National Institutes of Health
National Institute of Diabetes and Digestive and Kidney Diseases
Data Management and Sharing Webinar Series
Virtual Meeting

Session 2: Finding a Repository for Your Data
May 31, 2023

SUMMARY

Welcome and Introductions

Michelle Engle, Research Triangle Institute (RTI) International

Dr. Michelle Engle, Bioinformatics Specialist, RTI International, welcomed the participants to the second part of the Data Management and Sharing (DMS) Webinar Series, hosted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) to educate and connect with the scientific community about the National Institutes of Health (NIH) DMS policy. Additional resources from NIDDK to support investigators implementing the DMS policy, including links to the webinar series and first session materials on writing a DMS plan, are available on the NIDDK DMS page.

Finding a Repository for NIDDK Study Data Sets

Jeran Stratford, RTI International

Dr. Jeran Stratford, Bioinformatician and Data Scientist, RTI International, commented that the previous webinar built a solid foundation for the series by outlining the “who,” “what,” “where,” and “when” of data sharing and reviewed the required elements of a DMS plan. Dr. Stratford encouraged attendees to view the recording of the previous session for more information on the six required elements. This webinar focuses on the “where” attributes—selecting a quality repository to deposit scientific data and determining how long the data will be stored. This corresponds to the Data Preservation, Access, and Associated Timelines element of the DMS plan.

Scientific data come in many formats and have significant reuse value. Historically, scientific data have been shared only through publications or by requesting access from the authors of a publication. Data repositories now are emerging as the leader in data sharing, providing access and making the data findable, accessible, interoperable, and reusable (FAIR). A wide variety of databases are available, with different features for making data more FAIR. Desirable repository characteristics include long-term data storage, protection of research participant privacy and confidentiality, mechanisms to allow access to the data, maintaining appropriate provenance, and housing rich metadata associated with the data sets. Metadata and data standards will be discussed further in webinar 3 in the series.

Many types of repositories are available. Examples include generalist, data type–specific, and disease-specific repositories. They can be hosted by NIH or by public or private institutions. The landscape of repositories is continually evolving in response to investigators’ needs, and the number of repositories is growing. Researchers should evaluate many repositories to determine which is the best fit for their data. The vast number of options makes identifying the appropriate resource for sharing data in line with the DMS policy a potential challenge for NIH investigators.

NIH has created several resources for finding repositories with characteristics that make them highly desirable and make the data FAIR, maximizing the value of shared data. NIH Institutes, Centers, and Offices (ICOs) sometimes specify in the funding announcement which repository should be used. In such cases, investigators should indicate the repository prescribed in the funding announcement in their DMS plan. If a repository has not been specified, the investigator should evaluate available repositories and select one (or more) which best meets their data management and sharing needs. Discipline- and data type–specific repositories are preferred because they generally offer infrastructure and tools to make the data as accessible, browsable, and findable as possible, which may enable and facilitate future reuse. Generalist or institutional repositories may not have such tools and may be appropriate if an investigator's particular data type does not have a dedicated repository.

NIH has created a resource, available at sharing.nih.gov, that lists many existing biomedical repositories. Users can search for repositories based on ICO, data types, or specific characteristics of the data. Other tools for identifying repositories include information from publishers about acceptable repositories, the Registry of Research Data Repositories (re3data), and the NIH Generalist Repository Ecosystem Initiative (GREI), discussed later in this webinar.

To supplement the NIH repositories catalog, NIDDK has created guidance on repository selection for NIDDK studies, which can be found on the NIDDK DMS site. The information on the fourth data element required in the data plan—Data Preservation, Access, and Associated Timelines—describes desirable characteristics of repositories and preferences when selecting a repository. The site also includes links to the NIDDK Central Repository, NIDDK Information Network (dkNET), GREI, and other NIH-supported repositories.

Another resource on the site is the Repository Selection Considerations Tool, which can help investigators identify both repository characteristics and NIDDK preferences. This tool is based on data type, as investigators are encouraged to identify in the DMS plan each type of data and the repository for that data. The tool follows a decision tree model, and a workflow asks specific questions about the data to lead users to information on the available repositories and how to search for them. Once an investigator has identified a potential repository for their data, they will need to review the repository's policies to confirm the study's eligibility for the repository, required identification or de-identification of data, acceptable file formats, data linkage and metadata, and whether access to the data should be controlled or unrestricted.

One key characteristic of many repositories is issuance of a unique persistent identifier (PID) for the data set, which can be used in publications or other reporting to provide access to the data. The PID always points to the data and provides a "treasure map" with metadata as the key to finding, opening, and using the scientific data. PIDs allow data to always be findable, even if the data location changes.

