CODATA RDM Terminology (2023 version)

This document provides a human-readable overview of the updated RDM Terminology (RDMT) as revised by the CODATA RDM Terminology Working Group, shared for public review, and then confirmed and finalised in 2023. More information on the WG is available here: https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/.

As WG convenor, I'm currently working to make available the Terminology in a FAIR, machine-actionable way with our technical partners Research Vocabularies Australia. This first instantiation is a time-consuming process and so in the meantime, this human-readable guide to the updated Terminology is provided as an interim measure, with a single DOI (10.5281/zenodo.10626170) for the entire 2023 Terminology to aid referencing. Please note that the machine-actionable version of the Terminology will include referenceable URIs for each term, and notes on review decisions and sources. The simple overview provided here is limited to a list of the revised Terminology until the machine-actionable version is available, for the convenience of the Terminology's user community.

RDMT Scope:
"The goal of this terminology is to gather the key terms needed for a common understanding of the research data management domain. Research data management (RDM) refers to the storage, access and preservation of data created or collected in the course of research. Research data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used to long term preservation of data deliverables after the research investigation has concluded.
Definitions are intended to be clear and unambiguous, and where possible, fit with common usage. Definitions should be apposite across research data management activities of key stakeholders, including those working on research data management within the context of research, data management, digital curation and preservation, research management, research policy, open data advocacy, computer science, information management, research administration, library, scholarly publishing, digital archiving and research funding roles. Some terms may have more than one definition, in which case the relevant context should be specified.
If a consensus definition can be easily found elsewhere, the term is out of scope. This terminology is limited to the specific concepts necessary for a common understanding of RDM. This scope statement has been revised and approved by the 2021-22 RDM Terminology WG."

I will update the metadata of the Zenodo page hosting this document with the location of the full, machine-actionable version of the RDMT as soon as it is available. In the meantime, I hope this overview is useful.

Laura Molloy, CODATA
E: Laura AT codata.org

Term	Definition	Example
Access	Continued availability and ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for. Users who have access can retrieve, understand, manipulate, and store copies.	
Accessibility	Users' ability to gain access to or retrieve data once it has been discovered. This includes data instances where access to the data is limited, such as when user requests need to be authenticated and authorised.	
Access controls	Definitions of the access relationships between the following metadata: data object name, a user name (or user group, or user role), and access permission(s). The information can be stored as metadata information associated with each data object. The information can be generated dynamically by applying the access controls of the collection that organises the data objects (if a collection sticky bit is turned on).	
Access workflow	Type of access entity that contains the services and functions which make the data object holdings and their information content and related services visible to data consumers.	
Active data	Research data actively accessible and modifiable during the active phase of the research project.	
Administrative data	Information collected primarily for administrative, and not research purposes. It includes profiles and curriculum vitae of researchers, the scope and impact of research projects, funding, citations, and information about research outcomes. This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service. These data are also recognized as having research value.	
Administrative metadata	Metadata used to manage administrative aspects of the digital objects such as intellectual property rights and acquisition. Also documents information concerning the creation, alteration, and version control of the metadata itself. This is sometimes known as meta-metadata.	
Aggregated data	High-level data that are expressed in a summary form.	Summary statistics
Aggregation	Compilation of elements, often from different sources. Types of aggregation differ by the nature of the processes by which elements are brought together and the intention for aggregating. Aggregations differ in the nature of relations between the constituent parts.	
Analogue data	Data created and presented in the form of physical materials.	Written notes; sketches; maquettes; specimens
Analogue materials	Non-digital materials that have a physical presence.	Written and printed materials, maquettes, specimens

Anonymity	Ethical principle that is applied in research to maintain the privacy of research participants by keeping their identity unknown through irreversible processes.	
Archive		
Archive (noun)	Curated collection or repository containing physical or digital static records, objects, metadata and data deemed suitable for permanent retention, set up and managed to established standards and models, such as ISAD(G), CoreTrustSeal, and the OAIS reference model, that ensure long term integrity, security, authenticity and accessibility of the records, objects, metadata and data.	
Archive (verb)	Engage in curation activity that ensures that records, objects, metadata and data are properly selected, stored, and can be accessed, and for which logical and physical integrity are maintained over time, including security and authenticity.	
Archivist	Person responsible for appraising, acquiring, arranging, describing, preserving, and providing access to records of enduring value, according to the principles of provenance, original order, and collective control to protect the materials' authenticity and context. Such persons may also have responsibility for management and oversight of an archival repository or of records of enduring value. There is international variance in how this term is used; some archivists primarily interact with inactive records, while others would have responsibility for both inactive and active records.	
At-risk data	Data that are at risk of being lost. At-risk data include data that are not easily accessible, have been dispersed, have been separated from the research output object, are stored on a medium that is obsolete or at risk of deterioration, data that were not recorded in digital form, and digital data that are available but are not useable because they have been detached from supporting data, metadata, and information needed to use and interpret them intelligently.	
Audit	Evaluation of an organisation, system, group, project or product with respect to its data and processes around this, often in accordance with a standard, guide, or framework used to structure the work. This can involve assessing, describing, and classifying any data held. An audit can be carried out internally by those who have access to the data or participate in related processes regularly, or by an independent, external actor.	
Authenticity metadata	Type of metadata that conveys information needed to link a data object to its original source.	
Big data	Voluminous amount of structured, semi-structured and/or unstructured data that have the potential to be mined for information, primarily characterised by big volume, extensive variety, high velocity (creation and use), and/or variability that together require a scalable architecture for efficient data storage, manipulation, and analysis. The definition is evolving and can vary by sector, depending on what kind of software tools are commonly available and what sizes of datasets are common in a particular discipline. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes.	
Bit sequence	Representation of digital content in an assembly of the fundamental unit of digital bits.	
Bit stream	Unstructured sequence of bits that is identified as a unit (e.g., bits in a communication transmission). It may be stored as a unit or may exist as a pattern and be generated. A digital object may be represented as a bit stream of finite length that encodes its informational content.	

Born digital	Digital materials that were originally created in digital form, as distinct from those created as a result of digitising analogue originals.	
Boundary value	Data value that corresponds to a minimum or maximum input or output value specified for a system or component.	
Canonical data collection	Data collection that has been normalised by some established criteria to allow for effective data management. Examples include: data files that belong to a certain experiment, all files that are created by one specific simulation, all files that belong to a specific observation (same day, same place, etc.).	
CARE Principles for Indigenous Data Governance	Set of principles for Indigenous data governance. CARE stands for Collective benefit, Authority to control, Responsibility and Ethics. These principles complement the existing FAIR principles.	
Catalogue		
Catalogue (noun)	Index describing, indicating the location of, and recording other details of resources, materials works, etc. Curated and organised using a formal metadata schema such as MARC, ISAD(G), Dublin Core, DataCite etc.	
Catalogue (verb)	Describe digital or analogue data in accordance with a formal metadata standard to create a record of that dataset's characteristics, provenance and location. Create, add to or edit a catalogue.	
Change log	Document, spreadsheet, or digital tool that tracks the progress of each change in a dataset, code or other research object.	
Checksum	Alphanumeric signature (similar to a fingerprint) calculated from a digital object's content and structure using a mathematical algorithm. The algorithm will always produce the same checksum unless any change, no matter how small, is made to the file. Comparing checksums over time facilitates the management of integrity and authenticity of digital content.	
Citable data	Standalone dataset that can be cited in a similar manner to other research outputs. The dataset appears in a data repository, data paper or project website, and has a Persistent Identifier. Most current referencing systems provide a format for citing datasets.	
Cloud computing	Large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualised, dynamically-scalable, managed computing power, storage, platforms and services are delivered on demand to external customers over the Internet. Key features are that: it is a specialised distributed computing paradigm; it is massively scalable; it can be encapsulated as an abstract entity that delivers different levels of services to customers outside the Cloud; it is driven by economies of scale; and the services can be dynamically configured (via virtualisation or other approaches) and delivered on demand.	
Cloud ecosystem	Ecosystem that includes software, infrastructure, consultants, integrators, partners, third parties and anything else in the specified environment that has a bearing on the other components.	
Collection management identification	Type of data provenance indication that adds metadata to identify data collections. The organisation doing the collection management is stated in the metadata along with the provenance of collection management events such as source of data acquisition, conservation, and movement.	

