve faculty, staff, and students within the medical school. Heather Coates is the Digital Scholarship and Data Management Librarian in the University Library Center for Digital Scholarship. University Library is responsible for serving fifteen schools across the IUPUI campus, excluding Medicine, Dentistry, and Law since each of these schools have their own dedicated library. Both serve as Data Catalog Managers and are the main contacts for those interested in contributing to and/or using the Data Catalog.

University Library launched an institutional repository IUPUI ScholarWorks in 2003.[39] In 2013, an institutional data repository IUPUI DataWorks was launched.[40] These repositories, along with other systems used by the Center for Digital Scholarship, are collaboratively maintained by internal IT staff, one of whom is a DSpace committer. The Center staff manage the services associated with these platforms, although collaboration with librarians across campus to provide customized service is common.

Infrastructure

The institutional data repository, IUPUI DataWorks, is built on DSpace. We selected DSpace for the data repository due to deep technical experience using DSpace for IUPUI ScholarWorks. Despite this, we anticipated that repository platforms better suited to meeting the needs of research data would be developed in the next decade. Thus, the data repository is operated in a separate instance of DSpace than our institutional repository. Recognizing the limitations of the DSpace platform, we opted to stick closely to the core code, rather than creating local customizations that would present challenges when upgrading or migrating from DSpace.

However, as we have learned more about the needs of researchers at IUPUI and existing models for controlled data sharing, we recognized that customization would be necessary. This has been done primarily in connection with the "Request a Copy" feature in DSpace.[41] This feature provides users with a mechanism to 'request' file(s) that are not openly accessible in the repository. Either the depositor or the repository manager must approve the request in order to

39. IUPUI ScholarWorks, https://scholarworks.iupui.edu/.
40. IUPUI DataWorks, https://dataworks.iupui.edu/.
41. Bram Luyten, "Request a Copy," DuraSpace, https://wiki.duraspace.org/display/DSDOC5x/Request+a+Copy.

provide access to the file(s). In order to make contextual files such as project and data documentation open, while still enabling the Request a Copy mechanism to function, we have configured the Collection level policies to restrict the ability of anyone to view the files (i.e., called a bitstream in DSpace). Once an item has been deposited, we will manually override the file level permissions to make the documentation files readable by anyone. An empty file, serving as a placeholder for the data, will remain restricted. When a user clicks on the orange lock icon next to the placeholder file, they will be directed to a form to initiate the request process (see Policies, processes, & procedures section).

The scope of our customization is necessarily limited by the capacity and staffing available. While we have internal developers and DSpace expertise, this project competes with support of ongoing services, such as a thriving institutional repository, and many other projects that demand developer and librarian time. Thus, we selected a satisficing approach for this pilot and have, as much as possible, attempted to make the service platform agnostic.

Policies, processes, procedures

Much of the effort necessary to develop the Data Catalog has been invested in developing policies, processes, and procedures. The Data Catalog is a service composed of a platform and a set of pathways for sharing data. In many cases, the policies, processes, and procedures that make up each pathway need to be defined and approved before we can accept data into the catalog. Where regulations and procedures exist (e.g., HIPAA Data Use Agreements), we identified ways to integrate the Data Catalog with existing processes. We also looked at successful models such as the Interuniversity Consortium of Political and Social Research (ICPSR) and other domain specific examples mentioned earlier (e.g., NCBI databases, ADNI,

AMP-AD). Policies help potential users to understand the scope and purpose of the service. Clearly communicating what will happen helps researchers to navigate the deposit and request processes and is a key part of building trust in the system. We have developed workflow diagrams to facilitate communication and to identify gaps in the procedures (Figures 1 & 2). The use of checklists help to fill these gaps and ensure that all required controls are followed.

