Is Your Research Data FAIR? Best Practices for Managing Data

On June 4, 2018, the National Institutes of Health (NIH) released a new Strategic Plan for Data Science.

In the Strategic Plan, the NIH outlines five key goals for addressing the storage, cataloging, sharing, and publication needs related to the vast amounts of research data currently being generated in the biomedical sciences. With the direction of a new Chief Data Strategist, the NIH will enact measures to support an efficient biomedical research data infrastructure; promote modernization of the data resources ecosystem; support the development and dissemination of advanced data management, analytics, and visualization tools; enhance workforce development for the biomedical data sciences; and enact appropriate policies to promote data stewardship and sustainability.

Since 2010, Northwestern University has received more than half of its sponsor funding from the Department of Health and Human Services, the parent organization of the NIH, meaning that many Northwestern researchers are producing large amounts of biomedical data. In recent interviews with Northwestern researchers, Galter librarians have discovered that many researchers have concerns about general data management principles, and some have an interest in data sharing and long-term data preservation and retrieval. The concerns are timely, and potential solutions align especially well with the last goal listed in the NIH's strategic plan: enacting appropriate policies to promote data stewardship and sustainability.

How do I Keep my Data FAIR?

As outlined within the fifth goal of the NIH plan, one important policy will be to promote adherence to the FAIR principles for data stewardship (for a definition of FAIR, see below). The FAIR principles represent a consensus among data and information security professionals about best practices to make data freely and safely available. Processes and systems have already been put in place to make FAIR data a reality at Northwestern. One way for investigators to ensure that research data will be available for future re-use is to implement a Data Management Plan at the beginning of the research process. Northwestern researchers can easily create a plan tailored to their own institution and funder using the online DMPTool. Aside from meeting Federal grant requirements, data management plans can also encourage use of file and folder naming conventions and the inclusion of metadata, efforts which will make data more findable and interoperable in the long run.

Some researcher data is findable now through tools currently hosted by Northwestern, such as the Arch institutional repository and Northwestern Medicine's DigitalHub Both have rich record descriptors to enhance data findability, and DigitalHub leverages MeSH to make its records interoperable. Like many modern institutional and data repositories, both tools use unique identifiers (DOIs) to ensure accessibility.

The accessibility principle includes an information security requirement, and that too is being met for Northwestern's research datasets. A great example of this is seen in the Northwestern Medicine Enterprise Data Warehouse (NMEDW), a joint initiative across Feinberg and Northwestern Memorial Healthcare Corporation. Its mission is to create a single, comprehensive, and integrated repository of all clinical and research data sources on the campus to facilitate research, clinical quality, healthcare operations, and medical education. Data from the NMEDW is released in adherence with a Permissible Use Policy, and is protected by Data Security Plans for all information used in clinical research. Northwestern University Information Technology publishes and enforces additional policies for the protection of data.

By offering tools, policies, and techniques to make researchers' data findable, accessible, interoperable, and re-usable, NU libraries, NUCATS and FSM show their alignment with the data management goals of the NIH, and with best practices agreed upon by an international community of research data professionals. This firm foundation will support and enable advancements in research data management. Curation, interoperability, and discoverability will follow.

The FAIR Principles

Data that is stored, curated, and shared according to the FAIR principles is:

Findable: described richly with metadata and have a unique identifier (often a URI)

Accessible: retrievable by their identifier through free and open Internet protocols that allow authentication and authorization where necessary

Interoperable: described with commonly used metadata standards and controlled vocabularies

Re-usable: have been assigned a license assuring their re-use and have a clear provenance