Comma-separated values	Values in a table presented as a series of ASCII text lines organised so that each column value is separated by a comma from the next column's value and each row starts a new line. Along with pipe- and tab-separated values, comma-separated values are a specific example of a record-oriented data structure (usually of fixed dimension) with fields separated by an agreed delimiter.	
Confidentiality	Duty and practice of ensuring that personal or sensitive information only flows from one entity to another according to legislated or otherwise broadly accepted norms and policies. This can be done by either restricting access to the data or certain variables in the data, and/or by protecting personal or sensitive information using an anonymisation method.	
Conformance	Satisfaction of the requirements of a specific standard(s) and/or specification(s). Conformance is used with respect to voluntary standards and specifications, whereas compliance is used with respect to mandatory standards and regulations.	
Consumer data	Information trail customers leave behind as a result of their (digital) interactions with an organisation. This data, which sometimes comprises personal information, comes from such sources and channels as social media networks, marketing campaigns, customer service requests, call centre communications, online browsing data, mobile applications, purchasing history and preferences, and more.	
Container		
Container (digital archiving)	Logical collection of objects, using a standard such as Bag-IT for archival management purposes.	https://en.wikipedia.org/wiki/BagIt
Container (computing)	Computing technology which allows application code, data, dependencies, and configurations to be packaged into a single object that can be deployed in any environment.	Examples include Singularity (https://en.wikipedia.org/wiki/Singularity_(software)) and Docker (https://en.wikipedia.org/wiki/Docker_(software))
Content replication	Type of digital migration where there is no change to the Packaging Information, the Content Information, or the Preservation Description Information (PDI). The bits used to represent these Information Objects are preserved in the transfer to the same or new media instance.	
Controlled vocabulary	List of standardised terminology, words, or phrases, used for indexing or content analysis and information retrieval, usually in a defined information domain.	
Corpus	Set of documents that has a scientific meaning. A corpus can be produced by an individual researcher's activity (including its archival materials) or from a laboratory's research, a field campaign, a survey, or any other discrete research activity.	
Corrupt data	Deterioration of computer data as a result of some external agent such as viruses, hardware or software incompatibility, flaws, or failures, power outages, dust, water, extreme temperatures, etc.	
Cross-disciplinary	Research approach that explains aspects of one discipline in terms of another (e.g., the physics of music; the politics of literature)	
Curation	Managing and promoting the use of assets from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse. For dynamic datasets this may mean	

	continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials.	
Curation workflow	Type of workflow that includes active steps to curate data as an aid to on-going management of data through their lifecycle.	
Dark data	Operational data that are not being used, such as information assets that organisations collect, process and store in the course of their regular business activity, but generally fail to use for other purposes. Such data are seen as an economic opportunity for companies if they can take advantage of it to drive new revenues or reduce internal costs. Examples include server log files that can give clues to website visitor behaviour; client call detail records that can indicate consumer sentiment; and mobile geolocation data that can reveal traffic patterns to aid in business planning.	
Data	Facts, measurements, recordings, records, or observations about the world, collected by researchers, that are yet to be processed/interpreted/analysed. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records.	
Data access protocol	System that allows users to be granted access to a database under specified conditions.	
Data access statement	Statement accompanying a research publication that describes whether supporting data is available to access, where this has been made available and under what conditions.	
Data acquisition	Process of acquiring data from some source. For example, data may be acquired by download from a repository, transfer from a data logger, data capture, etc.	
Data analysis	Techniques that produce synthesised knowledge from organised information. Process of inspecting, cleaning, transforming, and modelling data with the goal of highlighting useful information suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.	
Data archive	Service providing long-term, permanent care and accessibility for digital objects with research value. The standard for such repositories is the Open Archival Information System reference model.	
Data availability	State when data are available in a timely manner in the place and form as needed by the user.	
Data capture	Process or means of obtaining and storing external data, particularly images or sounds, for use at a later time. In biometric security systems, for example, capture is the acquisition of, or the process of acquiring an identifying characteristic such as a finger image, palm image, facial image, iris print, or voice print. In order to capture the data, a transducer is employed that converts the actual image or sound into a digital file. The file is then stored. At a later time, it can be analysed by a computer, or compared with other files in a database to verify identity or to provide authorization to enter a secured system. Screen capture is the acquisition and storage of an image on a monitor or display exactly as it appears at a specific time. This can sometimes (but not always) be done by hitting the “print screen” key, in which case the image appears as a bitmap file in the clipboard. It can also be done by photographing the display screen with a digital camera external to the computer.	

	Electronic signals from scientific instruments, dataloggers, sensors, etc., can also be captured, converted to data, and stored for use at a later time.	
Data catalogue	Curated collection of metadata records describing datasets and their data elements. Curated and organised using a formal metadata schema appropriate to data and data sets (e.g. ReCollect etc).	
Data centre		
Data centre (computing)	Facility providing IT services, such as servers, massive storage, and network connectivity.	
Data centre (research data)	Facility holding large scale data repositories.	
Data citation	Process of citing a dataset in a similar manner to other research outputs. The dataset must be a standalone output that appears in a data repository, data paper or project website, and has a Persistent Identifier. Most current referencing systems provide a format for citing datasets.	
Data cleaning	Process of detecting and correcting corrupt or inaccurate records from a dataset. Data cleaning is a continuous process that requires corrective actions throughout the data lifecycle. Data cleaning involves identifying, replacing, modifying or deleting incomplete, incorrect, inaccurate, inconsistent, irrelevant, and improperly formatted data. Typically, the process involves updating, correcting, standardising, and de-duplicating records to create a single view of the data, even if they are stored in multiple disparate systems.	
Data collection		
Data collection (grouping)	Logical grouping of (research) datasets that share a common aspect or concept. Highest entity in the hierarchy of data groupings (data collection, dataset, row or record in a dataset). Comprises a grouping of datasets that have a strong connection and it is organised coherently around a single element or concept such as a model or instrument.	
Data collection (process)	Act or process of creating, recording, acquiring, or linking to data or type of data. Includes the specification of file formats, naming conventions, data structure, and what provisions have been made for version control, data re-use, sharing, and long-term access to the data.	
Data completeness	Degree to which all required measurements are known. Values may be designated as “missing” in order not to have empty cells, or missing values may be replaced with default or interpolated values. In the case of default or interpolated values, these must be flagged as such to distinguish them from actual measurements or observations. Missing, default, or interpolated values do not imply that the dataset has been made complete.	
Data compliance	Ongoing processes to ensure adherence of data to both enterprise business rules (government department, university, industry, or agency), and to legal, regulatory and accreditation requirements. Includes five areas: controls, audit, legal compliance, regulatory compliance, and accreditation conformance.	
Data container	Software stack that chunks digital objects at a physical layer. Typical containers are file systems, database management systems, content management systems, clouds etc. The software stack implies some form of encapsulation of the digital object.	

Data curation	Managed process throughout the data lifecycle, by which data/data collections are cleansed, documented, standardised, formatted and inter-related. This includes versioning data, or forming a new collection from several data sources, annotating with metadata, adding codes to raw data (e.g., classifying a galaxy image with a galaxy type such as “spiral”). Higher levels of curation involve maintaining links with annotation and with other published materials. Thus a dataset may include a citation link to publication whose analysis was based on the data. The goal of curation is to manage and promote the use of data from its point of creation to ensure it is fit for contemporary purpose and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Special forms of curation may be available in data repositories. The data curation process itself must be documented as part of curation. Thus curation and provenance are highly related.	
Data custodian	Individual or organisation responsible for the IT infrastructure providing and protecting data in conformance with the policies and practices prescribed by data governance.	
Data de-noising	Removing noise from data.	
Data deletion	Process of destroying data stored on tapes, hard disks and other forms of electronic media so that it is no longer there and it cannot be restored.	
Data destruction	Process of destroying data stored on tapes, hard disks and other forms of electronic media so that it is completely unreadable and cannot be accessed or used.	

Data dictionary	Collection of descriptions of the data objects or items in a data model. After each data object or item is given a descriptive name, its relationship is described (or it becomes part of some structure that implicitly describes relationship), the type of data (such as text or image or binary value) is described, possible predefined values are listed, and a brief textual description is provided. This collection can be organised for reference into a data dictionary.	
Data dredging	Data mining practice in which large volumes of data are analysed seeking any possible relationships between data. The traditional scientific method, in contrast, begins with a hypothesis and follows with an examination of the data. Data dredging often circumvents traditional data mining techniques and may lead to premature conclusions. Uncovered patterns may be presented as statistically significant without any specific hypothesis as to the underlying causality.	
Data driven decision management	Approach to governance that values decisions that can be backed up with data that can be verified. The success of the data-driven approach is reliant upon the quality of the data gathered and the effectiveness of its analysis and interpretation. Errors can creep into data analytics processes at any stage of the endeavour and serious issues can result when they do.	
Data driven disaster	Serious problem caused by one or more ineffective data analysis processes.	
Data element	Unit of data for which the definition, identification, representation (term used to represent it), and permissible values are specified by means of a set of attributes.	For example, the data element “age of a person” with values consisting of all combinations of 3 decimal digits; A personnel record may include the data elements “name” and “address”. In the context of the personnel record, “name” and “address” function as an indivisible unit, e.g., the data element “name” and the data

		element “address” each can be stored and retrieved as an indivisible unit. However, in a different context, “address” itself may be considered a record that contains its own data elements “street address”, “city”, “postal code”, “country”.
Data entity	Object, event or phenomenon about which data are stored in a database and which has intermediate representation in a Data Model.	
Data ethics	Branch of ethics concerned with the moral implications of practices involving data, including data collection, description, storage, accessibility, rights, ownership, and uses. This also includes any corresponding practices, such as the use of data in algorithms, artificial intelligence, or in policy development, which have the potential to positively or negatively impact individuals, groups, or society.	
Data exploration	Summarising the main characteristics of a dataset using visualisation.	
Data file format	Layout of a file in terms of how the data within the file are organised and encoded for storage.	
Data governance	Exercise of authority, control and shared decision making (planning, monitoring and enforcement) over the management of data assets. Overall management of the availability, usability, integrity, and security of the data employed in an organisation. A sound data governance program includes a governing body or council, a defined set of procedures, and a plan to execute those procedures.	
Data harmonisation	Making data from different sources comparable. The processes involved in producing inferentially equivalent data. The data are interchangeable between different information systems with uniform and unambiguous mutual understanding.	
Data hygiene	Collective processes conducted to ensure the cleanliness of data. Data are considered clean when they are relatively error-free.	
Data ingestion	Process of obtaining, importing, and processing data for later use or storage in a database. This process often involves altering individual files by editing their content and/or formatting them to fit into a larger document. An effective data ingestion methodology begins by validating the individual files, then prioritises the sources for optimum processing, and finally validates the results. When numerous data sources exist in diverse formats (the sources may number in the hundreds and the formats in the dozens), maintaining reasonable speed and efficiency can become a major challenge. To that end, several vendors offer programs tailored to the task of data ingestion in specific applications or environments.	
Data integration	Combining diverse datasets from disparate sources into one unified dataset or database. Data are accessed and extracted, moved, validated, cleaned, transformed and loaded.	
Data integrity		
Data integrity (access)	Assurance that information can only be accessed or modified by those authorised to do so.	