Generally, each data sharing path provides access to data for a limited period of time to conduct specific research analyses, after which the data are appropriately destroyed. Ownership is retained by the Data Depositor(s) and restrictions regarding the Data Requester(s) use and redistribution of the data are described in the appropriate agreement. Where necessary, these pathways allow for customization. Let's take ePHI data as an example. Though Indiana University has a template agreement and a designated signature authority, there are elements of the data transfer process that can vary. A Depositor, or researcher depositing data into the Data Catalog, can choose where to store the data that will be available for reuse. They can do so on an Indiana University approved storage platform used by their department or group, with the assurance that it will not be discarded until the terms of the deposit agreement have passed. Or they may choose to have the Data Catalog Managers deposit a copy of the data package to a Box Health Data Account, which is university-approved for storing critical data (e.g., PHI), provided appropriate HIPAA-compliant workflows are used. For datasets that are static (e.g., completed projects, or fully cleaned and transformed for analysis), storage in Box Health Data Accounts, which are controlled by IT administrators, ensures continuity and long-term access to the data even after researchers have left the institution. However, in some cases, it may not be feasible (e.g., size of the data, distribution) to store a packaged copy of the data in Box at IU. In these

cases, the deposit agreement will include an additional statement that the Data Creator(s) or a designated unit within the University will retain the data described.

When the data to be shared are not subject to HIPAA regulations, a non-HIPAA Data Use Agreement will be used. Both the DUA and the pathway allow for customization of the permissions granted, security requirements, restrictions, and training requirements. The template DUA will include language appropriate for the most common options. In some cases, specific customizations may require review and approval by General Counsel. Release of data through this mechanism may also require authorization by the appropriate institutional staff (e.g., data managers or data stewards) at both the providing and receiving institutions. One example is in the sharing of critical data, for which special security requirements may be included in the non-HIPAA DUA.

The third pathway developed is a Data License. This is appropriate for sharing data that do not have additional security requirements for storing and managing data. Data License Agreements (DLA) typically clarify ownership of the data, the permissions for use, restrictions on use and distribution, expiration date of the license, and disposal requirements. The core conditions in the template Data License Agreement include those mentioned previously, as well as the requirement for attribution in the form of citation. The Data License may include specific conditions of use, within reason, set by the Data Owner(s).

In order to support these pathways to data sharing, we have developed several new processes and procedures. We use the term processes to refer to the general way that data will be deposited and

transferred. The workflow diagrams created (Figures 1 & 2) document these processes in a user-friendly way that communicates to potential Depositors and Requesters. In combination with the deposit checklist and request checklist, these documents help to set expectations about the service. These workflow diagrams, user documentation, and checklists will be publicly available and linked from the Data Catalog entry page to maximize transparency and help us build trust of our users.

