Deposit Options to Enhance the FAIR Principles

John E Marcotte, Sarah Rush, Kelly Ogden-Schuette
University of Michigan
IASSIST 2022

Research data repositories must adhere to the FAIR Principles to make data Findable, Accessible, Interoperable, and Reusable as a minimum standard rather than a goal. In addition to making data findable, repositories should also strive to make data discoverable. Discoverability goes beyond cataloging data with just a Direct Object Identifier (DOI) to make it findable. Discoverable data turn up in many kinds of searches via contexts such as topics, variables, and funders. The keys to data discoverability are thorough metadata and effective search engines. Metadata is also important in making data accessible and interoperable.

Research data, as delineated from other data, is specifically information for producing aggregate results such as summary statistics and regression coefficients. These aggregate results must meet disclosure protection thresholds for cell sizes for tables, sample sizes for regressions, and other specified conditions such as the suppression of certain variables or disallowed sub-samples. Although research data may contain information about individuals and organizations, they are not intended for identifying particular individuals or organizations.

Research data are sometimes available from multiple sources. One example is ongoing studies with multiple waves that often provide access to data through a project website. These projects frequently, however, want to make their data findable to researchers via sources other than just the project website.

Most repositories for research data have only one method of deposit and access. That is, the repository hosts the only accessible version of the research data and documentation including the codebook or data dictionary. These repositories make the research data available for download or through a restricted access mechanism. It is essential for research data repositories to adhere to the FAIR Principles:

> **F**indable: Data and metadata should be easy for both humans and computers to find
> **A**ccessible:Public access through web download or controlled access through an application
> **I**nteroperable: Both data and metadata should interoperate with workflows for analysis, storage, and processing

Reusable: Data and metadata should be so well-described that they can be replicated and re-analyzed

A key goal of any repository is to make research data *Discoverable*, which is an enhanced version of the Findable principle. A DOI makes data findable but not necessarily discoverable, so a DOI is not sufficient. Discoverability enables searches by title, source, author, and topic as well as by variable and question text. A discoverable system allows researchers to compare search results to ascertain differences and similarities.

Discoverability is increasingly important as data repositories are tasked with accommodating a growing variety of data types. This situation is a challenge as no single repository can manage all types of data; however, researchers may want to analyze these different data types for the same project. By sharing data and metadata, data repositories increase discoverability and facilitate the FAIR principals in support of these goals.

ICPSR is one of the largest repositories of research data. ICPSR has a catalog of research data as well as a variables database. Through ICPSR, researchers can search for data by topics or variables. ICPSR offers online analysis and exploration of the data as well as a bibliography of publications based on data in its holdings. The bibliography helps researchers build on the previous work of others.

With the goals of FAIR and Discoverability in mind, the Data Sharing for Demographic Research (DSDR) project at ICPSR has expanded its deposit options. DSDR is funded by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Research data available through DSDR focus primarily on maternal and child health, lifecourse, and supporting international comparisons.

DSDR has developed four types of deposits to enhance discoverability. While many repositories only accept research data, DSDR also ingests metadata. The deposit options offered by DSDR are Standard, Mirror, Extensive Metadata, and Study Descriptor:

(1) *Standard* is the most common type of deposit in which the data are deposited exclusively with one repository. With a DSDR standard deposit, for example, we are in the position to add the most value to the data by enhancing the data with online analysis, a bibliography of data-related literature, and data guides. DSDR only provides access to restricted data with this type of deposit because administering Data Use Agreements from multiple sources undermines protocols for the protection of human

subjects. Restricted data agreement language and data security plans may vary among repositories.

(2) *Mirror deposit* is the option used for making data available through multiple repositories. Different repositories have the same data and documentation with one repository designated as the primary while other repositories mirror the primary. Mirror deposits are public access only because of the problem of administering restricted Data Use Agreements from multiple sources. A Mirror deposit enhances the Findable, Accessible, and Interoperable attributes of data. Because the research data are available from multiple sources, they have a greater degree of findability and accessibility. Moreover, in order to mirror between repositories, the data and metadata must have a high degree of interoperability.

(3) *Extensive metadata* deposits are for situations where the data and documentation are available through another repository. Metadata about the study and variables can be incorporated into multiple repository catalogs. This type of deposit is ideal for ongoing studies with multiple waves where data are available from a project website (as mentioned earlier), but where the project wants to increase the findability of the data. This type of deposit enables DSDR to make data that it does not host discoverable through searches of topics, variables, and funders in the ICPSR catalog and variables database. Moreover, ICPSR presents search results in a manner that facilitates comparisons across studies and variables. An extensive metadata deposit enhances Findable and Interoperable attributes of the data.

(4) *Study descriptor* is the easiest and most general deposit. While the data and documentation are available from another source, this deposit includes a description of the study and the type of data. The study descriptor metadata can be incorporated into multiple repository catalogs. Researchers can search and compare studies, so this type of deposit enhances the Findable aspect of research data. Study Descriptor deposits enable DSDR to make studies with diverse types of data discoverable through the ICPSR catalog, such as genomics data and brain images hosted in other repositories.

The deposit options are dependent on the interoperability of metadata. Metadata standards vary by academic discipline and type of data. ICPSR is a leader in the Data Documentation Initiative (DDI) and currently uses DDI version 2 as its metadata schema. For metadata, repositories also employ other standards, including *CEFI, Dublin Core, DCAT, DataCite, DIF, FITS, and PROV*. Harmonizing metadata elements among these different schemas is an ongoing challenge for organizations and data sharing, in general.

Data archives may comprise both data repositories and catalogs or only catalogs. Research data catalogs, such as the one at ICPSR, are essential for making data discoverable since catalogs are the foundation for searches. Mirror, Extensive Metadata, and Study Descriptor deposit options enhance the discoverability of research data by expanding where data are cataloged beyond the hosting locations alone.

These four deposit options allow DSDR to promote the discoverability and accessibility of all types of data and studies as well as catalog more data than it hosts. A replete catalog is essential for enhancing the FAIR principles.