FAIR handbook for the data steward community in the Czech Republic

Editor-in-Chief:

- Karolina Podloucká (NTK) https://orcid.org/0000-0002-5623-2949

Editors:

- Tereza Kolmačková (CAS Library) https://orcid.org/0000-0003-0829-0770
- Věra Franková (CU) https://orcid.org/0000-0002-1927-9596
- Anna Tůmová (CU) https://orcid.org/0000-0002-6716-0987
- Matěj Uchytil (CU)

Authors:

- Karolina Podloucká (NTK) https://orcid.org/0000-0002-5623-2949
- Adéla Jílková (NTK) https://orcid.org/0000-0002-3158-9626
- Jan Skůpa (BUT) https://orcid.org/0000-0001-8033-9634
- Jindřich Fejfar (CAS Library) https://orcid.org/0000-0002-2048-3682
- Eliška Pospíchalová (CAS Library) https://orcid.org/0009-0008-6092-7146
- Eliška Blažková (CAS Library) https://orcid.org/0009-0004-5454-9514
- Matyáš Hiřman (CU) https://orcid.org/0000-0003-0153-9414
- Martin Schätz (NTK/UCT Prague) https://orcid.org/0000-0003-0931-4017

This handbook was created based on the materials used in the FAIR data handbook. We would like to thank the authors for their work so far.

Editors:

- Eva Hnátková (NTK/UCT Prague) https://orcid.org/0000-0002-3237-9305
- Zdenka Dudová https://orcid.org/0000-0002-7615-1396
- Dagmar Hanzlíková (CU) https://orcid.org/0000-0001-9287-399X

Authors:

- Eva Hnátková (NTK/UCT Prague) https://orcid.org/0000-0002-3237-9305
- Zdenka Dudová https://orcid.org/0000-0002-7615-1396
- Tereza Šímová (CAS Institute of Philosophy/CZU) https://orcid.org/0000-0002-1774-4335
- David Šlosar (CAS Library/CU Institute of Information Science and Librarianship) https://orcid.org/0000-0001-5168-5327
- Veronika Zemanová https://orcid.org/0000-0002-2544-6872
- Jiří Marek https://orcid.org/0000-0003-2132-762X
- Tereza Šírová (CAS ISC) https://orcid.org/0000-0001-5927-1065
- Daniela Tršová https://orcid.org/0000-0002-7019-8310
- Kristýna Zychová (CZU) https://orcid.org/0000-0002-2021-7894

DOI for Czech language version: https://doi.org/10.71495/hxfc-6f57
DOI for English language version: https://doi.org/10.71495/ejrs-8t41

Tato práce vznikla za podpory projektu Národní repozitářová platforma pro výzkumná data, reg. č.: CZ.02.01.01/00/23_014/0008787 financovaného z EFRR.

This handbook was created with the support of the project "National Repository Platform for Research Data", reg. no. CZ.02.01.01/00/23_014/0008787, funded by the ERDF.

Contents

Introduction

This document is intended for data stewards who have a basic understanding of and knowledge in this area. It provides an overview and description of each element of the FAIR Principles (F = Findability, A = Accessibility, I = Interoperability, and R = Reusability), which were first defined and published in 2016 and are now an integral part of the research data cycle. More and more emphasis is placed on these elements.

This handbook is based on the FAIR Data Handbook. It develops these materials and enriches them with practical components.

The aim of the document is to familiarize data stewards with the definition of the FAIR Principles. In addition, it provides practical examples in the individual sections so that they may better understand the topic. Beyond the guiding principles, the handbook offers additional sections that the authors consider essential to illustrate the issue as a whole. Thus, more technical topics, such as data anonymization, semantic interoperability, or machine actionability, are also addressed. These topics are essential in interdisciplinary collaboration or even for data stewards working in an institution with a broad research scope.

The FAIR Principles in the context of open science and research data management

Research data and their management

Research data are usually of great value because it takes a lot of time, effort, money, and other resources to obtain and the research data results can have a big impact. Good research data management (RDM) is important throughout the entire research process and involves various procedures and strategies. The steps of the research data management life cycle include data planning, organization, documentation, storage, and archiving. These steps ensure that data are used effectively and are adequately protected against any loss or misuse. RDM is also the key to data reuse and sharing in research teams as well as at the national and international level. There are a number of domain-specific or institution-specific guidelines and instructional materials relating to RDM. The new position of data steward has been created in many projects and at many institutions to provide technical assistance, coordination, and oversight with respect to data management. The FAIR Principles have been formulated as general guidelines for data management[1] and ensure that data are findable, accessible, interoperable, and reusable.

Open research data

Open science (OS) is a concept in the research process that promotes open access to research results using new digital technologies and tools.[2, 3] The main objective of this concept is to improve the accessibility and reuse of research results, to conserve financial resources spent on science and research, to encourage new collaboration, and to ensure research transparency. Thus, open science enables the successful replication of results, leads to greater confidence in research, facilitates the evaluation of outputs relating to scientific debate, and increases the impact of researchers and research institutions. All of this contributes to accelerating and streamlining research and improving research quality.[2, 4, 5]

Open science covers several areas, including open access to research publications ("open access") and research data ("open data").[6–9] Open data are generally data that are freely accessible and that may be used, modified, and shared by anyone for any purpose.[10] However, there may sometimes be legitimate situations where access to data must be restricted and data kept inaccessible for justifiable reasons, e.g. to protect intellectual property, personal data and privacy, human rights, security, legitimate commercial interest, etc.

For data generated by publicly funded research, the European Commission recommends making the data accessible in accordance with the FAIR Principles and the principle "as open as possible, as closed as necessary".[11] Open access to data according to the FAIR Principles is increasingly required to varying degrees by funding providers (e.g. the Technology Agency of the Czech Republic, the Czech Science Foundation, etc.) and is being integrated into the policies of universities and research institutions (e.g. Charles University Research Data Policy). When publishing research articles, publishers often require disclosure of the data on which the research results are based and information on data accessibility (e.g. PLoS, Springer Nature, Wiley).

