Key concepts in Open Science

Access control (management)

Mechanisms which are used to manage user identification (authentication) and by which information will be accessible to each user (authorisation).

(Remote) access system

A remote access system allows researchers to gain access to data (usually via the Internet), without having to transfer data to the researcher's own computer. The researcher can check, combine and analyse data using a remote desktop. Original data remains protected behind a firewall, thus ensuring stringent data protection requirements.

Anonymisation

Process of removing confidential information (direct and indirect personal identifiers) from research data which may be needed for ethical (to protect people's identities in research), legal (to not disclose personal data) or commercial reasons. The process is time consuming, costly, and can use different techniques and toolkits.

Authentication

Process of verification of the identity of a user who requests access to a resource.

Authenticity

Set of criteria for determining the goodness, reliability, validity, and rigor of qualitative *research.*

Authorisation

Permission given to a verified user to use certain data, protected items or other service elements.

(Open) Access

The policy and practices of providing unrestricted access to research data. Initially, this concept was mostly related to the open accessibility of scientific publications. See also Open Science.

Creative commons license

One of several public copyright licenses that enable the free distribution of an otherwise copyrighted work. Under a Creative Commons license, the author relinquishes a portion of his or her copyrights, giving the desired rights for users, viewers or experimenters to use the work.

Data

Digital material that does not involve interpretation (compare with the term information, which refers to interpreted information).

(Meta)data

A set of data that describes and gives information about other data including the context, content and structure of data as well as its management and processing for the entire lifecycle. Metadata are widely used in recourse discovery (finding by criteria, identifying resources, its locating, regrouping); organizing resources (organizing links to resources, building these pages dynamically from metadata stored in databases); facilitating interoperability (using defined metadata schemes, shared transfer protocols, cross-system search etc.); digital identification (elements for standard numbers, e.g., ISBN , locating a digital object by a file name , URL , some persistent identifiers); archiving and preservation as a key to ensuring that resources will survive and continue to be accessible into the future. Metadata can include descriptive information, which describes a resource for purposes such as discovery and identification; structural entries which indicate how compound objects are put together (for example, how pages are ordered to form chapters), and administrative data, which provide information to help manage a resource (data of creation, file type, access rights. preservation data and other technical information).

(Micro)data

Microdata is any data gathered on a very small scale rather than aggregate data. E.g., in statistics microdata is usually collected from individual response data in surveys or censuses, or can be also taken from population registers.

(Official) data

Official data may be printed or digital text, images, solutions, draft documents or documents, registers and register data submitted to or produced by an authority in the course of carrying out its mandate. Data becomes public domain when the processing of a matter has been concluded or a report has been completed, even if the matter it addresses is still in progress. (Source: CSC 2010 Research Data – Essentials for Decision-makers)

(Raw) Data

Unrefined data which is produced and collected in a variety of forms by observations, simulations, conducting experiments and next can undergo numerous transformations and various processing phases during its lifecycle. E.g., in natural sciences, raw data is often produced by research instruments, such as telescopes, satellites, synchrotrons and, increasingly, by computer simulations. In the humanities and social sciences, raw data comes from, for example, interviews and surveys.

(Research) Data

Information, records, and tangible products arising from or associated with research conducted. This can include raw or primary data, processed data and published data.

(Research) Data in TTA

In the National Research Data Project (TTA), research data comprehends national digital data and data reserves produced by means of public funding. Data includes data produced in research and data used in research. Therefore, it should be kept in mind that all data useful to research was not necessarily originally collected for

research use, but rather for administrative monitoring, control, reporting and statistics.

Data balance sheet

An organisation's report on what type of information/data they are managing, how it is being distributed and used.

Database and data warehouse

A database is a storage of data in its most generic form. Data, such as text, audio and video in different formats can be stored in a database (DB). Building a database is based on the modelling and presentation of the data being stored in it, generally using a certain description language or technology. Databases are categorised according to the programming model. Data warehousing (DW, DWH) is the storage of data, typically summarized and prepared for analytical purposes (reprting, data analysis, business soulutions), in contrast to "operational" databases, which are used in the real-time.

Database software/environment, Database software/management system

Collection of utility programs supporting the actual database engine: creation, querying, update, and administration of databases. Database management systems also include auxiliary programs that facilitate maintenance, which make it possible for the data stored in the database engine to be backed up and restored as well as exported to other systems and the user interface.

Data conversion and migration

Data conversion deals with changes required to move or convert data from one physical environment format to that of another for purpose of interoperability, like moving data from one electronic medium or database product onto another format. When making conversion, the preservation of the readability and comprehensibility of the digital data must be quarantined.