Data integrity (fit for purpose)	Assurance the data are clean, traceable, and fit for purpose.	
Data item	Type of data element that expresses a proposition that binds one or more property values to some data entity.	
Data lake	Collection of generally unrefined or pre-processed (or raw) data captured and consolidated from multiple sources. It contains data with limited structure where data models are applied a posteriori, that is, after loading data into the lake. That is also known as data-first or schema on-read. Data in the lake might not be harmonised or integrated, and curation and quality assurance are not required.	
Data landscape	The broad communities of research data management and related areas that influence researcher incentives and behaviours concerning data.	
Data librarian	Librarian who manages the sharing and publishing of datasets as openly as possible and as closed as necessary, and the management and curation of repositories required to achieve this. Broad role requirements include support for sharing and publishing datasets, finding, accessing, interoperating and re-using these datasets, reviewing and supporting Data Management Plans and training delivery.	
Data linkage	Bringing together, from two or more different sources, data that relate to the same individual, family, place or event.	
Data management infrastructure	Infrastructure used to provide data management and enforce data management policies. A data management infrastructure would include resources such as a data repository and an information catalogue.	
Data management plan	Statement describing how research data will be managed throughout a specified research project's life cycle - during and after the active phase of the research project - including terms regarding archiving and potential preservation of the data in a data repository. The DMP is considered to be a 'living' document, i.e., one which can be updated when necessary.	
Data management policy	Statement of an organisation’s processes for the management of a specified set of data assets.	
Data mart	Repository of data designed to serve a particular community of knowledge workers. A data mart contains harmonised, highly structured, quality data integrated from multiple sources (which is also a characteristic of the data warehouse). It's also optimised to support well-known, predefined and repeatable analytic queries, rather than ad-hoc analysis.	
Data migration	Transfer of data between storage types, formats, information technologies, or computer systems. A data migration project is usually undertaken to replace or upgrade servers or storage equipment, for a website consolidation, to conduct server maintenance or to relocate a data centre.	
Data mining	Analysing multivariate datasets using pattern recognition or other knowledge discovery techniques to identify potentially unknown and potentially meaningful data content, relationships, classification or trends. Data mining parameters include: Association (looking for patterns where one event is connected to another event); Sequence or path analysis (looking for patterns where one event leads to another later event); Classification (looking for new patterns); Clustering (finding and	

	visually documenting groups of facts not previously known); Forecasting, or predictive analytics (discovering patterns in data that can lead to reasonable predictions about the future.	
Data model	Model that specifies the structure or schema of a dataset. The model provides a documented description of the data and thus is an instance of metadata. It is a logical, relational data model showing an organised dataset as a collection of tables with entity, attributes and relations.	
Data modelling	Formalising and documenting existing processes and events. A first step in analysing a system of objects with which users interact is to identify each object and its relationship to other objects. This process is called data modelling and results in a picture of object relationships. Data modellers often use multiple models to view the same data and ensure that all processes, entities, relationships and data flows have been identified. There are several different approaches to data modelling, including: Conceptual Data Modelling (identifies the highest-level relationships between different entities); Enterprise Data Modelling (similar to conceptual data modelling, but addresses the unique requirements of a specific organisation); Logical Data Modelling (illustrates the specific entities, attributes and relationships involved in a business function. Serves as the basis for the creation of the physical data model); Physical Data Modelling (represents an application and database-specific implementation of a logical data model).	
Data organisation	Set of measures that are used by a repository to form aggregations of data objects (including collections and metadata) to describe the properties of data objects, to register PIDs, to build the PID records, to link between all components, and to set up the containers (in the form of the software stack) that are used to store all components.	
Data paper	Peer-reviewed academic publication focussed on the description of a dataset.	
Data perturbation	Method of data anonymisation by adding 'noise' to the data so that single living individuals cannot have their identities disclosed.	
Data policy	Set of high-level principles that establish a guiding framework for data-related issues. A data policy may contain subsidiary sections that define approaches to strategic aspects such as data access; data protection; data management, custodianship or stewardship; data preservation; relevant legal matters; data acquisition; data ethics and other data-related issues. Distinct from a data management policy which is the subset of data policy addressing the management of a specified set of data assets.	
Data preprocessing	Any type of processing performed on raw data to prepare it for another processing procedure. Preprocessing may include: data sampling, data transformation, de-noising, data normalisation, data standardisation, or feature extraction.	
Data processing	Generic concept referring to all kinds of procedures being executed on data at any point in the data lifecycle.	
Data product specification	Detailed description of a dataset or dataset series together with additional information that will enable it to be created, supplied to and used by another party. A data product specification provides a description of the universe of discourse and a specification for mapping the universe of discourse to a dataset. It may be used for production, sales, end-use or other purposes.	
Data production	All activities involved in the planning, collecting, processing, analysis and maintenance of data in the original research project. Among these activities are selecting a study design, constructing	

	instruments for data collection, conducting data collection/creation, performing data editing/verification/validation, analysing data, backing up data versions and preparing and tagging metadata.	
Data profiling	Statistical analysis and assessment of the quality of data values within a dataset for consistency, uniqueness and logic. The data profiling process cannot identify inaccurate data; it can only identify rule violations and anomalies. The insight gained by data profiling can be used to determine how difficult it will be to use existing data for other purposes. It can also be used to provide metrics to assess data quality and help determine whether or not metadata accurately describes the source data. Profiling tools evaluate the actual content, structure and quality of the data by exploring relationships that exist between value collections both within and across datasets.	For example, by examining the frequency distribution of different values for each column in a table, an analyst can gain insight into the type and use of each column. Cross-column analysis can be used to expose embedded value dependencies and inter-table analysis allows the analyst to discover overlapping value sets that represent foreign key relationships between entities.
Data publication	Release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publication occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are findable, accessible, interoperable and re-usable.	
Data quality	Reliability and application efficiency of data. Perception or assessment of a dataset's fitness to serve its purpose in a given context. Aspects of data quality include: Accuracy, Completeness, Update status, Relevance, Consistency across data sources, Reliability, Appropriate presentation, Accessibility. Data quality is affected by the way data are entered, stored and managed. Maintaining data quality requires going through the data periodically and scrubbing it. Typically this involves updating, standardising, and de-duplicating records to create a single view of the data, even if it is stored in multiple disparate systems.	
Data quality assurance	Process of verifying the reliability and effectiveness of data.	
Data recovery	Process of restoring data that have been lost, accidentally deleted, corrupted or made inaccessible for any reason. An organisation's disaster recovery plan should make known who in the organisation is responsible for recovering data, provide a strategy for how data will be recovered and document acceptable recovery point and recovery time objectives.	
Data reduction	Process of reducing the amount or size of stored data. This may be achieved by eliminating redundant copies of data files, deduplicating data files by removing redundant records, or by compressing the data files.	

Data registration	Curation process on a data object by which it receives a persistent identifier (PID) from a trusted registration authority. Registration must be accompanied by the step(s) to create and submit metadata describing the object to the registry.	
Data repository management	Management of a national, discipline or institutional repository of published datasets. Provision of infrastructure, curation, policy and training that govern the organisation, control, and properties of the repository such as: required file formats, access control restrictions, integrity, replication, retention, disposal, etc.	
Data representation	Object describing the context of the data, including provenance, description, structural, and administrative information.	
Data rescue	Recovery and/or transformation and digitization of dark data and at-risk data so that they can be preserved, accessed, shared, and used. Data rescue also involves the addition of rich metadata to make the content understandable and more easily re-usable.	
Data residency	Physical or geographic location of an organisation's data or information. Data residency also refers to the legal or regulatory requirements imposed on data based on the country or region in which it resides. Cloud computing, which allows organisations to deliver hosted services over the Internet, can create data residency concerns.	
Data retention policy	Established protocol of an organisation for retaining information for operational or regulatory compliance needs. The objectives of a data retention policy are to keep important information for future use or reference, to organise information so it can be searched and accessed at a later date, and to dispose of information that is no longer needed. A data retention policy must consider both the value of data over time, and regulations to which the data may be subject.	
Data review	Activity through which the correctness conditions of the data are verified. It also includes the specification of the type of the error or condition not met, and the qualification of the data and its division into "error-free" or "erroneous" data. Data review consists of both error detection and data analysis, and can be carried out in manual or automated mode.	
Data sampling	Selection of a statistically representative subset from a large population of data.	
Data scaling	Techniques used to deal with parameters having different units and scales.	
Data selection	Process that creates a new dataset from an original source.	Examples include: creating a subset of the data, querying a database.
Data sharing	Practice of making data available for checking, reproducing or reusing. The mechanisms available for achieving this are: making data available on request, as supplementary information to journal articles, or as published datasets in data repositories.	
Data splitting	Protecting sensitive data from unauthorised access by encrypting the data and storing different portions of a file on different servers. An unauthorised person would need to know the locations of the servers containing the parts, be able to get access to each server, know what data to combine, and how to decrypt it. Data splitting can be made even more effective by periodically retrieving and recombining the parts, and then splitting the data in a different way among different servers, and using a different encryption key.	