As a part of the National Research, Development and Innovation Policy of the Czech Republic 2021+, the Czech Republic strives to ensure open access to the results of research and development in accordance with European legislation. In addition, the management of and access to research data in the Czech Republic is governed by Act no. 130/2002 Sb., on support for research and development.

The FAIR Principles

The FAIR Principles were published in 2016 in an article by Wilkinson et al. 2016, in which the four guiding principles are formulated, stating that data should be: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. The principles are further broken down into a total of fifteen elements that describe in more detail what characteristics current data, metadata, and the used tools and infrastructures should exhibit in order to make the data and/or metadata as findable, accessible and reusable as possible.[1, 12] The FAIR Principles themselves are not a standard or norm and do not mandate the use of specific tools or technologies, as these can change over time and depending on the domain.[1] However, the FAIR Principles are increasingly becoming a part of the requirements of funding providers or institutional policies to ensure good research practices.[13]

The FAIR Principles are very general and are not domain-specific so that they can be applied to a wide range of research results. They are relevant for both data and metadata. The different FAIR elements are interrelated but at the same time independent and separable. This modularity allows them to be used in a variety of circumstances, e.g. providing only metadata when working with sensitive data.[1] FAIR is described as a spectrum on which varying degrees of "fairness" (derived from the word FAIRness in the FAIR Principles) can be achieved and gradually improved.[14]

None of the FAIR Principles require data to be open or free of charge. However, clear and transparent conditions for access to and reuse of data are required. Hence, FAIR data need not be open, but should have a clearly assigned licence.[1]

Data should adhere to the FAIR Principles in terms of both human- and machine-driven activities. Particularly important is machine actionability, where computing systems are capable of recognizing the type of data and its usefulness, evaluating the conditions of use according to a licence, and then processing the data without human assistance or with minimal human intervention. Humans are not capable of working at the scale and speed required by the volume and complexity of current research projects and data, and so this area relies on the action of computing systems.[1]

It can be useful for a wide variety of target groups to follow the FAIR Principles. These groups include, for example, researchers who want to share their data or reuse the data of others, professional data creators, software developers, funding providers, or the research community, including data stewards.

The process of creating FAIR data is referred to as "FAIRification", and there are various online tutorials for preparing the workflow, for example:

- FAIRToolkit (Pistoia Alliance)
- FAIRification framework (FAIR Cookbook, ELIXIR, FAIRplus)
- FAIRification Process (GO FAIR)

Tools are also available to assess the extent to which research data are FAIR, for example:

- F-UJI
- FAIR DataSet Maturity
- FAIR Data Self Assessment Tool
- FAIR checker

FAIR vs. open data vs. data management

FAIR, open data, and research data management are three different concepts. However, they overlap with each other. Each of them focus on something different, and the best results can

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

be achieved when they are used together. It is important to start with good research data management in the early stages of research (even better, during the research preparation period), otherwise, it could be more difficult to achieve open and FAIR data in the later stages. The principles of open and FAIR data can also help engage researchers in data management with the motivation that more findable data are more visible and can improve the impact of research results.

FAIR and open data are not the same thing. Data can be FAIR and open at the same time, FAIR only, open only, or neither. Both FAIR and open data principles focus on sharing data. However, FAIR data are not necessarily open. The ideal situation is to make the data as FAIR and open as possible. If open data cannot be provided in certain situations for legitimate reasons, it may be possible to at least meet the FAIR requirements. Neither FAIR nor open data say anything about the quality of the data.[15]

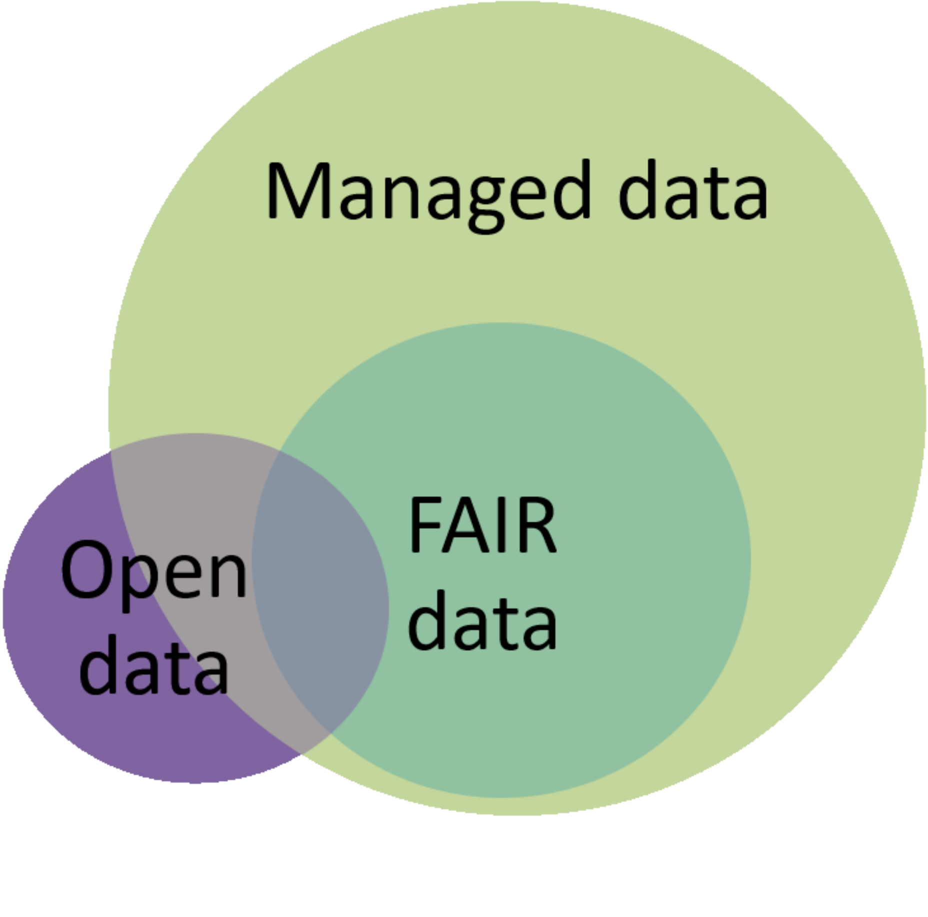


Figure 1: Overlapping of three different data concepts.[72]