(Data) Curation

Data curation is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation. In science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format, such as an entry of a database.

Data infrastructure

Data infrastructure is part of an e-infrastructure that contains all interoperable basic services and tools for the production, storage, sharing and use of data.

Data mining

A set of methods, which are used to find key information within large volumes of data. Data mining can be used in a wide variety of applications, as only raw data is needed. Typically, the data used in data mining involves, for example, measurements from industrial processes, extracts from a customer database or log data from a web server. The definition of data mining does not limit the methods that can be used.

Data life cycle management

Data life cycle management is a policy-based approach to managing the flow of data throughout its life cycle: from the production of raw data to the final processing consumption and storage till the time when data becomes obsolete and is deleted.

Data management plan

A data management plan is a document that describes overall plans for data creation, organization, documentation, its storage and sharing. It takes into account issues such as data protection, preservation, and curation, and provides a framework that supports researchers and their data throughout the course of their research and beyond. The management plan is required by the Academy of Finland.

Data policy

In a TTA context, data policy involves a nationally co-ordinated policy plan and programme, which steers the development of operations and practices in various administrative fields, and includes a specification of various actors' roles and delegation of assignments for the management of plans and intellectual property, data protection issues, pricing and the specification of accessibility and terms of use.

Data protection

Data protection is a practice of protecting personal data from unauthorised use during their processing.

Data reuse

Private persons and legal persons may copy, edit, publish and distribute data for different purposes. A large percentage of raw data and research data collected by means of public funding can be used in research as well as in the development of various everyday products and services.

Data system architecture

Data system architecture describes the structural components for data target area, their externally visible properties and the connections and dependencies between them. The architecture forms the framework for the design and execution of the system, and it steers development of the system structure throughout its lifecycle. It also serves as a platform for making bilateral solutions by system development and maintenance stakeholders (management, users, and).

Data security

Administrative and technical measures for ensuring that: data is only accessible to persons authorised to use it; no data can be altered by unauthorised persons; and data and data systems are only available to authorised persons. Data protection concepts include access control, confidentiality, privacy, verification, integrity and security.

Data type

Data type is a data storage format that can contain a specific type or range of values. E.g.; in computer programming data are storing in variables, and each variable must be assigned a specific data type. Some common data types include integers, floating

point numbers, characters, strings, and arrays. They may also be more specific types, such as dates, timestamps, Boolean values, and other formats. Data types are also used by database applications. The fields within a database often require a specific type of data to be input. For example, a company's record for an employee may use a string data type for the employee's first and last name. The employee's date of hire would be stored in a date format, while his or her salary may be stored as an integer. By keeping the data type uniform across multiple records, database applications can easily search, sort, and compare fields in different records.

Digital information

Digital information refers to data processes and stored in only digital form. Physical data is converted from analog to digital using analog to digital converters (e.g., scanning printed document). This digital data is what is processed to give digital information.

Document publicity

According to the Act on the Openness of Government Activities, public documents are public domain. Public documents are drafted and issued by authorities as well as documents sent or given to them, which are in their possession. The concept of 'document' is extremely broad in law. In addition to written documents, technical records and visual presentations can also be considered documents. A document that is being drafted by an authority is not yet public, nor is any proposal, draft, report, statement, memorandum or other account produced within the confines of that authority. Any matter or document that is specified as such by law must be kept confidential. In terms of publicity, documents can be divided into public documents, (non-public) draft documents and confidential documents.

E-infrastructure

The term e-infrastructure refers to an advanced IT resources and services environment that supports researchers through distributed global collaborations enabled by the Internet. E-infrastructure brings together research equipment, high performance computing and network resources and services, and research data produced as well.

ESFRI - European Strategy Forum on Research Infrastructures

Co-operative body of EU Member and Associate countries which functions as a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach. Its mission to leads policy-making on research infrastructures in Europe and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and international level.

E-science

Science that uses big data collections and advanced scientific instruments and methods, which are based on e-infrastructure services. E-science will refer to the large scale science researches that are conducted through distributed global collaborations enabled by the Internet. E-science is based on e-infrastructure' services which provide an access to big data collections and high performance computing resources back to the broader academic community.

Extraction costs

Costs incurred by the editing or copying of data for submission, delivery or other actions involving the transfer of data.

Federated system

A large-scale system (business lines, IT system, database or application), which is comprised of several semi-autonomous de-centrally organized subsystems and allows controlled sharing and exchange of information among these autonomous components by communication messages. Federation may also involve a control system that maintains user identification and personal data.