Data standard	Technical specification that defines how data should be structured and formatted to ease interoperability across different systems, publishers and users.	
Data standardisation	Conversion of multiple datasets to a single common format and structure.	
Data stewardship	Course of action taken by a person or group to manage and supervise organisational data assets with responsibility and commitment. Good stewardship involves adequate care, making use of the FAIR principles, and holding ownership and regulation to provide high-quality data (including metadata), combining trust and ethical practice.	
Data store	Repository for persistently storing collections of data, such as a database, a file system or a directory. The data stored can be of any type that can be rendered in digital format and placed in electronic media.	
Data stream	Sequence of digitally encoded, coherent signals used to send or receive a representation of information content as transmitted.	
Data structure	Specialised format for organising and storing data. General data structure types include the array, the file, the record, the table, the tree, and so on. Any data structure is designed to organise data to suit a specific purpose so that it can be accessed and worked with in appropriate ways. In computer programming, a data structure may be selected or designed to store data for the purpose of working on it with various algorithms.	
Data structure continuum	Continuum of data structure that includes unstructured data, semi-structured data, and structured data.	
Data table attribute		
Data table attribute (field)	Field or column in a database table. It is an abbreviation for 'physical data attribute' which is a single data element related to a data object, such as a table in a database. The database schema associates one or more attributes with each database entity (i.e. table).	
Data table attribute (logical)	Logical or conceptual attribute such as in an entity-attribute-relationship (EAR) data model.	
Data traceability	Data traceability follows the lifecycle of data to track all access and changes to the data. It helps demonstrate transparency, compliance and adherence to regulations. Data traceability, along with data compliance, can be considered part of a data audit process. Data traceability is fundamental to reproducible research.	
Data transformation	Manipulation of raw data to produce a single output.	
Data type registry	Registry for data types supporting their standardisation, uniqueness and discoverability. Data types range from complex digital objects to simple categories that occur in digital objects. An additional functionality may be to link data types to executable data processing functions.	Examples include: complex file types in biology (diagnosis), registering categories that appear in PID records to describe data properties.

Data upload database	Collection of interrelated data often with controlled redundancy, organised according to a scheme to serve one or more applications; the data are stored so that they can be used by several programs without concern for data structures or organisation.	
Data validation	Checks that data are valid, sensible, reasonable, clean, usable, and secure before they are processed. Provides well-defined guarantees for fitness, accuracy, and consistency for any of various kinds of user input into an application or automated system. Failures or omissions in data validation can lead to data corruption, security vulnerability. Improperly validated data can cause computer code processing the data to crash, generate error messages, behave in an unanticipated manner, or generate incorrect results that may be difficult or impossible to detect.	
Data warehouse	Central repository for all or significant parts of the data that an organisation's various business systems collect, containing harmonised, highly structured, quality data integrated from multiple sources. Data warehousing emphasises the capture of data from diverse sources for useful analysis and access but does not generally start from the point-of-view of the end user who may need access to specialised data marts. There are two approaches to data warehousing: The top-down approach spins off data marts for specific groups of users after the complete data warehouse has been created. The bottom-up approach builds the data marts first and then combines them into a single, all-encompassing data warehouse.	
Data wrangling	Manually or semi-automatically converting or mapping data from one form into another format that allows for more convenient consumption of the data with the help of semi-automated tools. Gathering and organising disparate data from different sources, often collected by many different investigators. Activities include developing and supporting search tools that utilise standardised metadata, harmonising the coding of data for specific variables, engineering new methods of combining data. with the help of semi-automated tools. The result of data wrangling is repurposed data.	
Data z-score scaling	Standardising data so that the mean is centred at zero. Therefore, participants' scores reflect their distance from the mean in standard deviations (i.e., whether their score is higher or lower than the mean).	
Database	Collection of data that is organised according to a conceptual structure/model describing the characteristics of these data and the relationships among their corresponding entities, supporting one or more application areas. A database allows its contents to be easily accessed, managed and updated. The type of database used depends on the requirements of the study. A common type is the relational database, where data are related to each other in a systematic manner so that they can be reorganised and accessed in a number of different ways. A database may house one or many datasets.	
Database administration	Managing the physical aspects of data resources, including database design and integrity, backup and recovery, performance and tuning.	
Dataset	Organised collection of data or objects in a computational format, that are generated or collected by researchers in the course of their investigations, regardless of their form or method, that form the object on which researchers test a hypothesis. This includes the full range of data: raw,	

	unprocessed datasets, proprietary generated and processed data and secondary data obtained from third parties. The presentation of the data in the application is enabled through metadata.	
Dataset series	Collection of datasets sharing the same product specification. A dataset series is a type of aggregation or collection with some “logical grouping” such as by a topic (specification) with the (product) unit being a dataset series.	Example: A series of earth observations. Each year, month or week (depending on the volume) might be a dataset and the series could run from a specified year to the present.
Datetime	Standard way to express a numeric calendar date that eliminates ambiguity, acceptable formats being defined by ISO 8601. ISO 8601 is applicable whenever representation of dates in the Gregorian calendar, times in the 24-hour timekeeping system, time intervals and recurring time intervals or of the formats of these representations are included in information interchange. It includes calendar dates expressed in terms of calendar year, calendar month and calendar day of the month; ordinal dates expressed in terms of calendar year and calendar day of the year; week dates expressed in terms of calendar year, calendar week number and calendar day of the week; local time based upon the 24-hour timekeeping system; Coordinated Universal Time of day; local time and the difference from Coordinated Universal Time; combination of date and time of day; time intervals; recurring time intervals.	
De-anonymisation	Reverse engineering process in which de-identified data are cross-referenced with other data sources to re-identify the personally identifiable information. This could occur if a de-identification process had not been not successfully performed, or had not been undertaken in the first place.	
De-identification	Techniques designed to make the risk of identifying a particular individual in a dataset negligible, whilst retaining the re-usability of the dataset. The purpose is to protect the privacy of the individual and comply with legislation, whilst enabling data sharing. Methods include removing direct and indirect identifiers such as names, addresses, social insurance numbers, or dates of birth, or using obfuscation methods such as encryption, hashing, generalisation, pseudonymisation, and perturbation.	
Denormalisation	In a relational database, an approach to speeding up read performance (data retrieval) in which the administrator selectively adds back specific instances of redundant data after the data structure has been normalised. After data has been duplicated, the database designer must take into account how multiple instances of the data will be maintained. One way to denormalise a database is to allow the database management system (DBMS) to store redundant information on disk. This has the added benefit of ensuring the consistency of redundant copies. Another approach is to denormalise the actual logical data design, but this can quickly lead to inconsistent data. Rules called constraints can be used to specify how redundant copies of information are synchronised, but they increase the complexity of the database design and also run the risk of impacting write performance.	
Derived data product	Result of applying a procedure to transform a data object in order to obtain a desired data product that is stored in a repository along with the provenance and descriptive metadata.	

Descriptive metadata	Metadata that describe a dataset or resource in such a way that people can discover and identify it. Contains information that aids with findability such as information (metadata elements) on the creator(s), affiliation(s), title, abstract, keywords, persistent identifier, related publications, etc.	
Digital archiving		
Digital archiving (libraries and archives)	Often used synonymously with 'digital preservation' in library and archiving professional communities.	
Digital archiving (computing)	In the context of computing, the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation.	
Digital data	Data in the form of digital materials.	
Digital infrastructure	Those layers that sit between base technology (a computer science concern) and discipline-specific science. Value-added systems and services that can be widely shared across scientific domains, both supporting and enabling large increases in multi- and interdisciplinary science while reducing duplication of effort and resources (including hardware, software, personnel, services and organisations).	
Digital materials		
Digital materials (surrogates)	Digital surrogates created as a result of converting analogue materials to digital form (digitisation).	
Digital materials (born digital)	Born-digital assets for which there has never been and is never intended to be an analogue equivalent.	
Digital materials (records)	Digital records.	
Digital object	Machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure helps to enable interoperability among diverse information systems. A digital object is composed of a structured sequence of bits/bytes. The bit sequence realising the object can be identified and accessed by a unique and persistent identifier or by use of referencing attributes describing its properties.	
Digital Object Identifier	Type of digital Persistent Identifier (PID) issued by the International DOI Foundation. This permanent digital identifier is associated with an object that permits the object to be referenced reliably even if its location and metadata undergo change over time.	
Digital preservation	Series of managed activities necessary to ensure continued access to digital materials for as long as necessary. All of the actions required to maintain access to digital materials beyond the limits of media failure or technological change. Those materials may be records created during the day-to-day business of an organisation; born-digital materials created for a specific purpose (such as teaching resources); or the products of digitisation projects. This definition specifically excludes the potential use of digital technology to preserve the original artefacts through digitisation.	