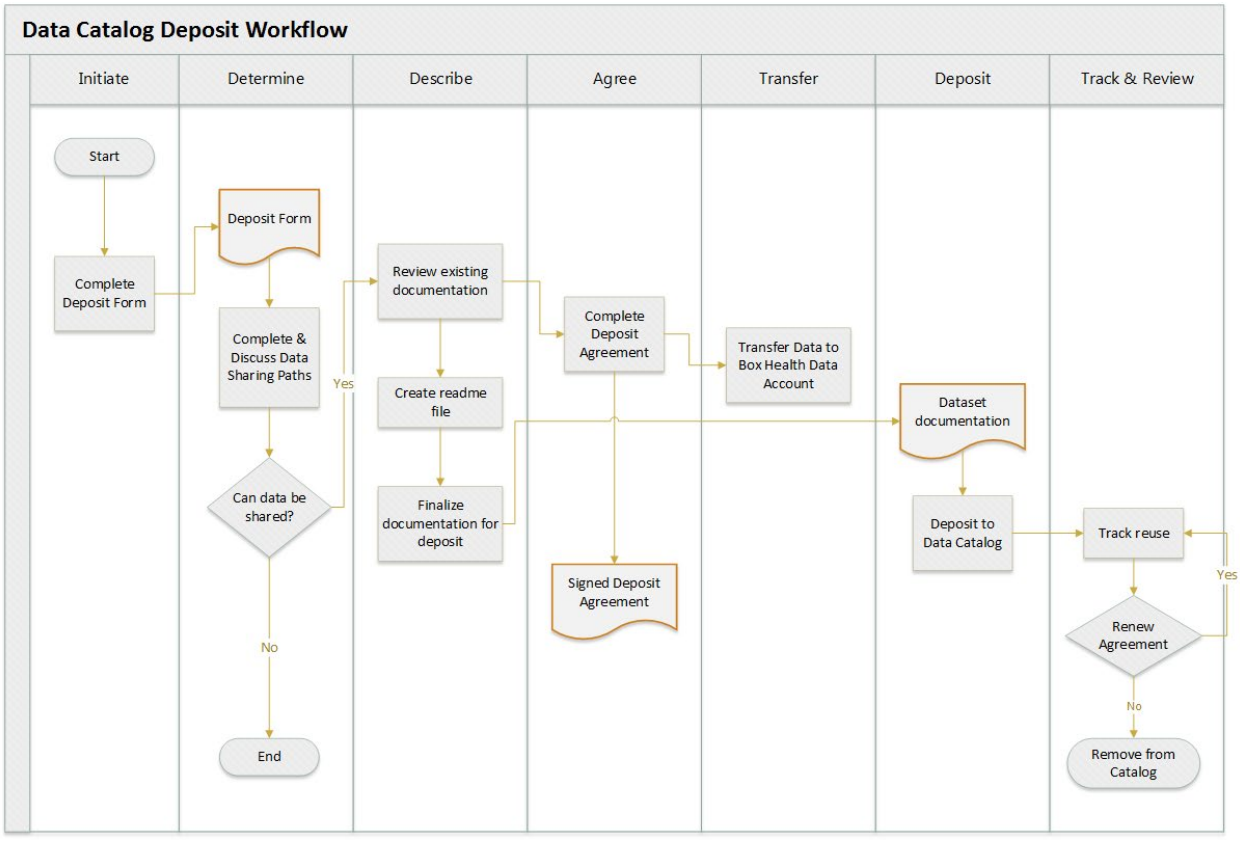

Figure 1: Deposit process

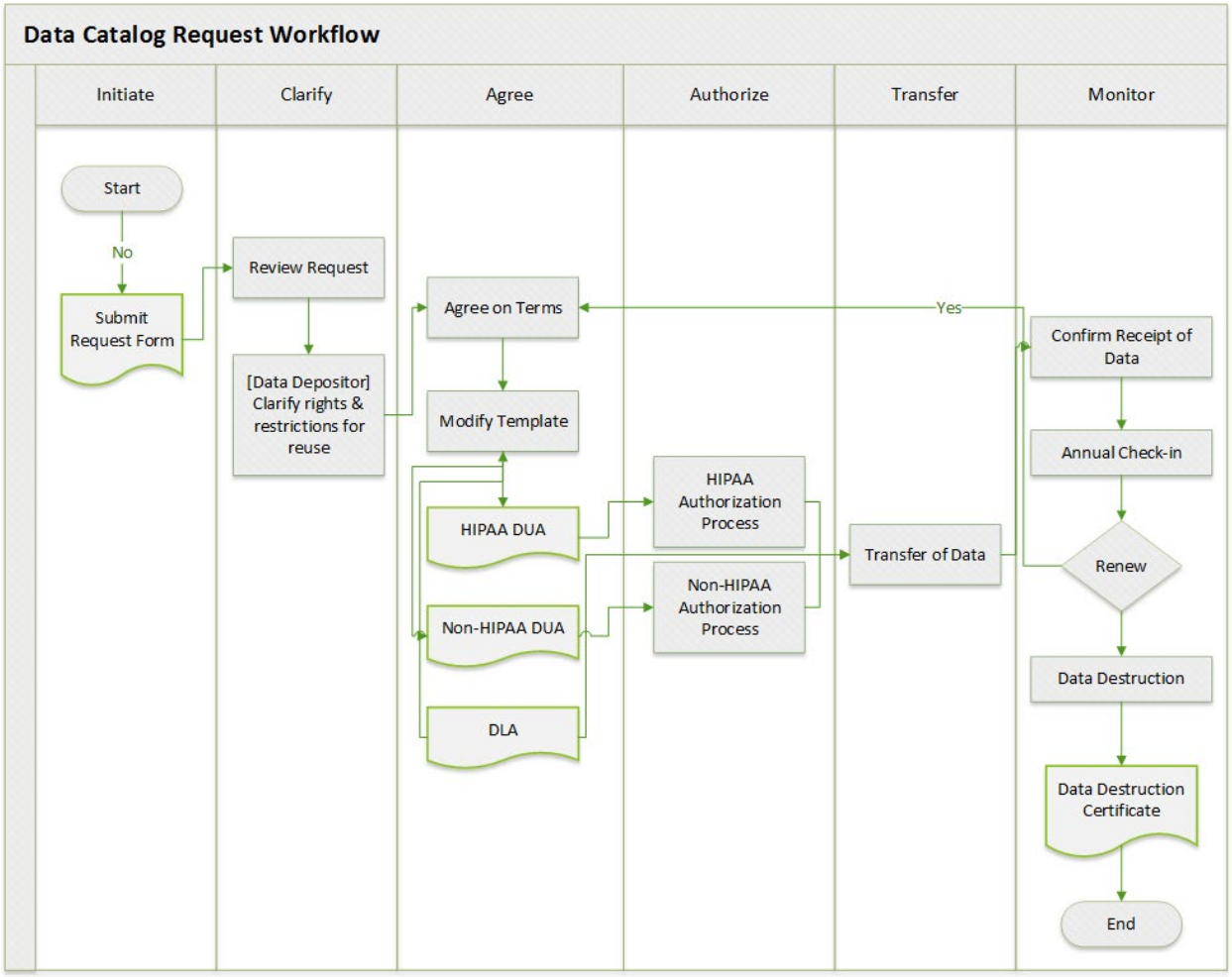

Figure 2: Request process

In contrast to the generalized processes described above, the Data Catalog Managers require detailed procedures and checklists describing the tasks in order to execute and manage the deposit and request processes. An electronic checklist will be completed for each potential data deposit and stored in REDCap (https://www.project-redcap.org/), a secure data management and data capture tool. This strategy enables us to track progress and make note of pertinent information while staying compliant with the relevant protections associated with various datasets. When a dataset is requested via the Data Catalog, the review and determination process will follow one of the three pathways described above. The use of these checklists for each

pathway will enable consistency in our procedures, as well as help us to document and monitor trends over time.

Another example of the need for specific procedures is the work being done to develop HIPAA compliant workflows for both deposit into the Box Health Data Account and transfer from it to data reusers. We have consulted with an analyst in the Indiana University Center for Advanced Cybersecurity Research[42] to develop workflows that both ensure systems are properly secured and clearly describe ways for people to interact with these systems that do not compromise the security and privacy of the data. Meeting that level of compliance as data flow across multiple systems requires a workflow designed to maintain adequate security at every step in the process. The HIPAA related checklists will be informed by these workflows to support effective management of the request process.

Finally, since DUA and DLA are typically term limited, we have developed a schedule and accompanying checklist for following up with users to ensure compliance with the agreement. This will take place annually, at a minimum. Prior to the agreement expiration, we will contact the user to remind them of the appropriate protocol for disposal of the data. We will also require that the user complete and return a Data Destruction Certificate. Scheduled procedures will be automated as much as possible using longitudinal events in REDCap.