Fifth Freedom

The free mobility of researchers, data and technologies that are realized through open science, open standards, and open access policies and supported by e-infrastructures.

Fourth Paradigm of science

In scientific research, the first three paradigms are experimental, theoretical and (more recently) computational science. According to Microsoft research, the forth paradigm is data-intensive scientific discovery, which is made possible by a new generation of computers, software and research methods and instruments. It provides researchers with tools to conduct "big data" experiments and make calculations at orders of magnitude, scales and quantities that were not possible before. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Geographic information, spatial data

All data that contains a direct or indirect reference to a certain place or geographic area(s). Spatial data describes a certain theme or phenomenon covering a specific geographic area. Spatial data themes include soil and bedrock, hydrography and climate, habitats and biotopes, land use, cadastral parcels, buildings and population, production and industrial facilities, transport and communication networks, power grids. Digital aerial and satellite images are key spatial data. Please see also Spatial data infrastructure.

Hosting service

Offering an operating environment as a service, such as providing web server functions and maintenance to outside parties.

INSPIRE Directive

The INSPIRE Directive 2007/2/EC establishes a spatial data infrastructure for the sharing and use of spatial information held by public administration organisations throughout Europe. Implementation of the Directive is based on the phased development of interoperability of national spatial data infrastructures. In Finland, the Spatial Data Infrastructure Act and supplementary Decree were enacted for implementation of the Directive.

Integration environment/architecture

Integration architecture describes how a set of different applications share and transfer data with one another. Integration architecture describes the principles by which application interfaces with other systems are to be made. When defining integration architecture, a decision must be made as to how the whole will be integrated uniformly, not just integration of a single interface on a case-by-case basis.

Interface (open/closed)

Interface is commonly software or a software component by means of which different programs, databases or data systems can exchange data with one another. This is known as an Application Programming Interface (API). The software, system or technical device interfaces can be set as either open or closed. An interface whose properties are all public and which can be used without restrictions (e.g., a program that uses the interface without requiring any specific consent from the interface manufacturer or compulsory license fees) is called open. The properties of a closed interface are not public or they may not be used without restrictions.

Interoperability

Generally, interoperability is seamless operability at different levels of operation within actors and among them. Levels of operation include strategies, services, processes, organisations, data, specifications, terminology and technology. Enterprise architecture planning has been undertaken in order to improve interoperability in public administration.

Long-term preservation

Long-term preservation is the reliable preservation of information for tens or hundreds of years. Although hardware, software and file formats age, the information must be kept usable.

Open Science

Open Science is a concept in which research methods, data and outcomes are all thoroughly documented and public accessible.

Open Standard

An open standard is a standard that is publicly available and has various rights to use associated with it, and may also have various properties of how it was designed (e.g., open process).

Persistent identifier (PID)

This identifier is a unique, object-specific character string, which is used to make clear references to the object. The persistent identifier is assigned to the object for its entire lifecycle. There are numerous identifier systems for digital data, such as Uniform Resource Name (URN) or Digital Object Identifier (DOI).

Primary use

The first intended use stated in the original research plan or otherwise. See also Secondary Research, re-use.

Public information

Public information is the information which is characterized as follows: firstly, it is comprised of anything falling outside the purview of copyright or other legislation; secondly, public information can be a certain type of data produced by public administration as part of its official duties. Data produced by public administration is part of a broader category, i.e. public sector information. See also *Public sector information*.

Public good

An economic concept referring to any good which multiple people may use multiple times without consuming it, or whose consumption does not prevent others from consuming the same good.

Public sector

Public sector is a part of the national economy owned by the state or municipalities. In its broadest sense, the public sector comprehends not only the actual state and municipal functions, but also public welfare funds (e.g., the Social Insurance Institution of Finland), joint municipal authorities, state research institutes, public enterprises and state-owned companies.

Public sector information (PSI)

Data collected and produced by public organisations (e.g., digital maps, weather information, traffic information, business and economic information, legal information).

Reference information

Metadata describing an object that is attached to the object by the author when transferring the data into storage.

Registers

In carrying out official duties, the public sector produces comprehensive data, which is collected in registers and data systems. Statistical operations also create statistical registers.

Register authorities

Register authorities maintain base registers or, where authorities for a specific area are concerned, registers and data systems containing administrative monitoring and control data. Authorities and institutions performing these kinds of duties are often required to report on situations.

(Base) Registers

A national administrative register maintained by authorities. Base registers include: Population Information System (personal data, building and residential information), Land Information System (cadastre, title and mortgage register), company and organisation information (Trade Register, Finnish Business Information System data, Register of Associations, and Business Register).