Digital research data	Research data in digital form. It may have been originally created in digital form, or it may have been converted from paper, or other non-digital form to a digital representation.	
Digital scholarship	Scholarship which is dependent upon digital methods, tools or resources. May include building a digital collection of information for further study and analysis; creating appropriate tools for collection-building; creating appropriate tools for the analysis and study of collections; using digital collections and analytical tools to generate new intellectual products; or creating authoring tools for these new intellectual products, either in traditional forms or in digital form.	
Digitisation	Process of creating digital files by scanning or otherwise converting analogue materials. The resulting digital copy, or digital surrogate, would then be classed as digital material and then subject to the same broad challenges involved in preserving access to it, as born-digital materials.	
Direct identifier	Data variable or value that directly discloses the identity of a single living individual, for example, a name, passport number or fingerprint.	
Dirty data	Data that contain errors. Can be caused by a number of factors including: inaccurate, incomplete or erroneous data such as spelling or punctuation errors, incorrect data or incorrect data type associated with a field, incomplete or outdated data, duplicate data, inconsistent data, incorrectly ordered data, improper parsing of fields from disparate systems, etc. Using a dirty dataset can lead to spurious associations, false conclusions and misdirected investments.	
Document type definition	Definition of the structure and the legal elements and attributes of an XML document.	
Documented data	Data that are delivered with all associated metadata, data dictionary, description of methods and instruments used to collect and process the data, and other supporting data (such as duplicate sample results, replicate analyses, percent recovery, etc.) with the purpose of providing the full context in which the data were created.	
Dublin Core	Widely used metadata element set, formally titled ISO 15836-1:2017, Information and documentation — The Dublin Core metadata element set — Part 1: Core elements.	
Dynamic data	Data that are changing frequently and at asynchronous moments.	Examples include: Data streams that are generated by sensors when it is unpredictable when data segments will appear in time (i.e. data streams have gaps); Data streams that are generated by humans in crowdsourcing scenarios where it is not clear when which cell in a database will be filled.
E-Research	(Historical) Computationally intensive, large-scale, networked and collaborative forms of research and scholarship across all disciplines, including all of the natural and physical sciences, related applied and technological disciplines, biomedicine, social science and the digital humanities.	

E-Research infrastructure	(Historical) Digital information-processing and/or computational technologies supported research to a significant degree. E-Science often involves intensive use of cutting-edge technologies that are advanced in technique, collaborative or on a large scale (over various possible measures: volumes of information, computational intensity, extent of distribution, variety of information types handled). The term 'Digital infrastructure' may be used in its place.	
E-Science	(Historical) Science supported to a significant degree by digital information-processing and/or computational technologies. E-Science often involves intensive use of cutting-edge technologies that are advanced in technique, collaborative or on a large scale, over various possible measures: volumes of information, computational intensity, extent of distribution, variety of information types handled.	
Ecosystem	Technical infrastructure available within a researcher's workflow, consisting of interoperable systems that impact how research data is handled and by whom.	
Electronic health record	Compilation of core electronic health data submitted by various healthcare providers and organisations, accessible by numerous authorised parties from a number of points of care, possibly even from different jurisdictions. Electronic health records typically include: contact information, information about visits to health care professionals, allergies, insurance information, family history, immunisation status, information about any conditions or diseases, a list of medications, records of hospitalisation, information about any surgeries or procedures performed.	
Electronic medical record	Electronic version of the paper record that doctors have traditionally maintained for their patients and which is typically only accessible within the facility or office that controls it.	
Encoding schema	Machine processable specifications which define the structure and syntax of metadata specifications in a formal schema language.	
Engineering and scientific support	Technical service involved in the performance, inspection and leadership of skilled technical activities.	
EXtensible Markup Language	Text format derived from Standard Generalized Markup Language or 'SGML' (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.	
Extensible resource identifier	Scheme used for identification of resources (including people and organisations) and the sharing of data across domains, enterprises, and applications. XRI TC will define a Uniform Resource Identifier (URI) scheme and a corresponding Uniform Resource Namespace (URN).	
Extract-Transform-Load	Process that involves the following steps: (a) Extract data from homogeneous or heterogeneous data sources which are often managed by different people. An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources have the correct/expected values; (b) Transform the data for storing it in consistent format or structure for querying and analysis purposes; An important function of transformation is the cleaning of data; and, (c) Load the data into the final target (database, operational data store, data mart, or data warehouse). ETL processes can involve considerable complexity, and significant operational problems can occur.	
FAIR data	Data that is managed in such a way that it is findable, accessible, interoperable, and reusable.	

FAIR data principles	Set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable.	
FAIR Guiding Principles for scientific data management and stewardship	Set of foundational principles [...] that all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people.	
Fair use	Legal concept in some jurisdictions that allows the reproduction of copyrighted material for certain purposes without obtaining specific permission and without paying a fee or royalty. Purposes permitting the application of fair use generally include review, news reporting, teaching, or scholarly research. When in doubt of whether a 'fair use' condition applies to a resource, the quickest and simplest thing may be to request permission from the copyright owner.	
Feature extraction	Selection of specific data that are significant in some particular context.	
Field	Data table column name.	
Findable	Data and metadata that can be located by humans and machines. They should be assigned a globally unique persistent identifier, be described with rich metadata, and ideally registered in a searchable catalogue or index.	
Fixed data	Data that are not, under normal circumstances, subject to change. Examples of fixed data include results from concluded research, medical records, and historical data.	
Foundational interoperability	Set of conditions that allow data exchange from one information technology system to be received by another and does not require the ability for the receiving information technology system to interpret the data.	
Golden record	Single, well-defined version of all the data entities in an organisational ecosystem. Encompasses all the data in every system of record within a particular organisation. A well-maintained, current golden record should be a fundamental element of the master data management policy for every enterprise. The word “golden” is sometimes used in information technology to express the importance of some type of source. In the context of virtualization, for example, a golden image is a template for a virtual machine, virtual desktop, servers, or hard disk drive.	
Hashing	Transformation of a string of characters into a usually shorter fixed-length value or key that represents the original string. Hashing is used to index and retrieve items in a database because it is faster to find the item using the shorter hashed key than to find it using the original value.	
Heat map	Two-dimensional representation of data in which values are represented by colours. Heat maps communicate relationships between data values that would be much more difficult to understand if presented numerically in a spreadsheet.	
High quality data	Data that are complete, timely, accurate, consistent, relevant, reliable, traceable, cleaned, validated, and well documented.	
Human-readable format	Data and code that are commented so that humans can understand what they represent, their design, and their purpose.	
Indigenous data	Data that pertains to (is created or gathered by, or is about) Indigenous peoples.	

Indigenous data sovereignty	Right of Indigenous peoples to own, control, access and possess data that derive from them, and which pertain to their members, knowledge systems, customs or territories.	
Indirect identifier	Data variable or value that, when combined with other identifiers, either within the same dataset or combined with another available dataset, directly discloses the identity of a single living individual, for example, age, geographical location or health condition.	
Informaticist	Person who identifies, defines, and solves information problems using information systems, systems integration, and the management of human interactions with machine and data.	
Information	Aggregation of data to make coherent observations about the world, meaningful data, or data arranged or interpreted in a way to provide meaning.	
Information management advisor	Person with broad knowledge of information management disciplines and who provides guidance and support to program and staff functions on all aspects of managing the information resource.	
Information management specialist	Expert in one or more of the information management disciplines that support the effective and efficient management of information.	
Information silos	Heterogeneous data sources.	
Information technology specialist	Information systems and technology infrastructure manager, expert, or technician.	
Instrument	Device used for making measurements, alone or in conjunction with one or more supplementary devices.	
Instrument output data	Raw electronic data generated by an instrument, analyser, or data logger before any human action on the data and before any processing of the data by automated or semi-automated 3rd-party software or algorithms.	
Integrated access management	Combination of business processes, policies and technologies that allows organisations to provide secure access to confidential data. Integrated access management software is used by enterprises to control the flow of sensitive data in and out of a network.	
Integrity	In the context of data and network security, assurance that information can only be accessed or modified by those authorised to do so. Measures taken to ensure integrity include controlling the physical environment of networked terminals and servers, restricting access to data, and maintaining rigorous authentication practices. Data integrity can also be threatened by environmental hazards, such as heat, dust, and electrical surges.	
Inter-disciplinary	A study undertaken by combining two or more distinct research disciplines. Research based upon a conceptual model that links or integrates theoretical frameworks from those disciplines, uses study design and methodology that is not limited to any one field, and requires the use of perspectives and skills of the involved disciplines throughout multiple phases of the research process.	
International chemical identifier	A non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations.	