42. Center for Applied Cybersecurity Research, Indiana University, https://cacr.iu.edu/.

Metadata

The metadata contained within the data catalog exists to enable discovery, communicate rights information, and enable controlled data sharing. It is not sufficient for preservation, as we do not anticipate that most of the datasets described in the IUPUI Data Catalog will be retained indefinitely. If some are determined to be good candidates for preservation, further description and curation will then be carried out. Since the metadata structure and crosswalk for Dublin Core to DSpace already exists, only slight modifications were necessary. The first task was to incorporate information about the contact person for the data item. Though Dublin Core does not have an exact field for the contact information needed for display in the Data Catalog, many repositories use the dc.contributor field to store such information (e.g., Dataverse[43], Omeka[44], and the University of Minnesota DRUM[45]). In consultation with our metadata librarian, we determined this was the most consistent with community practice and the easiest to implement without customizing the repository code. The metadata will also specify the associated legal protections, which affects the possible data sharing pathways available. Due to ongoing development of a new platform for the institutional data repository - and, therefore, the Data Catalog - to a different repository platform, we determined that significant effort to customize metadata was not an effective use of limited resources.

Next Steps

Over the next 12-18 months, we plan to expand recruitment of pilot participants beyond biomedical researchers, with the goal of testing this service against a diverse range of protected

43. The Dataverse Project, https://dataverse.org/.
44. Omeka, https://omeka.org/.
45. Data Repository for the University of Minnesota (DRUM), https://www.lib.umn.edu/datamanagement/drum.

data. In the process of working with pilot participants, we will listen carefully to describe how they experience both barriers and drivers for sharing their data. This information will help us to gain insight into their experiences, which, in turn, will help us to test and refine our service model, particularly the approval processes and internal procedures. It will also inform how this service might scale up, or not.