Findability

The main prerequisite for making data accessible and reusable is data retrieval, which will be the focus of this chapter. The importance of findability can be demonstrated by the following example. A research article was written. This takes quite a lot of time, effort, and work. Although the article may be of excellent quality, if there is insufficient information about it, people will not know about it and it will not reach the research community, which may result in low citation rates and insufficient knowledge of the authors and the prestige of the institution. It may also slow down future research since the results published in the article cannot be followed up on. However, if the article is stored in a suitable place (a repository), is sufficiently described (metadata) and assigned a persistent identifier (making it easily findable), the authors will contribute to the awareness of their work and others will be able to easily find, disseminate, or cite the article thanks to the identifier. A similar approach should be followed with research data.

F1. (Meta)data are assigned a globally unique and persistent identifier

The **persistent identifier (PID)** is a tool used to uniquely identify people, organizations, and other objects (e.g. books, articles, datasets) in a research communication system. The unique links associated with an entity and/or its metadata do not change over time and allow for persistent retrieval, access, citation, linking, and reuse of research results. These are globally unique, persistent, and machine-readable digital identifiers using a metadata schema.[16]

In this context, most of the identifiers consist of a web link, and clicking on them in a browser will allow access to a web page (landing page) with additional information or directly with the data itself. Identifiers are essential for the human-machine interoperation that is key to the vision of open science. In addition, identifiers will help others properly cite our work when reusing our data. Some digital objects may already have a PID assigned when they reach us. Thus, it is advisable to use it consistently and not to assign another identifier.[17]

Functional requirements for persistent identifiers

- Globally unique;
- Globally readable as a URI (uniform resource identifier) in line with http (e.g. https://doi.org/10.48813/k2xs-y923);
- Persistent – intended to work regardless of the lifetime of the systems or organizations;
- Managed by a dedicated organization with defined governance and decision-making mechanisms;
- Linkable/interoperable with other identifiers through metadata elements that describe their relationships;
- Must contain metadata that describe their most important properties;

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

- Indexable and searchable by its metadata elements along with all other trusted identifiers.[16]

In academic environments, a variety of persistent identifiers are used to identify objects (e.g. authors, digital objects, publications and journals, datasets, organizations, etc.). The table below provides examples of the most common types of unique identifiers. Some identifiers are still in the development phase, e.g. RAiD (research activity), ePIC (research data prior to publication), ConfIDent (professional conferences).

Type of identifier	**Type of use**	**Example**
ORCID iD	Person (author, researcher)	https://orcid.org/0000-0001-8888-635X
ResearcherID	Authors of a publication that is indexed in the Web of Science (WoS) database have an author record automatically created and an assigned ResearcherID.	B-6035-2012
Scopus Autor ID	Identifier and author profile belonging to the Scopus bibliographic and citation database operated by Elsevier publishing company.	6603082428
DOI	Digital objects (articles, datasets, DMPs (Data Management Plans), audiovisual recordings, conference papers, preprints, software, standards, etc.)	https://doi.org/10.1038/sdata.2016.18
Handle	Digital objects	https://hdl.handle.net/20.500.14391/1549
ROR	Research organizations	https://ror.org/028txef36
IGSN	Research samples	10.60510/ICDP5054ESYI201
ISSN	Journals	1214-8790

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generála Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

ISBN	Books	978-80-247-2279-5
ISMN	Musical works	979-0-2600-0043-8

Table 1: Examples of persistent identifiers

Examples of globally unique and persistent identifiers:

- One particular person on planet earth has this globally unique and persistent identifier: https://orcid.org/0000-0001-8888-635X.
- An identifier that unambiguously refers to an article about the "FAIRness" of the FAIR Principles: https://doi.org/10.2218/ijdc.v12i2.567.
- The human polycystin-1 protein has a globally unique and persistent identifier given by the UniProt database: http://purl.uniprot.org/uniprot/P98161.
- Polycystic kidney disease Type 1 has a globally unique and persistent identifier given by the OMIM database: http://omim.org/entry/173900.[17]

PID graph – Not only are PIDs important for uniquely identifying a publication, dataset, or person, but the metadata of these persistent identifiers can provide unambiguous links between persistent identifiers of the same type, e.g. journal articles citing other journal articles, or different types, e.g. linking a researcher to datasets they created[18] – illustrated in the graph below.