Dr. Stratford reiterated that repositories have become the primary way to share data and ensure they are FAIR, and that choosing an appropriate repository is a multifaceted decision. NIDDK has developed several resources, tools, and guidance, available on the DMS website, to help investigators make this decision. Investigators should reach out to their program officers with any questions.

dkNET DMS Resources

Jeffrey Grethe, University of California, San Diego (UCSD)

Dr. Jeffrey Grethe, Principal Investigator, dkNET, Co-Director, FAIR Data Informatics Laboratory, UCSD, provided an overview of dkNET resources. dkNET is funded by NIDDK to provide a single point of access to information on materials, tools, and available services, including resources related to FAIR data and the new DMS mandate. Dr. Grethe demonstrated how to access the data portal through the FAIR Data Services menu and recommended visiting the area titled "Getting Help With the NIH DMS Plan." The site has an informational area and links to other sites that may be useful. dkNET staff, in collaboration with NIDDK, have developed a list of available repositories categorized by discipline, with direct links to various repositories' information on submission of and access to data, as well as potential costs. dkNET also provides links to FAIRsharing.org, a resource with information on the standards for these repositories.

Several tools and resources are available through dkNET. The program's repository wizard will be released this summer; it rates 80 repositories against FAIR principles and other standards and offers users a decision tree to help select an appropriate repository. The Summer of Data student program occurs annually and teaches undergraduate, graduate, and postdoctoral researchers about the importance of rigor and reproducibility, data management, and FAIR data. Lessons associated with the program are available on dkNET. dkNET also hosts a webinar series with information about informatics resources and repositories, and the archives are freely available via slides and recordings on YouTube.

A new resource this year is the Office Hours program, which focuses on interacting with researchers about the DMS plan; many of the questions received have been about selecting an appropriate repository. Office Hours are held quarterly, and recordings of the first two are available on YouTube. Questions received via Office Hours—as well as through a data repository inquiry form researchers can use to request a discussion with dkNET about repositories—will inform future training efforts. Dr. Grethe emphasized that dkNET is dedicated to helping researchers share their data.

Submitting Resources to the NIDDK Central Repository
Rebecca Rodriguez, NIDDK

Dr. Rebecca Rodriguez, Director, NIDDK Central Repository, explained the prominent role the NIDDK Central Repository plays within NIDDK's data-sharing landscape. NIDDK promotes sharing FAIR resources to increase their scientific value and encourages connections to increase the scientific value of the data. The Central Repository strives to be NIDDK's model of high-quality resource sharing and supports eligible investigators throughout their projects' life cycle by promoting the ethical and equitable archival and redistribution of specimens and data, thus enabling novice and seasoned investigators to test new hypotheses without having to collect new data or biospecimens. This leads to faster development of impactful health outcomes. The Central Repository works continuously to improve the discoverability of all resources under its guardianship and to minimize barriers to resource access.

The Central Repository is in NIDDK's Office of the Director within the Office of Clinical Research Support, which provides the NIDDK community with resources to support NIDDK's mission. The repository is administered by NIDDK staff with contractor support for operations. The biorepository has several accreditations and certifications, with more in progress. The Central Repository serves more than 20 countries, supporting active and concluded interventional or observational clinical studies funded by NIDDK. It provides end users with high-quality quantitative and qualitative clinical phenotype data for a wide range of biospecimens, maintained under uniform and standardized conditions. This facilitates secondary use, which increases the impact of these resources. The Central Repository has more than 5,000 registered users and 3,000 research projects, and has accepted 95 percent of all requests. About 230 public releases citing the repository resources have been received since the repository began tracking.

The Central Repository aligns with FAIR and TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles, and its policies and procedures are consistent with desirable repository characteristics. Recent system enhancements include data package versioning controls, digital object identifiers (DOI), and a metadata map hosted through schema.org, which helps index resources through Google's data set search, FAIRsharing.org, and re3data, thereby improving discoverability. The repository also has improved its business logic and consolidated its documentation, a proactive effort to emphasize its trustworthy standards and demonstrate its commitment to the spirit of the DMS policy.

Dr. Rodriguez recommended that investigators selecting a repository should understand what services the repository offers, how the resources will be managed and shared, and any possible restrictions on eligibility or submission. Those eligible and approved to submit to the Central Repository receive assistance with planning their data archive and DMS plan, as well as guidance on good practices throughout the entire life of the project, consistent with NIDDK and repository policies and procedures. For studies that have been approved to submit to the Central Repository, prior to enrollment of the first participant, investigators must develop, and the NIDDK repository must approve, a plan that describes the timeline for submission and public sharing. The data and biospecimens must be deposited in accordance with approved timelines. Dr. Rodriguez reminded attendees that the submission timelines are being revised to align with the DMS policy and NIDDK's guidance. NIDDK's policies and procedures are designed to encourage timely analysis and reuse of the data and balance the interests of data generators and data users.

The Central Repository is revising its policies around who is eligible to submit. The priority has been extramural NIDDK-funded clinical studies that receive significant programmatic support and ancillary or secondary research directly resulting from NIDDK resources. Some non-NIDDK–funded studies within NIDDK's research mission and of significant impact or benefit to the scientific community may also be eligible to submit. Eligibility does not guarantee acceptance, which will hinge upon compliance with requirements. Dr. Rodriguez showed the sequence of events for data submission, noting that after submission, data are curated before being made publicly available under controlled access.