IUPUI University Library, Indiana University Bloomington Libraries, and University Information Technology Services have been collaborating to develop a common Samvera-based solution for hosting a range of digital content such as video, images, paged media, and, more recently, research data. The features necessary to create a functional catalog will inform the data repository evaluation and development processes, particularly where these features align with data repository functions. For example, in order to make the Request a Copy feature in DSpace function as a user might expect, we have had to create collection level policies that restrict access and expand the deposit workflow to modify bitstream access after the deposit is submitted. Neither of these choices are optimal and require additional time on the part of library staff to alter and check. Ideally, a new repository platform would allow configuration of bitstream level access during the deposit process. In spite of these workarounds, the Data Catalog will facilitate controlled sharing of protected data and allow willing researchers to personally experience the benefit of data sharing and citation.

Challenges

Developing this pilot has highlighted three sets of challenges for scaling the data catalog approach. The first is that sharing of protected data cannot yet be automated. Though projects,

like the DataTags Project[46] are promising and aim to streamline the sharing of protected data, the associated legal protections and authorizations require human review and approval. Additionally, ingest procedures are heavily dependent on human review to confirm that requirements are met and that the files made public do not in fact contain protected data. Even if we were able to clearly articulate the policies so that they could be automated, ethical considerations demand that humans who understand the potential for harm which inappropriate access may cause - and who recognize their responsibility to protect affected individuals - have full control in the release of data.[47]

A second area of challenge is the varied way in which laws apply to research data. In the case of intellectual property law, they may have a strong influence on how data are shared, depending on the type of data.[48] In other cases, where the data are deemed factual, intellectual property laws do not apply to the data themselves, but may apply to the arrangement and organization of the data. These issues become more difficult as datasets are aggregated and derived from existing datasets with differing legal protections and requirements.

Finally, considering the proliferation of institutional and domain repositories for publications and data, we must consider how to function as an integrated part of the broader data exchange and scholarly communication ecosystems. What does it mean to be integrated into those ecosystems? What does it mean to be interoperable? Do we focus on the interoperability of metadata, or

46. Sweeney, Crosas, and Bar-Sinai.
47. Dorothea Salo, "The Memory of Research," in *2018 Sage Assembly* (Seattle, Washington, USA2018).
48. Michael W. Carroll, "Sharing Research Data and Intellectual Property Law: A Primer," *PLoS Biol* 13, no. 8 (2015).

should we also consider the related metrics? What data sharing models might be sustainable for such a broad range of stakeholders - academic research institutions, research institutes, professional societies, government agencies, public entities, and others? Those managing health data subject to HIPAA, in part driven by the HITECH Act and associated funding, have implemented approaches such as data warehousing to manage and make usable massive amounts of data from varied sources. Though the specific requirements and regulations of HIPAA do not apply to other types of data, we can look to the lessons demonstrated by health information exchanges (HIE) failures and successes for guidance and potential models.[49] The Coalition of Open Access Repositories (COAR) report on next generation repositories offers food for thought about the types of functionality that repositories should have to participate in "a distributed, globally networked infrastructure for scholarly communication".[50] But it remains to be seen how researchers actually engage with and navigate this messy, emergent data sharing environment.

Conclusion

We believe that institutional data repositories need to be able to support controlled data sharing as well as open data. The Data Catalog at IUPUI exists as an option between open data repositories and secure data enclaves, enabling the registration and discovery of research data for reuse and citation. Though similar projects exist across Indiana University, they tend to focus on big data. Some of these include the Indiana University Network Science Institute Web of Science

49. Vest and Gamm.
50. Confederation of Open Access Repositories, "Next Generation Repositories: Behaviours and Technical Recommendations of the Coar Next Generation Repositories Working Group," (2017).