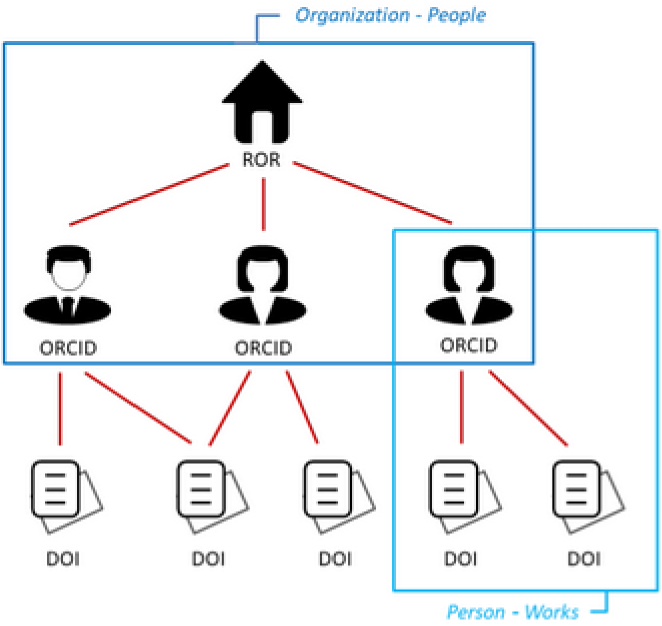


Figure 2: The linking of different types of persistent identifiers to facilitate academic communication[19]

Support for persistent identifiers in the Czech Republic:

- Czech National ISBN and ISMN Agency, operated by the National Library of the Czech Republic
- National Centre for Persistent Identifiers, operated by the Czech National Library of Technology, which includes the following centres:
  a) National ISSN Centre
  b) National ORCID Centre, which was established in 2023. Member institutions can work with the verified ORCID iDs of researchers in their systems and record information on affiliation, publication activity, etc. on the ORCID profile of their researchers.
  c) National DOI Centre, which was established in 2023. Member institutions can register DOIs for various types of objects, including IGSNs for samples.[16, 20]

The following website can also help you understand persistent identifiers more easily: https://identifikatory.cz. Here you can find up-to-date information on PIDs and the types of support available at the national level.

F2. Data are described with rich metadata

The probability of finding research data is greatly enhanced by rich metadata descriptions, such as descriptive information about the context, quality, status, or characteristics of the data (described in more detail in section R1). Rich metadata allow computers to automatically perform searches that are very time-consuming for humans. The premise of the F2 principle is that a user should be able to find data based on the information provided in the metadata, even if the data do not have a persistent identifier assigned. Particularly in research activities, a good metadata description increases the chance of discovery and additional use of the data.[17]

F3. Metadata clearly and explicitly include the identifier of the data they describe

The dataset and descriptive metadata are usually separate files. As the name of this principle suggests, it is important that the two are linked. That is, the globally unique and persistent identifier of the dataset should also be included in the metadata. In cases where the data are no longer (or can no longer be) available, it is recommended to at least mention that the data exist. Many repositories today automatically generate PIDs when data are deposited, and these can be used for this purpose.[17]

F4. (Meta)data are registered or indexed in a searchable resource

Identifiers and rich metadata alone will not make data “findable” on the Internet without further action. Perfectly good data resources may go unused simply because no one knows they exist. If a dataset is not available, then nobody (not even a machine) can discover it. Thus,

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

the data and metadata need to be deposited into a suitable and trusted (data) repository that is accessible, searchable, and indexed. Depending on the need, either an institutional, disciplinary (domain), or multi-domain (also generic, general, or orphan) repository can be used.[16, 17] The following registry may be consulted to select a suitable repository: Registry of Research Data Repositories. In this web-based tool, you can search by title, domain/research discipline, country in which the repository is hosted, or by content type (e.g. datasets, text files, structured text, images, configuration data, etc.). Another tool that can help you find a suitable repository is FAIRsharing, where you can search in the Databases section. Recommendations from certain publishers and journals can also help authors choose the appropriate repository, e.g. https://www.nature.com/sdata/policies/repositories.

Accessibility

This principle ensures the accessibility of data and metadata. From a practical standpoint, not all data can be openly accessible (due to personal or sensitive data, national security, commercialization, etc.), so it is recommended (and often specified in project terms and conditions) to follow the "as open as possible, as closed as necessary" rule.

It is now quite common practice for repositories to allow a distinction to be made between making data accessible to anyone or making the data accessible only to a limited range of users. If it is not possible to make the actual data accessible, the metadata should at least be accessible in the repository without restrictions, because even information about the existence of data with restrictions is very valuable. All of this should be done using standard communication protocols.

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol

The Identifiers used for metadata are mainly DOI, URN, PURL, HDL (Handle). The assignment of DOIs is provided, for example, by repositories that register DOIs based on the creation of a record or the storage of a dataset. The identifier ideally refers to a landing page. The standardized communication protocol here is primarily a Hypertext Transfer Protocol – http(s).[21, 22]

Name	Value	Link
DOI	10.5281/zenodo.4420115	https://doi.org/10.5281/zenodo.4420115
Handle	1813/7895	https://hdl.handle.net/1813/7895
URN:NBN	urn:nbn:cz:tst02-000008	https://resolver.nkp.cz/urn:nbn:cz:tst02-000008h

Table 2: Examples of persistent identifiers

A1.1 The protocol is open, free, and universally implementable

For access to (meta)data, access must be provided using an open protocol without financial or implementation constraints. In the vast majority of cases, this means using an HTTP or HTTPS protocol, but it is also possible to use API, FTP, FTPS, SFTP, or SCP protocols and others. It should be added that the use of HTTP and FTP protocols is not recommended today, because all communication takes place over these protocols in an unencrypted form. In the case of access through, for example, APIs, it is important to provide the necessary specification of possible operations, documentation and, in the case of sensitive data with restricted access, authentication of the accessing party, which will ensure the possibility of providing automated access to data or their integration into other services for authorized entities.[23]

Example of API documentation: https://docs.ckan.org/en/2.10/api/index.html

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

Not all data can be accessible to everyone without authentication. Thus, secure authentication and authorization must be ensured when accessing research data. Access options should be clearly stated in the metadata or in the ReadMe file. This should include information on who can use the data, any restrictions on use, and how the data can be accessed and under what conditions. User authentication and authorization is usually managed by the repositories themselves. When programming your own systems, it is advisable to use standardized protocols, such as OAuth, SAML.[24]

In addition to direct user authentication using a password, federated identities can be used (e.g. eduID.cz, mojeID, eduGAIN). When working with APIs, users can be authorized using API keys or OAuth tokens. Organizations often use Role-Based Access Control (RBAC) where users are granted permission based on their role in the system.