International standard	Standard that is used in multiple nations and whose development process is open to representatives from all countries.	Some international standards are promulgated by multinational treaty organisations (e.g., the International Telecommunications Union (ITU); the United Nations Food and Agriculture organisation (FAO)). Some international standards are promulgated by multinational non-treaty organisations (e.g., the International organisation for Standardization (ISO); the International Electrotechnical Commission (IEC)). Some international standards are promulgated by organisations that originated as national industry associations, professional societies, or standards developers, but over time evolved into a global presence with multinational participation (e.g., ASTM International, SAE International, and NFPA International). Annex 4 of the World Trade organisation (WTO) Committee on Technical Barriers to Trade Report 2000 contains a good discussion of what constitutes an international standard. In short, the WTO suggests that a standard may be considered international if the processes and procedures used to

		develop it are transparent, open, impartial, and provide meaningful opportunities for WTO members to contribute to the development of the standard so that the standard does not favour any particular suppliers, countries, or regions. Equally important, the standard must have a global relevance and use.
Interoperable	Data and metadata that can be integrated with other data/metadata and can interoperate with applications or workflows for analysis, storage, and processing. Semantic and syntactic interoperability are the two main types of interoperability.	
Knowledge	Rules and organising principles gleaned from aggregated data. The internalised or understood information that can be used to make decisions.	
Legacy data	Older data that can no longer be accessed or processed easily because they are stored in obsolete formats or systems.	
Linked open data	Data where relationships/connections between them are available to allow easy data access.	A typical case of a large Linked dataset is DBPedia (http://dbpedia.org/), which essentially makes the content of Wikipedia available in RDF. This related collection of interrelated datasets is stored on the Web and available via a common format RDF.
Long-term preservation	Continued access to digital materials, or at least to the information contained in them, indefinitely.	
Machine-actionable	Machine-readable dataset or file format that is structured in such a way as to allow machines to take automated programmed actions as a result.	
Machine-readable	In a form that can be used and understood by a computer.	
maDMP	Machine-actionable Data Management Plan. DMP with actionable functions such as automatic requesting of additional storage, or automatic ingest of dataset publication details.	
Masking	Application of a set of data transformation techniques to de-identify data without any concern for the analytical utility of the data. Applied to direct identifiers such as name and phone number.	

	Masking techniques include, among others, removal of direct identifiers or replacement of direct identifiers with pseudonyms.	
Meaningful use	In the context of health information technology (HIT), minimum government standards for using electronic health records (EHR) and for exchanging patient clinical data between healthcare providers, between healthcare providers and insurers, and between healthcare providers and patients.	
Medium-term preservation	Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.	
Metadata	Data about data. It is data (or information) that defines and describes the characteristics of other data. It is used to improve the understanding and use of the data.	
Metadata catalogue	Catalogue containing metadata records that enables services to find data and services.	
Metadata profile	Document that modifies a metadata standard. May reduce the overall number of metadata elements defined by a standard. May further restrict the optionality of a metadata element, making it mandatory where before it was optional; however, a profile cannot make mandatory elements optional. May further restrict the values allowed in a metadata element. Can be adopted by a standards body, agency, or organisation in place of a metadata standard.	
Metadata record	Set of metadata elements and their values that describe an object. Metadata elements in the record may derive from a metadata profile or standard, and may include different types of metadata (descriptive, administrative, etc.). A metadata record is typically stored within a metadata catalogue or repository.	
Metadata standard	High level, shared representation of the metadata elements related to a dataset, collection, or other digital object. May also provide an XML schema describing the format in which the elements should be stored. Typically, a standard XML format is defined using XML Schema or document type definition (DTD). Standards are typically ratified by national or international standards bodies.	
Migration	Means of overcoming technological obsolescence by transferring digital resources from one hardware/software generation to the next, to preserve the intellectual content of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. Differs from the refreshing of storage media in that it is not always possible to make an exact digital copy or replicate original features and appearance and still maintain the compatibility of the resource with the new generation of technology.	
Minimal metadata	Description of a digital object with a limited number of fields including at least a name and persistent identifier.	
Missing data	Data that are missing on a variable. The missing data can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). The reason why the data are missing informs how it should be curated (e.g., whether to impute the missing data or not).	
Namespace	Uniquely identifies a set of names so that there is no ambiguity when objects having different origins but the same names are mixed together. Using the Extensible Markup Language (XML), an XML namespace is a collection of element type and attribute names. These element types and attribute names are uniquely identified by the name of the unique XML namespace of which they	

	are a part. In an XML document, any element type or attribute name can thus have a two-part name consisting of the name of its namespace and then its local (functional) name.	
Noisy data	Meaningless data, including any data that cannot be understood and interpreted correctly by machines such as unstructured text; any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it.	
Non-identifiable data	Data that could not lead to the identification of a specific individual, to distinguishing one person from another, or to personally identifiable information. These may be data that have been de-identified, or that could not lead to personally identifiable information in the first place.	
Normalisation	Organising data into tables so that the results of using the database are always unambiguous and as intended. Normalisation is typically a refinement process after the initial exercise of identifying the data objects that should be in the database, identifying their relationships, and defining the tables required and the columns within each table. First normal form (1NF) is the "basic" level of normalisation: Data and information are contained in two-dimensional tables with rows and columns. Each column corresponds to a sub-object or an attribute of the object represented by the entire table. Each row represents a unique instance of that sub-object or attribute and must be different in some way from any other row (that is, no duplicate rows are possible). All entries in any column must be of the same kind. For example, in the column labelled "Date," only dates are permitted. In Second normal form (2NF), the tables are in first normal form and, in addition, each column in a table that is not a determiner of the contents of another column must itself be a function of the other columns in the table. At the second normal form, modifications are still possible because a change to one row in a table may affect data that refers to this information from another table. In Third normal form (3NF), the tables are in second normal form and, in addition, there is no transitive functional dependency. For example, if A is functionally dependent on B, and B is functionally dependent on C, then C is transitively dependent on A via B. In Domain/key normal form (DKNF), a key uniquely identifies each row in a table. A domain is the set of permissible values for an attribute. By enforcing key and domain restrictions, the database is assured of being freed from modification anomalies. DKNF is the normalisation level that most designers aim to achieve.	
OAI repository	Type of repository with a network accessible server that can process the 6 OAI-PMH requests in the manner described in the OAI Implementation Guide.	
Object attribute	Object model that is the logical attributes or properties associated with a particular object. In a data object this would be the associated properties.	
Object model	Collection of descriptions of classes or interfaces, together with their member data, member functions, and class-static operations.	
Object property	Characteristics of any digital object can be described by a number of properties which are typically stored in metadata and/or PID records.	
Ontology	Shared and standardised list of words, terms and phrases to describe components of a particular discipline or domain, along with a taxonomy of their relations. Compare this to a controlled vocabularies, which tend not to include a structure of relations between their terms. Ontologies are	

	typically developed by domain-specific institutions or communities to aid in the precise referencing of elements. RELATED TERM. Controlled vocabulary.	
Open Archives Initiative Protocol for Metadata Harvesting	Low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP.	
Open data	Data that are accessible, machine-readable, usable, intelligible, and freely shared. Open data can be freely used, re-used, built on, and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.	
Open science	Scientific knowledge that is openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community.	
ORCID	Unique identifier for researchers. Disambiguates researchers and allows researchers to connect their ID with additional professional information including affiliations, grants, and publications.	
Original repository	Type of repository where the original copy of data was stored and probably a data identifier registered.	
Persistent identifier	Long-lasting digital reference to an object that gives information about that object regardless of what happens to that object. Developed to address link rot, a persistent identifier can be resolved to provide an appropriate representation of an object whether that object changes its online location or goes offline.	
Persistent uniform resource locator	URL that points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client.	
Personal information privacy	The World Wide Web Consortium's Platform for Personal Privacy Project (P3P) offers specific recommendations for practices that will let users define and share personal information with Web sites that they agree to share it with. The P3P incorporates a number of industry proposals, including the Open Profiling Standard (OPS). Using software that adheres to the P3P recommendations, users will be able to create a personal profile, all or parts of which can be made accessible to a Web site as the user directs. A tool that will help a user decide whether to trust a given website with personal information is a Statement of Privacy Policy that a website can post.	
Personally identifiable information	Data that relate to a living individual who can be identified from those data or those data plus other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual. Any information that can be used to distinguish one person from another and can be used for de-anonymising anonymous data can be considered personally identifiable data.	
PID attribute	Single data element related to a PID and part of its record content.	
PID domain	For a single identifier, the class of entity it refers to. For a PID system, the typical class of entities it is intended to be used for, such as digital objects, physical objects, bodies, actors.	

PID record	Type of record (and organisation) that stores an instance of an executable/understandable PID. The content of a PID record distinguishes a registered digital or data object from other digital objects. A PID record is a type of record that includes property information that characterises the digital object it is identifying. Important parts of a PID record are location and checksum. However there is a large variation in usage. In some data models the PID is simply used as a unique label with an empty record. A PID record has a lifecycle including creation, publication, curation and destruction.	
PID resolution	The process of resolving a PID to a useful state of information about a digital object by using a globally available system.	
PID service	Service that provides a connection between a PID and its target object.	
PID system	Consists of at least one PID resolver, a name schema and a defined mechanism for issuing PIDs that conform to the name schema.	DOI, Handle System, URN, ARK, PURL, ORCID, ROR, etc.
Pipe-separated values	Values in a table presented as a series of ASCII text lines organised so that each column value is separated by a pipe (|).	
Preservation	An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.	
Preservation metadata	Documents actions that have been undertaken to preserve a digital resource such as migrations and checks sum calculations.	Metadata Encoding and Transmission Standard (METS).
Privacy governance	Monitoring the risk to privacy posed by data requests from researchers, and the practices of data custodians in providing data (information governance) to ensure that confidentiality is protected. Such governance requires specialised knowledge of technology, law, and statistical methods.	
Privacy-preserving data linkage	Data linkage where the resulting product has been de-identified.	
Proportionate governance	Keeping the procedural mechanisms that researchers and data custodians must follow when engaged in data sharing and linkage proportional to the degree of risks associated with such practices. Proportionate governance operates in situations that are too variable to be regulated by hard laws (e.g., custom data access requests). It requires that analytical judgments be performed to ensure that the governance mechanisms deployed for a given research proposal correspond to the level of risk it entails. Proportionality is an important cross-cutting consideration across all types of governance that are put in place.	
Proprietary	File formats of datasets that are specific to a company, organisation or individual that are not in wider use, and can therefore impact on the re-usability and preservation of datasets.	
Protocols	A formal or official record of scientific experimental observations.	
Provenance	A type of historical information or metadata about the origin, location or the source of something, or the history of the ownership or location of an object or resource including digital objects. For example, information about the Principal Investigator who recorded the data, and the information concerning its storage, handling, and migration.	