Dr. Rodriguez reiterated that the full potential of the resources and the spirit of the DMS policy can be fulfilled only if high-quality resources are shared and able to be used in secondary research. The Central Repository ensures that the data-sharing workflow focuses on providing easy access that allows continuous data augmentation, meaning that any additional data or knowledge generated using resources from the repository are deposited back and associated with the originating resources.

NIDDK and the Central Repository are indexing NIDDK-funded studies to support semi-federated discovery of resources deposited elsewhere for those not able to be hosted by the Central Repository. The Central Repository is also partnering with other repositories to reduce redundancies and bridge gaps, improving the data-sharing ecosystem for the scientific community. Dr. Rodriguez pointed out that although the Central Repository helps investigators comply with the DMS policy, investigators must recognize the value of using metadata standards and common data elements to maximize connections in the ecosystem and the impact of their research. Dr. Rodriguez noted the upcoming Central Repository 20th Anniversary Seminar, which focuses on promoting secondary research to accelerate breakthroughs.

NIH GREI

Ishwar Chandramouliswaran, Office of Data Science Strategy (ODSS), NIH

Mr. Ishwar Chandramouliswaran, Program Officer, ODSS, explained that generalist repositories accept data regardless of data type, format, content, or disciplinary focus, providing a low barrier to entry. One of the pillars of the NIH Strategic Plan for Data Science is modernizing the data ecosystem to reduce or eliminate silos, optimize costs, and disentangle data from specific projects for reuse. Modernizing the ecosystem allows better sharing, discovery, and reuse of data and supports the NIH DMS policy.

ODSS first engaged with generalist repositories with a pilot project to identify gaps in data management and sharing support and how easier data sharing would change the culture. A subsequent workshop focused on the "coopetition" concept, which creates ecosystems in which researchers collaborate on common capabilities while providing unique value propositions to their communities. GREI was launched in 2022 based on these efforts and in anticipation of the NIH DMS policy.

The goals of GREI are to develop collaborative approaches for data management and sharing by the generalist repositories and to better enable search and discovery of NIH-funded data in the generalist repositories. The seven GREI awardees—a mix of commercial, academic, and nonprofit repositories—already existed, but the goal of the initiative is to develop collaborative approaches. GREI's mission is to develop a common set of cohesive and consistent capabilities, services, and metrics; to facilitate a social infrastructure across generalist repositories; and to raise awareness of and facilitate FAIR principles.

Objectives include implementing consistent capabilities, improving access to and discovery of NIH-funded data, conducting outreach and training on FAIR data practices, and engaging the research community. Expected outcomes are making data sharing easier, improving discoverability, increasing the reproducibility of research, and encouraging secondary use of data. Additional objectives include aligning with the desirable characteristics of data repositories, developing consistent metadata models between and across repositories, and enhancing the browse and search capability for NIH-funded data sets in those repositories. The repositories also practice quality control, enable connectivity to prevent duplication, implement open metrics, develop educational materials, and provide outreach activities.
Mr. Chandramouliswaran encouraged attendees to visit datascience.nih.gov for more information.

Question and Answer Session

- In response to a question about preferred repository characteristics and characteristics to help researchers get credit for their data, Dr. Stratford suggested that researchers should look for repositories that have a DOI, which can help researchers get credit for and track their data. Repositories also should include indexable and searchable metadata; if a repository requests the same type of metadata that researchers plan to provide, this indicates that other users of that repository also will search for and find that data. Dr. Grethe added that research papers also use DOI, so applying the same citation standards to data sets is important for crediting researchers for sharing their data. Dr. Rodriguez pointed out that some repositories offer Open Researcher and Contributor Identification (ORCID), which can be especially important for researchers with common names, and data sets can be added to a researcher's ORCID record. Mr. Chandramouliswaran added that persistent identifiers are an important part of the public access plan distributed by the White House Office of Science and Technology Policy, so learning to incorporate identifiers into the entire research life cycle is useful.
- When asked if GREI has provisions for researchers who prefer using generalist repositories, Mr. Chandramouliswaran emphasized that repository choice should be driven by the science, but each GREI repository has unique capabilities; he encouraged researchers to contact the repositories for more information. Dr. Rodriguez pointed out that the Central Repository has partnered with one of the GREI repositories to accept data that are not eligible for inclusion in the Central Repository. She recommended that researchers visit dkNET to review the resources or contact the Central Repository for more guidance.
- When asked about sharing large data sets when a designated repository is not available, such as when a new data type is developed, Dr. Rodriguez recommended that researchers contact their program officials, who can advise on adequate repositories; dkNET and Central Repository also offer resources to help researchers find an appropriate repository. Mr. Chandramouliswaran pointed out that some NIH programs use cloud capabilities to share large data types, which can be cross-linked in general repositories to make very large data sets findable. Cloud workspaces also avoid the need for researchers to download and upload the data frequently.

Adjournment

Dr. Engle thanked the panelists and invited attendees to return for the third webinar, on June 27, and the fourth webinar in July.