A2. Metadata should be accessible, even when the data are no longer available

Not all data can be available to everyone, and not all data can be available for an unlimited period of time. In such cases, at least the metadata should remain accessible, thus providing evidence of e.g. the experiment being measured. It is important (even necessary) to store and publish metadata in standard formats such as Dublin Core or RDF, which have an accessible online specification. Metadata should always be accessible via the persistent identifier type DOI or Handle.[25]

Other aspects relating to data accessibility

Long-term accessibility of research data

Research data may be relevant many years after being created. Thus, it is essential to ensure that the data are protected and managed so that they are usable in the future. This issue is not only important from the standpoint of scientific knowledge, but also for research efficiency and repeatability. Proper storage of data enables its reuse, minimizes the need for expensive repeated experiments, and increases the credibility of the research results.

In addition to the correct choice of format, which is discussed in more detail in the following chapter Interoperability, there are other measures that can help ensure the long-term accessibility of data. Repositories, which can provide specialized services to protect and preserve data over time, play an important role. Thus, when selecting a repository, it is crucial to check whether it offers a guarantee of long-term accessibility, what data protection strategies it has in place, and whether it meets international standards for digital archiving. Suitable repositories often have certifications, such as CoreTrustSeal, which confirm their ability to store data in a reliable and sustainable manner.

Elements that guarantee or promote long-term accessibility:

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generála Piky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

Data bit integrity

One of the key elements to ensuring the long-term accessibility of data is an integrity check, which is performed using bit integrity. This process involves calculating checksums such as MD5, SHA256 or the more modern SHA3 algorithms to verify that data have not been tampered with or corrupted during transmission, storage, or archiving.

A checksum is created when a file is uploaded to a repository and can be compared against the current version of the file in the future to ensure that the data have not been altered without be detected. Should any integrity breach occur, the repository system should allow data recovery from redundant backups. Many repositories, such as Zenodo or Dataverse, actively implement mechanisms to detect and correct errors in files, minimizing the risk of loss or irreversible damage to research data.

In addition to checksums, it is also advisable to use techniques such as **data integrity auditing**, which consists of periodically recalculating the hash values of files and comparing them with the original values. Some advanced archiving systems even allow for **automatic data correction** if a deviation from the original file is found, which greatly contributes to protecting information from gradual degradation.[26]

Migration strategy

Digital technology is constantly evolving, and some file formats can become outdated over time, which can lead to problems with readability and usability. Thus, it is important to have a **migration strategy** in place to ensure that data are regularly converted to up-to-date and sustainable formats. This process is particularly crucial for specific fields of science that use proprietary or less common file formats that may not be supported in the future.

Migration strategies usually include several primary measures. One of them is to **monitor the evolution of software standards** and to regularly check the compatibility of formats with new versions of programs. If a format is found to be at risk of obsolescence, repositories can automatically convert files to newer open formats, such as **CSV instead of proprietary spreadsheet formats**, or **XML and JSON instead of specific binary data formats**.

It is also important to ensure that the migration of data does not have a negative impact on the correctness of its content. Any data conversion should be accompanied by detailed documentation that describes the changes made and preserves information about the original format. In this respect, it is worthwhile to engage with repositories that have clearly defined and transparently communicated migration strategies, such as the **DataverseNO Preservation Policy**, which sets out precise procedures for regular data conversions to current formats.[27]

Geographic redundancy

No matter how reliable a data infrastructure may seem, there is always a risk that stored data may be irretrievably lost due to technical failures, natural disasters, or cyber-attacks. Hence, one of the most effective measures to ensure the long-term accessibility of data is **geographic redundancy**, i.e. storing copies of data in various geographical locations.

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generála Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

With this strategy, data are not stored in just one repository or data centre, but distributed among multiple independent repositories located in different regions. If one repository breaks down, data can be recovered from another location, making the data **more resilient to disasters**, such as fires, earthquakes, or floods.

Interoperability

The **interoperability of (meta)data** is simply the ability to work and interact with other data, applications, and operating systems for the purpose of processing, analyses, comparison, and storage. This involves the following requirements:

- Data are provided in commonly used and ideally open formats;
- The provided metadata are regulated by the relevant standards;
- Where possible, controlled vocabularies, thesauri, and ontologies should be used for description;
- References and links are provided to other related data.[28–30]

Examples of good practice in data management include considering the interoperability aspect of the data (controlled lists, syntax or format issues) and the metadata of the digital file (schemas, variable names, metadata tools).

A commonly agreed approach to providing European public services in an interoperable manner is defined in The European Interoperability Framework (EIF). The basic levels of interoperability are:

1. Legal

Legal interoperability refers to a shared interpretation and understanding of the laws regulating the exchange of information and collaboration (e.g. whether it is permissible to share information about research participants, how to approach privacy, or anonymization).[31, 32]

2. Organizational

In the EIF, organizational interoperability is defined as the manner in which organizations align their business processes, responsibilities, and expectations to achieve commonly agreed and mutually beneficial goals. In view of the overall agreed objective of open science, according to the document EOSC Interoperability Framework, organizational interoperability should focus on documenting, integrating, or aligning the processes of various service provider organizations to ensure that researchers can achieve their open science goals. It should also be clear who is responsible for the provision (and the development, maintenance, and management) of common interoperability services, such as, but not limited to, service catalogues, registries, and common PID services.[32–35]

3. Semantic

The purpose of semantic interoperability, according to the European Commission, is to ensure that the exact format and meaning of the data and information exchanged between parties is preserved and understood. According to the EIF, semantic interoperability includes both semantic and syntactic aspects.