enclave,[51] a secure data enclave for ePHI, and the Addiction Data Commons,[52] which is in development. Unlike these examples, the Data Catalog is a solution designed for individual researchers and teams generating protected data. As we have described, controlled data sharing requires the development of standard processes and infrastructure to facilitate these mechanisms in a sustainable way. This is no small task, particularly in light of continued budgetary and staffing constraints. Development of the policies, processes, and workflows for this project has required approximately 100 hours of work. This has included identifying the key decision makers, discussing needs with stakeholders, gathering information about current practices, exploring the technical feasibility for certain features within the DSpace platform, and obtaining expert review of the workflows. Still, this work is not yet complete. Development of policy and agreement documents continues, including final approval by university general counsel. We anticipate that this work will be complete by the end of 2018 in order to launch the service in early 2019. Despite the substantial time investment, we feel the Data Catalog is an important service for incrementally increasing the openness and transparency of research conducted at IUPUI and offers a model for libraries with similar goals. Even for libraries without a dedicated data repository, this approach can be modified to take advantage of other storage options to make available the protected data generated at their institutions. By providing increased opportunity for data sharing, we strive to expose the institution's researchers to the benefits of data sharing in hopes that they may adopt more open research practices and become champions of openness.

51. Indiana University Network Science Institute, "Web of Science," https://iuni.iu.edu/resources/web-of-science.
52. Grand Challenges - Addiction, "Phase 1," Indiana University, https://grandchallenges.iu.edu/addiction/phase-one.html.

Bibliography

Addiction - Grand Challenges. "Phase 1." Indiana University, https://grandchallenges.iu.edu/addiction/phase-one.html.

Appelbaum, Mark, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. "Journal Article Reporting Standards for Quantitative Research in Psychology: The Apa Publications and Communications Board Task Force Report." [In eng]. *Am Psychol* 73, no. 1 (Jan 2018): 3-25.

Australian National Data Service (ANDS). https://www.ands.org.au/.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D. Pruitt, and Eric W. Sayers. "Genbank." [In eng]. *Nucleic Acids Res* 46, no. D1 (Jan 4 2018): D41-d47.

Carroll, Michael W. "Sharing Research Data and Intellectual Property Law: A Primer." [In eng]. *PLoS Biol* 13, no. 8 (Aug 2015): e1002235.

Center for Applied Cybersecurity Research. Indiana University, https://cacr.iu.edu/.

Confederation of Open Access Repositories. "Next Generation Repositories: Behaviours and Technical Recommendations of the Coar Next Generation Repositories Working Group." 2017.

Data.gov. "Open Government." https://www.data.gov/open-gov/.

Data Repository for the University of Minnesota (DRUM). https://www.lib.umn.edu/datamanagement/drum.

European Union. "The Transition to an Open Science System: Council Conclusions." 2016.

FAIRsharing. "Data Sharing." https://fairsharing.org/policies/.

Grabus, Sam, and Jane Greenberg. "Toward a Metadata Framework for Sharing Sensitive and Closed Data: An Analysis of Data Sharing Agreement Attributes." Paper presented at the MTSR: Research Conference on Metadata and Semantic Research, Tallin Estonia, 2017.

Health and Human Services. "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (Hipaa) Privacy Rule." https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

HealthIT.gov. "Health Information Exchange." https://www.healthit.gov/topic/health-it-basics/health-information-exchange.

Hodes, Richard J., and Neil Buckholtz. "Accelerating Medicines Partnership: Alzheimer's Disease (Amp-Ad) Knowledge Portal Aids Alzheimer's Drug Discovery through Open Data Sharing." [In eng]. *Expert Opin Ther Targets* 20, no. 4 (2016): 389-91.

Indiana University Purdue University Indianapolis (IUPUI). "History." https://www.iupui.edu/about/history.html.

Indiana University Network Science Institute. "Web of Science." https://iuni.iu.edu/resources/web-of-science.

Interuniversity Consortium of Political and Social Research (ICPSR). https://www.icpsr.umich.edu/icpsrweb/.

IUPUI DataWorks. https://dataworks.iupui.edu/.

IUPUI ScholarWorks. https://scholarworks.iupui.edu/.

Luyten, Bram. "Request a Copy." DuraSpace, https://wiki.duraspace.org/display/DSDOC5x/Request+a+Copy.

McDonald, Clement J., J. Marc Overhage, Michael Barnes, Gunther Schadow, Lonnie Blevins, Paul R. Dexter, Burke Mamlin, and INPC Management Committee. "The Indiana Network for Patient Care: A Working Local Health Information Infrastructure. An Example of a Working Infrastructure Collaboration That Links Data from Five Health Systems and Hundreds of Millions of Entries." [In eng]. *Health Aff (Millwood)* 24, no. 5 (Sep-Oct 2005): 1214-20.