Provenance metadata	Information concerning the creation, attribution, or version history of managed data. Provenance metadata indicates the relationship between two versions of data objects and is generated whenever a new version of a dataset is created. Provenance information is gathered along the data lifecycle as part of curation processes. A finer level of provenance metadata would be concerned only with data flowing between various stores such as curated databases and managed repositories. Provenance metadata is designed to allow queries over the relationship between versions, and includes either or both fine-grained and coarse-grained provenance data. Different applications may store different provenance data.	Examples include: (i) the name of the program that generated the new version, (ii) the commit id of the program in a code version control system like GitHub, (iii) the identifiers of any other datasets or data objects that may have been used in creating the new version.
Quality assurance	The process or set of processes used to measure and assure the quality of a product.	
Quality control	The efforts and processes put in place to ensure that the management of research data is of sufficient quality to allow researchers to carry out their work effectively.	
Raw data	Data that have not been processed for meaningful use. Although raw data have the potential to become information, they require selective extraction, organisation, and sometimes analysis and formatting for presentation.	
Re-use	The re-analysis of a dataset or combination of datasets outside of the original research purpose for which the dataset was created.	
Real-time data	Data that are being received, processed and stored at the time of their occurrence with only small delays. Examples include: stock quotes, manufacturing statistics, Web server loads, data warehouse activity and sensor feeds to data collectors. Real-time data are often used for navigation or tracking. Real-time data are data streams that are typically generated by sensors and received via direct networking connections.	
Record	(noun) Sometimes called a row, a group of fields (sometimes called columns) within a table that are relevant to a specific entity. Multiple records are contained in a file or dataset.	For example, in a table called 'Client contact information', a row would likely contain fields such as: ID number, name, street address, city, telephone number, etc.
Record provenance information	Information for a data object that includes: the person who deposited the data object in the repository, the source of the data object, the date when the object was deposited, and authenticity information needed to link the data object to its original source.	
Record standardisation	Process in which files are first parsed (assigned to appropriate fields in a record) and then translated to a common format. Data often lack consistency simply because there are many ways of saying the same thing. Standardising the record ensures that when a query is run for a particular field, accurate results will be returned.	
Records retention schedule	Policy that depicts how long data items must be kept, as well as the disposal guidelines for these data items.	

Referable data	Type of data (digital or not) that is persistently stored and which is referred to by a persistent identifier. Digital data may be accessed by the identifier. Some data object references may access a service on the object.	
Reference model	Design covering a class of frameworks with the following characteristics: (1) it can be used to generate more specific models that still belong to the class and (2) it can be used to compare a concrete framework design to identify whether it belongs to the same class.	
Reformatting	Copying information content from one storage medium to a different storage medium (media reformatting) or converting from one file format to a different file format (file reformatting).	
Refreshing	Copying information content from one storage media to the same storage media.	
Registered data	Data that have gone through a registration process and have been assigned an identifier metadata to aid in their search and retrieval.	
Registry	Database containing information about trusted repositories that are provided by repository managers and are useful for human and machine users. These registries do not contain information about all metadata descriptions of digital objects, nor do they offer a list of PIDs of all stored digital objects. They do offer information based on standardised types on how to retrieve such information (e.g., the port under which OAI-PMH can be accessed to offer metadata). A registry requires the assignment of a permanent, unique and unambiguous identifier to each item.	
Reliability	Likelihood of observing the same result if the data were to be collected in another sample. That is, we can rely on the results to be accurate at a different time or in a different context.	
Remote data access	Ability to access and download data from a repository.	
Replica number	Type of metadata used as part of a replication process or access.	
Replication		
Replication (object)	Generation of a copy of a data object that is referenced by the same name, but with a different replica number. When changes are made to the data object, the replica can be updated to track the changes. As part of replication, data may be given a PID for a repository. Enhanced metadata may be stored in a repository as part of replication. A PID should allow replicated objects from different communities to be identified as such.	
Replication (method)	Repeated measurement or observation of the same object or phenomenon.	
Repository	Physical or digital storage location that can house, preserve, manage, and provide access to many types of digital and physical materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the physical and digital material to be authentic, reliable, accessible and usable on a continuing basis.	
Representation	Resource that conveys either the content of a resource (if it is a digital object instance), or provides a digital object that conveys the intention of the resource in a form useful to a user (machine or human).	
Representation object	Contains provenance, description (such as format, encoding scheme, algorithm), structural, and administrative information to provide context for a data object. This is a form of metadata.	

Reproducibility	Ability to replicate the results of a study using the same input data and procedures used by the original investigator.	
Reproducible research	Results that can be replicated using the documented data, code, and methods employed by the author or provider without the need for any additional information or needing to communicate with the author or provider.	
Repurposed data	New datasets obtained by combining data appropriately from a variety of existing files, generating new data products that did not previously exist. Repurposed data result from data wrangling.	
Requirements analysis	Process of determining user expectations for a program, system, dataset, or product. Requirements analysis is a team effort that must take into account hardware, software, end use, and human factors engineering expertise. Requirements analysis also requires skills in dealing with people. Requirements analysis involves frequent communication with end users to determine specific feature expectations, resolution of conflict or ambiguity in requirements as demanded by the various users or groups of users, avoidance of feature creep and documentation of all aspects of the project development process from start to finish. Energy should be directed towards ensuring that the final system or product conforms to client needs rather than attempting to mould user expectations to fit the requirements.	
Research data	Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.	
Research data lifecycle	Entire period of time that research data exists. This lifecycle describes the flow of research data starting from planning, collecting, processing, analysing, preserving, sharing and finally reusing the research data. Research data often have a longer lifespan than the research project.	
Research data management	Storage, access and preservation of data created or collected in the course of research. Research data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used to long term preservation of data deliverables after the research investigation has concluded. Specific activities and issues that fall within the category of data management include: File naming (the proper way to name computer files); data quality control and quality assurance; data access; data documentation (including levels of uncertainty); metadata creation and controlled vocabularies; data storage; data archiving and preservation; data sharing and reuse; data integrity; data security; data privacy; data rights; notebook protocols (lab or field) as required.	
Research data management infrastructure	Configuration of staff, services and tools assembled to support data management across the research lifecycle and to provide comprehensive coverage of the stages making up the data lifecycle. It can be organised locally and/or globally to support research data activities across the research lifecycle.	
Research data publication workflow	Activities and processes in a digital environment that lead to the publication of research data, associated metadata and accompanying documentation and software code on the Web. In contrast	

	to interim or final published products, workflows are the means to curate, document, and review, and thus ensure and enhance the value of the published product. Workflows can involve both humans and machines and often humans are supported by technology as they perform steps in the workflow. Similar workflows may vary in the details depending on the research discipline, data publishing product and/or the host institution of the workflow.	
Research governance	Activities and policies that ensure that the benefits to society of research outweigh any risks, from both an ethical and legal perspective.	
Research metadata format	Acceptable formats for transmitting and sharing research metadata.	ISO 19115-2:2009
Research Organization Registry	Community-led registry of open, sustainable, usable, and unique identifiers for every research organisation in the world.	
Research results	Findings and assertions resulting from, and justified by, scientific method.	
Research software	Software written to create, generate, analyse or display research data.	
Resource authorisation	Process of deciding if a subject (person, program, device, group, role, etc.) is allowed to have access to or take an action against a resource. Authorisation relies on a trusted identity (authentication) and the ability to test the privileges held by the subject against the policies or rules governing that resource to determine if an action is permitted for a subject.	
Retention period	Metadata operation to create state information for a data object that defines the date when retention of the data object should be evaluated. The retention period must have an associated disposition policy for deciding what to do when the retention period expires.	
Reusable	One of the four FAIR principles. Reusable data is data that can be utilised to replicate research findings and/or can be analysed in settings outside of the original context in which it was produced or collected. The reusability of research data can depend on its format, licensing, and the richness of the relevant metadata.	
Schema		
Schema (database)	Organisation or structure for a database. The activity of data modelling leads to a schema. (The plural form is schemata.) The term is used in discussing both relational databases and object-oriented databases. The term sometimes refers to a visualisation of a structure and sometimes to a formal text-oriented description.	Two common types of database schemata are the star schema and the snowflake schema.
Schema (AI)	Formal expression of an inference rule for artificial intelligence (AI) computing. The expression is a generalised axiom in which specific values or cases are substituted for each symbol in the axiom to derive a specific inference.	
Science and technology data	Qualitative or quantitative attributes of a variable or set of variables. Data refers to representations of physical, biological or chemical facts, typically the results of measurements/observations. It also includes related socio-economic and cultural representations. Data are normally in a structured, tabular, numeric, character, geo-referenced, and/or computer-readable format.	
Scientific data infrastructure	Facilities and systems to enable researchers to create, store and share the data resulting from their experiments, and to find, access and process the data they need.	