Co-funded by the European Union
MINISTRY OF EDUCATION, YOUTH AND SPORTS
National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generala Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

The semantic aspect is the common underlying models and the codification of data, including the use of data elements with standardized definitions from publicly available value sets and coding vocabularies (terminology standard), which promotes shared understanding and meaning. In most cases, these vocabularies are internationally agreed, thus ensuring the unambiguous interpretation of information.

The syntactic aspect refers to the description of the exact format of the information to be provided in terms of grammar and format.[32]

4. Technical

Technical interoperability should in practice make it possible for systems or applications to receive (meta)data from each other and from other entities. In addition, it can be understood as the execution of specific tasks in a suitable and satisfactory manner without the need for additional data processing requirements or "operator" intervention in the context of the full automation of data exchange activities.

The aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and processing, and secure communication protocols. The main obstacles to interoperability are primarily legacy systems unable to interconnect with newer systems or the general compatibility of devices with each other.[32]

Practical examples: software standards, physical hardware components, and systems and various platforms supporting computer-to-computer interaction.[36]

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

Computers should be able to exchange and interpret data. Data should generally be machine-readable without the need for specialized (proprietary) or ad hoc algorithms, compilers, or mappings. In this context, interoperability means that each computer system should have at least a minimum knowledge of the data formats of the other system. The structure of data that is considered FAIR must be documented in both a human-readable manner and a machine-readable schema.

The optimal form of standardization is to define a data format independently of the specific software with a detailed description of the format structure and a precise explanation of its parts. It is advisable to supplement the definition with a data validation tool that allows you to determine whether the produced data conform to the format, e.g. XML validator for XML. Formats approved by a recognized international standards organization are the highest form of standardization, e.g. a standard for the format CSV, ISO, OASIS – Open Document, etc.[37–39] Such standardized data formats are suitable for data exchange in terms of long-term readability and interpretability.

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generála Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

Extensively used data formats with support implemented in a wide range of software from different manufacturers have a higher probability of **good interoperability and long-term readability**. The openness and good definition (standardization) of data formats has a positive impact on the expansion of data formats and their implementation in software from different manufacturers.[40]

Examples:

Suitable open formats:

- Tabular data – CSV, TSV, JSON, RDF (JSON-LD, Turtle, …);
- Hierarchical data – XML, JSON, RDF (JSON-LD, Turtle, …);
- Graph and linked data – RDF (JSON-LD, Turtle, …);
- Geodata (spatial data) – GeoJSON, ESRI Shapefile, OGC GML, OGC GeoPackage.

Recommended languages for defining schemas:

- CSV – CSV schema, CSV on the Web;
- XML – XML schema;
- JSON – JSON schema;
- RDF – RDFS, OWL, SHACL.[41–48]

Some organizations list recommended formats on their websites, e.g. DANS or UK Data Service.

I2. (Meta)data use vocabularies that follow the FAIR Principles

Controlled vocabularies used to describe datasets must be validated and defined using globally unique and persistent identifiers for the individual terms and ontologies. Everything must be easy to find and accessible to anyone using the datasets.

Examples of vocabularies "Knowledge Organization Systems Types": AGROVOC, GeoNames

Examples of metadata vocabularies: Dublin Core, W3C Data Catalog Vocabulary, Multi-Crop Passport Descriptors

How to find the most useful available vocabularies

1. FAIRsharing – https://fairsharing.org
   Registry of terminology artifacts, models/formats, reporting guidelines, and identifier schemas.

2. Linked Open Vocabularies (LOV) – https://lov.okfn.org/dataset/lov
3. The Basel Register of Thesauri, Ontologies and Classifications (BARTOC) – http://bartoc.org [49–51]

I3. (Meta)data include qualified references to other (meta)data

A qualified reference is a cross-reference that explains its intent. The goal is to create as many meaningful links as possible between (meta)data resources to enrich contextual knowledge about the data, balanced against the time/energy involved in making a good data model.

Specifically, you should specify if one dataset builds on another dataset, if additional datasets are needed to complete the data, or if the complementary information is stored in a different dataset. In particular, the links between the datasets need to be described. In addition, all datasets need to be properly cited (i.e. including their globally unique and persistent identifiers). Last but not least, the PIDs themselves can be referenced to each other during their creation as a part of their metadata descriptions. These linkages between datasets give rise, among other things, to linked data, considered as extended data interoperability. See the linked data vs. FAIR data relationship described on the website The Road to FAIR.[52]

The linked data paradigm postulates four rules according to The Road to FAIR:

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL). For example, the SPARQL query language can also be practically used via wikidata.

4. Include links to other URIs, so that they can discover more things.[52]

The linked data paradigm according to **Jonathan Blaney, "**Introduction to the Principles of Linked Open Data (LOD)**"**:

1. Use a recognized LOD standard format. In order for LOD to work, the data must be structured using recognized standards so that computers interrogating the data can process it consistently. There are a number of LOD formats, some of which are discussed below.
2. Refer to an entity the same way other people do. If you have data about the same person/place/thing in two or more places, make sure you refer to the person/place/thing the same way in all instances.
3. Publish your data openly. By openly I mean for anyone to use without paying a fee and in a format that does not require proprietary software.[53]

In 2010, **Tim Berners-Lee** published a system for evaluating the openness of linked data: 5-Star Linked Open Data. He believes that it is good to publish data openly using open formats and public standards. Data acquire information value when linked to other data.[54]