Register information

Register information is data on a certain, mandatory destination set maintained by register authorities. At the unit level, register information involves data on

phenomena under review (a person, business, the environment, etc.), which is recorded in administrative or statistical registers compiled by authorities.

(Public Sector Data) Registers

Public Sector Data Registers are key reserves of public administration data, which are used broadly across agency lines. These data registers are a key data reserve. Public Sector Data Registers also include base register-type registers, registers serving a certain purpose and other, important spatial databases.

Register research

Register research is research that uses register data. The research may be entirely based on register data or the register data can be used as supplementary information for other data (e.g., interview or survey date or clinical and sampling data).

Representation information is metadata that to be attached to a part of the information being stored. Representation information is information that makes the data readable - the file format description, semantics and other information, which are used to change the bits stream into a human or machine-readable and comprehensible form.

Research infrastructure (RI)

Research infrastructure is entire complex of research instruments, equipment, data and services, which makes research and development innovation possible in various phases, promotes organised research work and maintains and develops research capacity. The term 'research infrastructures' refers to facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields. This definition covers: major scientific equipment or set of instruments; knowledge based-resources such as collections, archives or structured scientific information; enabling ICT-based e-Infrastructures such as Grid, computing, software and communication networks; any other entity of a unique nature essential to achieve excellence in research. Such infrastructures may be 'single-sited' or 'distributed' (a network of resources).

(Secondary) Research, re-use

Secondary research refers to any use in which the primary or secondary users apply research data for research purposes other than originally intended, i.e. as stated in the research plan.

(Sector) Research Institute

Coming into wider use in 2006, Sector Research institute is a general term for research institutes that fall within the budget of Finnish ministries. Sector research institutes are often data producers within their respective areas of administration. See also *State research institutes*.

(State) Research Institute

State Research Institute is an institute that produces data primarily on the state and trends of its respective fields. The State research institute also maintains data systems that contain monitoring data for its own administrative area. It can also be

referred to, for example, research authorities, authorities conducting research or organisations conducting state research.

SHOK – *Strategic* Centres for Science, Technology and Innovation

Comprised of enterprises, universities and research institutes, SHOK is a co-operation platform for refining expertise. Establishment of the strategic centres is based on a policy set by the Science and Technology Policy Council of Finland (currently known as the Research and Innovation Council) in 2006. The goal of SHOK is to offer top research units and companies using research results a new way of their working in close, long-term co-operation. The strategic centres are application-based and promote multidisciplinary approaches. Tekes – the Finnish Funding Agency for Technology and Innovation is primarily responsible for the public funding of SHOK.

Spatial data infrastructure

As part of the social information infrastructure, spatial data infrastructure involves jointly produced spatial data and spatial data services, their description and technical execution, and the principles and processes concerning the availability and use of data.

Statistics authorities

Statistics authorities are entitled to collect data for statistical purposes based on the obligation to provide data as specified for in the Statistics Act. Statistics authorisation bodies are the Ministry of Agriculture and Forestry Information Centre (Tike), National Institute for Health and Welfare (THL), Statistics Finland and Finnish Customs. In addition to statistics authorities, the state statistics function also includes other state agencies, institutes, and organizations producing statistics, such as the Finnish Civil Aviation Authority, Finnish Meteorological Institute, Agrifood Research Finland - MTT, National Land Survey of Finland, Finnish Forest Research Institute (Metla), Game and Fisheries Research, Finnish Environment Institute SYKE and the Ministry of Employment and the Economy.

Statistical data

Unit level data (e.g. personal level or company level data) acquired for keeping statistics or aggregated data based on the statistics compiled. When using this term, it should be specified whether the data in question is unit level data or aggregated data.

Statistical data protection methods

General term for methods whose goal is to protect the privacy of statistical unit data to be published or submitted for research use. (See statistical unit.)

(National) Statistical Service

Statistics authorities and other authorities producing statistics at a national level. These are responsible for the production of Official Statistics of Finland (OSF) and statistics belonging to the European Statistical System (ESS). The purpose of the National Statistical Service is to serve society's general need for data by producing statistical data describing social conditions and their trends and making it available

for public consumption. The basis of the National Statistical Service is the Statistics Act.

Statistical unit

Basic unit for analysis, i.e. a natural person, household, financial actor or enterprise, to which the data refers.

Sui generis protection

In data, a sui generis database right is considered to be a property right, comparable to but distinct from copyright, that exists to recognize the investment that is made in compiling a database, even when this does not involve the 'creative' aspect that is reflected by copyright.