Scientific data services	Services that assist organisations in the capture, storage, curation, long-term preservation, discovery, access, retrieval, aggregation, analysis, and/or visualisation of scientific data, as well as in the associated legal frameworks, to support disciplinary and multidisciplinary scientific research.	
Scientific method	Series of steps to follow for the systematic discovery of knowledge: 1. Ask a research question which is grounded in existing research and/or theory, 2. Generate a hypothesis, 3. Collect/retrieve the data, 4. Analyse the data, 5. Interpret the results, 6. Report the results.	
Scientific workflow	Set of chained operations used in the execution of the scientific method. Workflows produce outputs that may include, for example, visualisations and analytical results. Preserved workflows are important for reproducible research. They simplify complex sequences of activities and enable researchers to automate and track the provenance of the work in workflow execution.	The simplest computerised scientific workflows are scripts that can involve several ingredients such as data, programs, models and other inputs such as human or sensor observations.
Semantic data	Data that are tagged with particular metadata that can be used to derive relationships between data.	
Semantic interoperability	Ability of computer systems to transmit data with unambiguous, shared meaning. Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems. Semantic interoperability is achieved when the information transferred has, in its communicated form, all of the meaning required for the receiving system to interpret it correctly, even when the algorithms used by the receiving system are unknown to the sending system. Syntactic interoperability is a prerequisite to semantic interoperability. Semantic interoperability ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties; in other words, what is sent is what is understood.	
Semi-structured data	Data that have not been organised into a specialised repository, such as a database, but that nevertheless have associated information, such as metadata, that makes them more amenable to processing than raw data. Semi-structured data lie somewhere between structured and unstructured data. They are not organised in a complex manner that makes sophisticated access and analysis possible. However, they may have information associated with them, such as metadata tagging that allows elements contained to be addressed.	Example: A Word document is generally considered to be unstructured data. However, metadata tags could be added in the form of keywords and other metadata that represent the document content and make it easier for that document to be found when people search for those terms — the data are now semi-structured. Nevertheless, the document still lacks the complex organisation of a database, so falls short of being fully structured data

Sensitive data	Data that must be protected against unintended access or disclosure. This includes information regarding an individual, organisation or other entity.	
Service object	Type of digital object containing executable code, considered as a unit.	
Short-term preservation	Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.	
Standard	Set of agreed-upon and documented guidelines, specifications, accepted practices, technical requirements, or terminologies that have been prepared by a standards developing organisation or group, and published in accordance with established procedures. These can be mandatory or voluntary and are distinct from Acts, regulations, and codes, although standards can be referenced in those legal instruments.	
Standard Operating Procedure for the collection of harmonised or integrated data	Written methods, instructions, and tools that, when applied in different data collection contexts produce data that are ready to be harmonised or integrated without further manipulation.	
Standardisation	Process of converting data to a common format to ease analysis and facilitate the comparison or collation of different sets of data.	
Statistical de-identification	Application of data transformation techniques to de-identify data in such a manner that the resulting transformed fields retain a very high analytic value.	
Sticky bits	User ownership access-right flag that can be assigned to digital objects such as directories. When the sticky bit flag is set, files added to the directory will inherit the access permissions associated with the directory.	
Storage location	Physical storage location where a data object will be stored upon ingestion into a data repository. This requires identifying the IP address and the physical path name within the storage location where a data object will be stored. The sequence of these chained activities is conceptualised as a workflow object. For retrieval, the data object location is specified by the storage location and the physical path name.	
Structural metadata		
Structural metadata (information)	Type of metadata that indicates how compound objects are put together.	Examples include how pages are ordered to form chapters; how data are organised in a table; how datasets are organised in a collection.
Structural metadata (computing)	Metadata type that tells a computer how to assemble a digital object.	

Structured data	Data elements that have been organised into a consistent format and data structure within a defined data model such that the elements can be easily addressed, organised and accessed in various combinations to make better use of the information, such as in a relational database.	
Syntactic interoperability	Defines the structure or format of data exchange and is achieved through tools such as XML or SQL standards.	
System metadata	Digital entity properties that are generated by the data management system (e.g., creation time; owner; storage location; data retention period; the length of time a digital entity will be retained).	
System of record	Information storage and retrieval system that serves as the authoritative source for a particular data element in a system containing multiple sources of the same element. To ensure data integrity, a single system of record must always exist for each and every data element.	
Tab-separated values	Values in a table in a series of ASCII text lines organised so that each column value is separated by a TAB from the next column's value and each row starts a new line.	
Table	Organised grouping of columns (i.e. fields). In a relational database, a table (sometimes called a file) organises the information about a single topic into rows and columns. The process of normalisation determines how data will be most effectively organised into tables.	
Tabular data	Data that are arranged in tabular forms, in rows and columns.	
Technical metadata	Information describing the technical processes used to produce, or required to use a digital object.	
Temporary version	Copy of a data object such as a file during the course of routine operations.	
Tombstone record	Metadata record relating to a dataset that has been destroyed, deleted, corrupted, or otherwise made unavailable that acknowledges the past existence of the data and preserves the details of it.	
Topical metadata	Describes the topic or "aboutness" of an information/data object – what are these data about. In order to make sense to an agent or systems, this may include the use of a variety of vocabularies for describing, subjects, topics, categories, etc.	
Transdisciplinary	Research efforts conducted by investigators from different disciplines working jointly to create new conceptual, theoretical, methodological, and translational innovations that integrate and move beyond discipline-specific approaches to address a common problem.	
Trusted Digital Repository	Infrastructure component that provides reliable, long-term access to managed digital resources. It stores, manages, and curates digital objects and returns their bit streams when a request is issued. Trusted repositories undergo regular assessments according to a set of rules such as defined by CoreTrustSeal or TRAC (ISO 16363). Such an assessment has the potential to increase trust from its depositors and users. Certain quality criteria need to be met to distinguish trusted repositories from other entities that store data, such as notebooks or lab servers.	
Unified data management platform	Centralised computing system for collecting, integrating and managing large sets of structured and unstructured data from disparate sources.	
Uniform resource identifier	String of characters used to identify or name a resource on the Internet. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols.	

Uniform resource namespace	Internet resource with a name that, unlike a URL, has persistent significance – that is, the owner of the URN can expect that someone else (or a program) will always be able to find the resource. A frequent problem in using the Web is that Web content is sometimes moved to a new site or a new page on the same site. Since links are made using Uniform Resource Locators (URLs), they no longer work when content is moved.	
Universal Numeric Fingerprint	Unique signature of the semantic content of a digital object. It is not simply a checksum of a binary data file. Instead, the UNF algorithm approximates and normalises the data stored within. A cryptographic hash of that normalised (or canonicalised) representation is then computed. The signature is thus independent of the storage format.	The same data object stored in, say, SPSS and Stata, will have the same UNF.
Universally Unique Identifier	128-bit number used to guarantee unique identity for objects on the internet over time.	
Unstructured data	Data that have not been organised into a format and identifiable data structure that makes them easy to access and process. These data can often be searched as long as they are digital, but they are difficult to use for computer analyses.	
Usable data	Data that can be used: delivered in a form that meets the needs of different end-user audiences, is ready for the tasks that the end-user needs to accomplish, and that has been adapted to the end-user's needs. Usable data have been cleaned, structured, are in machine readable format, fully documented, and ready for analysis and interpretation.	
Use case	Methodology used in system analysis to identify, clarify, and organise system requirements. The use case is made up of a set of possible sequences of interactions between systems and users in a particular environment and related to a particular goal. It consists of a group of elements (e.g., classes and interfaces) that can be used together in a way that will have an effect larger than the sum of the separate elements combined. The use case should contain all system activities that have significance to the users. A use case can be thought of as a collection of possible scenarios related to a particular goal, indeed, the use case and goal are sometimes considered to be synonymous.	
Verify checksum	Generate a unique reduced representation for a data object by applying a procedure and compare the result to the original reduced representation that has been stored as provenance information.	Examples include: a checksum, a hash, a digital signature.
Version control	Control over time of data, computer code, software, and documents that allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting mistakes. Version control generates a (changed) copy of a data object that is uniquely labelled with a version number. The intent is to track changes to a data object, by making versioned copies. Note that a version is different from a backup copy, which is typically a copy made at a specific point in time, or a replica.	
Visualisation	The representation of a dataset in visual form, for example, a chart, diagram or picture, used to gain insights that tabular data would not provide.	
Web resource	Addressable unit of information that is addressed through Uniform Resource Identifiers (URIs). The early notion of static addressable documents or files has evolved to a more generic and abstract definition. Every 'thing' or entity that can be identified, named, addressed or handled in any way	Examples include: an electronic document or data stored on the Web, an

	whatsoever in the web at large or in any networked information system. Each resource must have a URI.	image, a service (e.g., “a weather report), a collection of other resources.