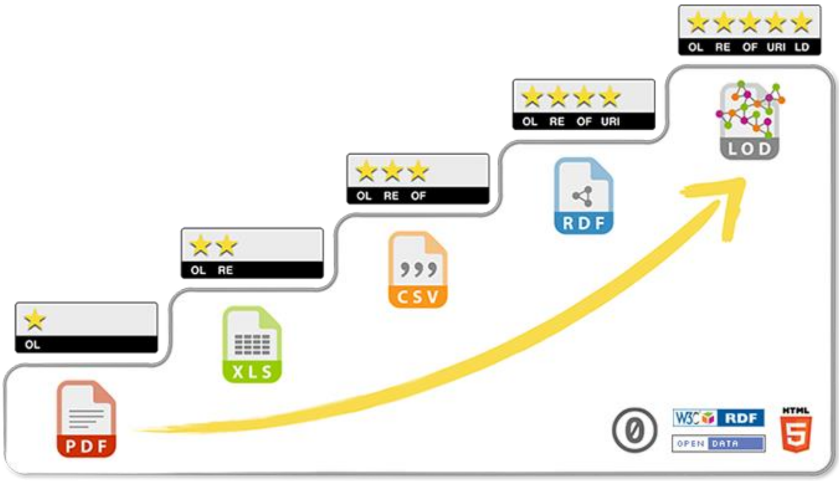


Figure 3: System for evaluating linked data according to Tim Berners-Lee.[54]

Evaluation system: 1–5 stars (5 being the best)

★ Make your stuff available on the Web (whatever format) under an open licence.

- See How to set the terms of use for datasets

★★ Make it available as structured data (e.g. Excel instead of an image scan of a table).

- The dataset is provided in a machine-readable format that enables automated machine processing.

★★★ Make it available in a non-proprietary open format (e.g. CSV instead of Excel).

★★★★ Use URIs to denote things, so that people can point at your stuff.

★★★★★ Link your data to other data to provide context.

- The dataset meets the linked data standard.[55–57]

Reusability

The aim of opening any content, especially digital objects, is not simply to make the object accessible, but rather to reuse it. This involves the reuse of existing resources to prepare, develop, or create something new that has economic and other benefits. In order for this to happen, there are legal, technical, and managerial (curatorial) recommendations.

Following the FAIR Principles, these include:

1. An appropriate (as open as possible) licence for the dataset, if necessary, or an indication that no rights are attached to the dataset (see section *R1.1 (Meta)data are released with a clear and accessible data usage licence*).[58]
2. Proposals to address potential obstacles to (meta)data publication (*Other legal obstacles to opening (meta)data*).[58]
3. The origin (provenance) of the dataset and its modifications (*R1.2 (Meta)data are associated with detailed provenance.*[59]
4. Use of generally known domain standards in terms of the technical and methodological aspects for the metadata description of the dataset (*R1.3 (Meta)data meet domain-relevant community standards.*[60]

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

To make data easier to find and reuse, it is crucial that rich metadata describing their individual attributes be linked to the data. The most accurate description of a dataset contributes to increasing the ability of the user (machine or human) to decide whether or not the data are really useful in a specific context. A detailed and rich description of (meta)data leads to the possibility of being automatically linked or integrated (with minimal human effort) to the relevant data resources.[61]

Published (meta)data should reference their resources with sufficiently detailed metadata and information on provenance to allow for correct citation.

It is beneficial to provide not only metadata that enhance the findability of the dataset (Principle F), but also metadata that richly describe the context in which the data were generated. This means supporting documentation for the research process. The creator of the data should not attempt to anticipate the needs of the data consumer/user.

In general, the (meta)data author should be as generous as possible in providing metadata, including information that may seem irrelevant.

The following topics, for example, should be dealt with as a part of this principle:

- During what period/in what location did the data collection take place;

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generála Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

- Information about tools or software versions that were used for collecting or analysing the data;
- Demographic data on respondents, the manner of addressing the respondents;
- A list of variables/field names in the table and their description (e.g. the values they can take);
- Explanations of abbreviations or code names;
- A laboratory diary;
- The questionnaire form;
- A template for informed consent;
- Licensing terms for using the collected data;[62]
- And much more information.

R1.1 (Meta)data are released with a clear and accessible data usage licence

Datasets that are copyrighted works in accordance with Act no. 121/2000 Sb., the Copyright Act (the “Copyright Act”) are automatically protected under copyright law. This means that a copyright protecting an author’s intellectual property is a certain **legal obstacle to the reuse of data** by other persons. **Public licences are the main tool for removing these legal obstacles.** The best-known and most widely used set of public licences is Creative Commons licences. The person authorized to grant the licence (typically the author, or an employer in the case of an employee work) may attach the selected public licence to the dataset, thereby granting permission to use the dataset to an unspecified and unlimited group of users. A contract is then concluded when the user begins to use the data in accordance with the terms of the attached licence.

Each Creative Commons licence includes some of the **four licensing elements** (the BY element is always present), indicating the conditions that must be observed when using the licensed dataset:

- **BY** – Must include the name of the author, the name of the work, the licence, and the source;
- **SA** – Identical licensing terms must be used;
- **NC** – Must be used for non-commercial purposes only;
- **ND** – No derivatives or adaptations of the work are permitted.[63]

Use of the least restrictive public licences is justified in the sense that the research data output of one experiment is the input data for another experiment. With this in mind, it is **not** appropriate to use licences containing the NC and ND elements, since these two elements prevent the effective reuse of research data.

On the other hand, it would be beneficial to use the following:

- To the greatest extent possible, the least restrictive licence Creative Commons Attribution 4.0 International (CC BY 4.0).

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generala Piky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

- For a database protected by special rights of the database creator, Creative Commons Zero Universal Dedication (CC0)[1]. The rights of the database creator is an exception where this licence can be used.
- If the dataset is **not** a copyrighted work or a database under the Copyright Act (it involves simple data, typically the results of instrument measurements), it is appropriate to include information in the metadata that the dataset is not subject to copyright protection and hence is free to use.