Meyer, Michelle N. "Practical Tips for Ethical Data Sharing." *Advances in Methods and Practices in Psychological Science* 1, no. 1 (2018): 131-44.

Mueller, Susanne G., Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. "Ways toward an Early Diagnosis in Alzheimer's Disease: The Alzheimer's Disease Neuroimaging Initiative (Adni)." [In eng]. *Alzheimers Dement* 1, no. 1 (Jul 2005): 55-66.

NCBI Resource Coordinators. "Database Resources of the National Center for Biotechnology Information." [In eng]. *Nucleic Acids Res* 46, no. D1 (Jan 4 2018): D8-d13.

Nosek, Brian A., G. Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, S. J. Breckler, Stuart Buck*, et al.* "Scientific Standards. Promoting an Open Research Culture." [In eng]. *Science* 348, no. 6242 (Jun 26 2015): 1422-5.

Novotney, Amy. "Reproducing Results." *Monitor on Psychology* 45, no. 8 (2014).

Obama, Barack. "Executive Order - Making Open and Machine Readable the Default for Government Information." (2013).

Office of Management and Budget. "M-13-13, Open Data Policy: Managing Information as an Asset." edited by Executive Office of the President, 2009.

Ohno-Machado, Lucila, Susanna A. Sansone, George Alter, Ian Fore, Jeffrey Grethe, Hua Xu, Alejandra Gonzalez-Beltran*, et al.* "Finding Useful Data across Multiple Biomedical

Data Repositories Using Datamed." [In English]. *Nature Genetics* 49, no. 6 (2017): 816-19.

Omeka. https://omeka.org/.

Read, Kevin, Jessica Athens, Ian Lamb, Joey Nicholson, Sushan Chin, Junchuan Xu, Neil Rambo, and Alisa Surkis. "Promoting Data Reuse and Collaboration at an Academic Medical Center.". *International Journal of Digital Curation* 10, no. 1 (2015): 260-67.

Reph, Mary G. "Nasa Climate Data Catalog." [In eng]. (1984).

Ritchie, Felix. "The 'Five Safes': A Framework for Planning, Designing and Evaluating Data Access Solutions." Zenodo, 2017.

Salo, Dorothea. "The Memory of Research." In *2018 Sage Assembly*. Seattle, Washington, USA, 2018.

Sansone, Susanna A., Alejandra Gonzalez-Beltran, Phillipe Rocca-Serra, George Alter, Jeffrey S. Grethe, Hua Xu, Ian M. Fore*, et al.* "Dats, the Data Tag Suite to Enable Discoverability of Datasets." [In eng]. *Sci Data* 4 (Jun 6 2017): 170059.

Scott, Anna. "What Is 'Open Data' and Why Should I Care?" In *Open Data Institute*: Open Data Institute, 2017.

Sieber, Joan A. "Data Sharing in a Historical Perspective." In *Social Science Space*, 2015.

Smith, Kent. "A Brief History of Ncbi's Formation and Growth.". In *The Ncbi Handbook*. Bethesa MD (US): National Center for Biotechnology Information (NCBI), 2013.

SPARC. "Data Sharing Requirement by Federal Agencies." http://datasharing.sparcopen.org/.

SPARC. "Open Data." https://sparcopen.org/open-data/.

Sweeney, Latanya, Merce Crosas, and Michael Bar-Sinai. "Sharing Sensitive Data with Confidence: The Datatags System." *Technology Science* (2015). https://techscience.org/a/2015101601.

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." [In eng]. *PLoS One* 10, no. 8 (2015): e0134826.

The Dataverse Project. https://dataverse.org/.

Tryka, Kimberly A., Luning Hao, Anne Sturcke, Yumi Jin, Zhen Y. Wang, Lora Ziyabari, Moira Lee*, et al.* "Ncbi's Database of Genotypes and Phenotypes: Dbgap." [In eng]. *Nucleic Acids Res* 42, no. Database issue (Jan 2014): D975-9.

United Kingdom (UK) Data Service. https://www.ukdataservice.ac.uk/.

Vest, Joshua R., and Larry D. Gamm. "Health Information Exchange: Persistent Challenges and New Strategies." [In eng]. *J Am Med Inform Assoc* 17, no. 3 (May-Jun 2010): 288-94.

World Wide Web Consortium. "Data Catalog Vocabulary (DCATS)." https://www.w3.org/TR/vocab-dcat/.