# GLOSSARY OF DATA MANAGEMENT TERMS

This glossary has been developed by The Centre for e-Research and Digital Innovation (CeRDI) at Federation University. It provides definitions for select terms that might be unfamiliar. This is not meant to be a comprehensive list of all data management terms. Rather, it is a list of ones that we have found, in our experience, to be most used.

This work is part of the Soil CRC's Improved Soil Data Management project led by CeRDI's Dr Nathan Robinson.

## TERMS

**anonymized data:** Data about individuals that does not reveal their identities nor links to other data that would reveal their identities. This term is not the same as "coded data" or "de-identified data."

**archive (verb):** The transfer of material to a facility that stores, retains, and provides access to that material on a long-term or permanent basis.

**attribution:** The act of referencing a dataset's original creator, usually signalled by copyright or licencing information accompanying the dataset.

**citation:** The practice of referencing research. A data citation includes key descriptive information about the data, such as the title, source, and responsible parties.

**code**: This applies to computer codes or technical scripts. In the context of data management, this may include code used in the collection, manipulation, processing, analysis, or visualization of data, but may also include software developed for other purposes.

**coded data**: Data accompanied by a linkage file that connects unique identifiers about an individual (such as their name, medical number, email address or telephone number) with a unique study ID code not associated with the individuals' personal information.

**confidentiality**: The right of privacy and security of disclosed personal information. This mainly applies to data collected on human subjects. Researchers are guided by strict legal requirements to prevent the release of private, personally identifiable information provided by research subjects.

**copyright**: A set of legal rights extended to copyright owners (the author, creator, or organisation to whom the rights have been assigned) that govern such activities as reproducing, distributing, adapting, or exhibiting original research.

**data:** Data is the information used to validate research findings.

**data governance**: a collection of processes, roles, policies, standards, and metrics that ensure the effective and efficient use of data

**data licensing:** a legal arrangement from the owner or repository of the data specifying what users can do with the data

**data life cycle:** Refers to the sequence of stages that data goes through. It encompasses data creation, data storage, data usage, data archiving and data destruction.

**data management plan:** A data management plan requirements may be defined differently by different funders, programs, or organisations. It is a plan that clearly outlines actions that will be taken throughout the data life cycle.

**data quality assurance process:** the process of identifying and eliminating anomalies by means of data profiling and data cleansing

**data stewardship:** the management and oversight of an organisation's data assets

**de-identified data:** Data for which all direct or indirect identifiers or codes linking the data to individual subjects' identities are destroyed.

**derivative (of data)**: Any data, publication, illustration or visualization, or other work that rearranges, presents, or otherwise makes use of an existing data set.

**digital object identifier (doi):** an electronic 'tag' created to uniquely identify objects or data

**domain specific data standards:** data standards that are designed and intended for use in a particular domain

**FAIR Principles:** Acronym for the main qualities of a dataset namely "Findable, Accessible, Interoperable, and Reusable."

**file format**: The technical structure used for encoding data in a computer file. File formats are usually identified by the file extension (e.g. .xlsx, .csv, .dbf).

**institutional repository**: A service hosted by an institution for storing and providing online access to digital content. The institution hosting this service is usually the institution that creates the content.

**intellectual property**: Legal rights applied to creative works, including (but not limited to) copyright, patents, trademarks, and trade secrets.

**license**: In the context of data management, a legal statement that expresses the terms of use of data

**machine-readable data**: Structured data that can be easily processed by a computer.

**metadata:** Documentation or information about a data set. It may be embedded in the data itself or exist separately from the data. Metadata may describe the ownership, purpose, methods, organisation, and conditions for use of data, technical information about the data, and other information.

**metadata standards:** a requirement which is intended to establish a common understanding of the meaning or semantics of the data

**open access**: Refers to online, freely available material that has few or no copyright or licensing restrictions.

**open data**: Data that can be freely accessed used, modified, and shared. Entails legal and technical permissions for users to access and modify the data.

**open-source software**: Software for which the source is available under an open license. Allows users to inspect the source code, run their own versions of the program, offer modification suggestions, fix bugs, develop new features, etc.

**persistent identifier**: A unique and long-lasting reference that allows for continued access to a digital object. Examples of persistent identifier systems include Digital Object Identifiers (DOIs), handles, and Archival Resources Keys (ARKs). Persistent identifiers support interoperability and the reliable citation of digital content.

**preservation (of data)**: Ensuring that data remain intact, accessible, and understandable over time. This requires preserving the integrity of digital files themselves.

**primary data:** Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments etc.

**privacy**: The protection of personal information from unauthorized access by others, usually governed by legislation.

**repository**: A facility that manages the appraisal, preservation, and accessibility to materials on a long-term or permanent basis.

**restricted data or restricted access data**: Data which are made available under stringent, secure conditions. Typically, confidential, or sensitive data.

**secondary data:** The data that have been collected for another purpose but has some relevance to your research needs and/or the data is collected by someone else instead of the researcher themselves.

**security**: Methods of protecting data from unauthorized access, modification, or destruction.

**standards:** Accepted methods or models of practice. In the context of data management, standards typically apply to data or file formats, and to metadata.

**tabular data:** Data that appear in a table format

**usage statement**: An expression of the conditions under which a data set may be used. May be formal, such as a license or contract, or informal based on the preferences of the data owner(s).