ROSA

Glossary for Data Sharing/Curation/Documentation

This glossary is meant to serve as a useful reference for terms that may not be familiar to all partners. It is an evergreen document and will be updated and added to as often as necessary.

Calibrated Data

Time series data that has been processed using software to apply instrument calibration (to compare acoustic measurements to a known standard). *From PAM Data Mgmt Best Practices.*

CARE Principles

The CARE Principles for Indigenous Data Governance were developed by the International Indigenous Data Sovereignty Interest Group, part of the Research Data Alliance.They were designed to complement the FAIR data principles and describe key aspects of working with Indigenous communities and cultures and their data, information, and knowledge. They only apply to Indigenous communities. The CARE principles are: Collective benefit, Authority to control, Responsibility, and Ethics. Published in Data Science Journal in 2020.

Controlled Vocabularies

Collections of terms used to improve the consistency and effectiveness of data entry, storage, retrieval, and interoperability. Their use removes ambiguity, makes implicit information explicit, and increases the interpretability of the data. See NISO Z39.19, and applications such as MEDIN.

Data Catalog

A persistent, findable, searchable virtual location that stores metadata, data or both, and facilitates searching and finding data across multiple data repositories or data storage locations.

Data Documentation

Noun: Metadata, READMEs, data dictionaries, and other materials that describe the data and the processes that created the data.
Verb: The process of creating the information that enables the sharing, use and reuse of data and software packages.

Data Governance

The legal structures and policies necessary to insure the desired management of data assets. This includes structures for how decisions are made about data and how people and processes are expected to behave in relation to the data.

Data Inventory

A data inventory is a fully described record of data assets. The inventory records basic information about a data asset including its name, contents, update frequency, use license, owner/maintainer, privacy considerations, data source, and other relevant details.
References: Bloomberg GovEx, Inventory.data.gov

Data License

A legal framework that specifies how data can be accessed, used, and reused.

Data Literacy

The knowledge that enables one to understand and communicate about data in an appropriate and meaningful way. This includes the principles, practices, and methods used throughout the data lifecycle.

Data Management

The tasks involved with documenting, maintaining, updating, and publishing data.

Data Products

Models, maps, predictions, analysis results, visualizations and other syntheses that are produced from raw and/or derived data.

Data Provenance

The detailed records that keep the history about the data from creation/origin, through transformations and modifications, to products. Data provenance is transparent, supports trust in the reliability of data, and allows reproducibility.

Data Repository

A persistent, findable, searchable entity that provides infrastructure for long-term storage and access to data. It should provide for data publication by data holders, as well as access for using/reusing data. Datasets in a repository are described with metadata that provides essential information about the data and enables efficient search and reuse. It should also employ unique identifiers, have redundancy, and have clear guidelines for what types of data it accepts. May or may not be federated with other repositories. See the *Encyclopedia of Big Data*.

Data Sharing Agreement

A legal agreement between two or more parties that defines the terms and conditions for sharing non-public data.

Data Standard

A technical specification that details the structure, organization, documentation, and format of data. Using data standards allows for the consistent collection of data, and aids in data aggregation, sharing and reuse, and interoperability of that data across different systems, sources, and users. Data standards can save time and effort if used during data collection, but may be applied at other points in the data life cycle by restructuring and re-documenting the data. References: data.gov, National Library of Medicine, and US Fish & Wildlife Service.

Data Stewardship

The actions involved in supporting, promoting, and guiding the use and reuse of data throughout the data lifecycle.

Data Type

A classification or category of various kinds of data, usually defined by common properties of the data, such as collection method (e.g. satellite telemetry data), scientific topic/question (e.g. fish length), or activity (e.g boulder relocation). Note that there are other definitions of this term not described here, specific to programming/coding and database management.

Database

An organized collection of data. Structured data are stored in relational databases, which are designed to make search and retrieval of data quicker and easier. Unstructured data can be stored in nonrelational databases, and are often found in web applications.

Derived Data

Existing data that have been processed, combined, assembled for the purposes of analysis, visualization, or publication. Derived data often refers to data resulting from the compilation or assembly of raw data into analysis-ready data to answer a particular question. Sometimes also referred to as processed data, though that often has a more limited meaning. Examples include text or data mining, and compiled datasets.

DOI

Digital Object Identifier. A unique identifier for a publication, including journal articles, datasets, software packages, etc.

ERDDAP

ERDDAP is a data server and broker that gives you a simple, consistent way to download subsets of scientific datasets in common file formats. It is used by organizations around the world to expose their data online. More information is available from NOAA.

Experimental Data

Data that are generated and/or collected under controlled conditions set up by experimental design or test method. They are often reproducible, but that can be expensive or time consuming. Examples include gene sequences, animal abundance and diveristy, plant physiological responses, chromatograms. See DMPTool definitions.

FAIR Principles

Findable, Accessible, Interoperable, Reusable. First published in Scientific Data in 2016.

Knowledge Graph

A digital structure that uses a graph model to formally represent entities (knowledge) and their properties, types, and relationships. Can include an ontology that allows humans and machines to understand and reason about its contents. *Ref: Stephen Young on Medium*

Metadata

Data about the data. The information that describes the data, both at the file-level (column headings, etc.) and the project level (data creator, geographic location, year, etc.), and enables someone to know enough about the data to be able to reuse it.
Federal Government Definition: Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource (NISO 2004, ISBN: 1-880124-62-9)

Observational Data

Data that are captured in real-time, typically in nature outside a lab. They are usually irreplaceable and therefore the most important to safeguard. Examples include sensor readings, telemetry, survey results, and images. See DMPTool definitions.

Ontology

A definition of terms and the relationships between them, often depicted as a flow chart. Frequently used to align term definitions in order to create a common vocabulary. Examples include the Scientific Variables Ontology and the Linked Earth Ontology. *Definition Ref: NIH*

OPeNDAP

Open-source Project for a Network Data Access Protocol is a server that enables sharing data over the internet. Serves data in ASCII, binary, or NetCDF formats. Useful information from NASA.

Open Data Science

The tools and practices enabling reproducible, transparent, and inclusive practices for data-intensive science. *Adapted from Ileana Fenwick's presentation*.

Raw Data

Data from a sensor, observation, or experiment that has not been cleaned or processed. This data may be uncalibrated, unpacked and compressed, and not QAQC'd.

Research Partnership

A data collaborative structure where a private-sector data holder engages directly with academic or government partners and shares private data in order to generate new analysis and knowledge. *Adapted from the GovLab.*

Reusable Data

Data that are formatted, documented, and shared in ways that make them understandable, interpretable, and obtainable by people other than the original data collector/holder. See also the FAIR definition.

Schema

An outline, diagram, or model of the structure of different types of data.
This can be within a relational database, where the schema describes the tables and fields contained within the database. Another type is the XML schema, which describes the elements in the XML file using a specific structure. This structure is frequently used for metadata.

Simulation Data

Data that are generated by machine from test models. These data are likely reproducible if the code, model parameters, and input data are preserved. Examples include climate or economic models. See DMPTool definitions.

THREDDS

The Thematic Real-time Environmental Distributed Data Services (THREDDS) server has features and interfaces that make it easier to explore and use data, both interactively and via automatic clients. THREDDS services include OPeNDAP and NetCDF Subset Service (NCSS). User guide from National Center for Environmental Information (NCEI) is here.

Trusted Intermediary

A data collaborative structure where a third-party actor supports collaboration between private-sector data holders and data users in the public sector, academia, NGO sector, and society. *Adapted from the GovLab.*