This tool can be used as a guide for choosing an appropriate Creative Commons licence for your own output: https://chooser-beta.creativecommons.org/

Other legal obstacles to opening (meta)data

The practical aspects of opening or “FAIRifying” research data are part of the life cycle of data and need to be considered during the creation, processing, and subsequent publication of datasets. Before making research data available for reuse, it is important to bear in mind that, in addition to copyright protection (see Section R1.1), the data may fall under one of the other **protection regimes** provided for under law.

This involves in particular cases where:

1. **Research data contain personal data.** Disclosure of personal data in the form of open data is not permitted. However, if the data are anonymized during processing, properly anonymized data can be made accessible, since they no longer fall under the GDPR protection regime; see the section Personal data issues.
2. **Research data contain other data protected under law.** Disclosure of such research data (e.g. trade secrets, classified information) is not permitted. Their non-disclosure is entirely legitimate in accordance with the principle of “as open as possible, as closed as necessary”.

If a dataset is not intended to be made accessible for any reason or is to be made only partially accessible, the reason for not disclosing the dataset or the conditions for increasing the “fairness” of the data should be stated in the metadata describing the dataset.

Personal data issues

The presence of personal data in a dataset (i.e. any data that can lead to the identification of a private individual) can be an obstacle to its disclosure. Personal data must always be handled in accordance with the content of the consent to the processing of personal data granted by

[1] This is a special type of the Creative Commons licence that provides a “waiver”. Thus, a work with a CC0 licence is available to anyone for free use.

the individual research participant. The requirements of consent are set out in EU Regulation no. 2016/679 on the protection of personal data (the GDPR). In addition, the consent usually regulates in what form (i.e. identifiable, anonymous, or pseudonymized – see below), for what purpose and with whom the data can be shared, and for how long the data can be stored. Hence, if a dataset with identifiable data cannot be disclosed based on consent, this obstacle can be addressed by anonymizing the data.

Data that allow the identification of a private individual are referred to as identifiers. Identifiers can be divided into two groups: direct and indirect.[64]

Direct identifiers

- Identifiers that allow the direct identification of a person
- For example, personal ID numbers, first names and surnames, etc.

Indirect identifiers

- Cannot be attributed to a specific person without the use of additional information
- For example, gender, age, addresses, IP addresses, types of illness, etc.

Anonymized data

- They do not allow the direct or indirect identification of a private individual, even retroactively.
- Thus, they cannot be linked to a specific person (i.e. they cannot be assigned to a specific research participant).[64]
- Hence, they are not subject to EU Regulation no. 2016/679 on the protection of personal data (the GDPR).

Data are anonymized primarily by removing **direct identifiers**. The disadvantage of anonymizing data is the loss of relevance and/or other useable features of the data. Thus, in some cases, data cannot be anonymized. The advantage is that anonymized data can be freely shared between researchers without the need to obtain the consent of the research participant to the processing of personal data (GDPR), since the data are no longer personal in nature.[66]

Data can also be anonymized for **indirect identifiers**. In this case, we use a process of generalization, in which data are summarized into general predefined categories. This process results in irreversible data loss, so it is important to carefully consider whether generalization is necessary. The greater the risk of re-identification, the higher the level of generalization applied to the entire dataset should be.[66]

For example, the date of birth for two individuals may change from A: 29. 1. 1948; B: 11. 9. 1948 to A: 01.1948; B: 09.1948, or even A: 1948; B: 1948.

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generala Piky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

Pseudonymized data

- Data that cannot be attributed to a specific person without the use of additional information.
- Subject to EU Regulation 2016/679 on the protection of personal data (the GDPR).[65]

The pseudonymization method is appropriate to use when there is a need to **maintain a connection between information/data and a particular individual** at a certain level. However, this connection will only be accessed by those who have mapping tables with direct identifiers and pseudonyms at their disposal. Pseudonymization can be used, for example, to allow a participant to be contacted again after their participation in a clinical trial has ended, for example, to supplement information.[64]

When pseudonymizing data, personal data are obscured in various ways, for example, by using a **random code** instead of a name. Thus, the identity of the research participants is only known by the GDPR data processor (usually a limited number of employees of the institution, e.g. the supervising physician, who enrolled the individual in the research), and the research team is only working with a sample designated with a random code, so the individual is always an anonymous person for the researchers. If necessary, the identity of the research participant can be "deciphered" and the personal data supplemented (e.g. with the results of other laboratory tests, diagnostic tests, etc.) based on an encryption key, which is always done by the data processor.

Unlike fully anonymized data, pseudonymized data are still subject to the **GDPR** and **Act no. 110/2019 Sb., on the processing of personal data**, because the data can be attributed to a specific private individual, and based on this, possible restrictions on data sharing must be taken into account.[65]

How to anonymize data
Data may be anonymized at any time during its life cycle, from the time of collection itself up to just before the data should be published. The method of anonymization and its scope need to be firmly established and applied to all relevant data.

Data may be anonymized in various ways:

- The data may be anonymized manually.
- A short script can be written, for example in Python, which will anonymize the dataset.
- The anonymization software AMNESIA can be used:
    - Open source software managed by OpenAIRE
    - The online demo version can handle files of up to 5,000 lines; the software itself can handle even larger files.
    - Intended for plain csv and txt datasets

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generala Píky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

- Another appropriate tool may be, for example, the software ARX.

What if data cannot be anonymized?
There are cases where data cannot be anonymized. In such a case, the following should be done:

- Address and control access to the data during the research process (legal permissions, safeguards for sharing in a collective, etc.);
- Do not share and describe this situation, for example, in the preparations for the project – DMP;
- Adequately justify why such data cannot be shared;
- Where appropriate, share the metadata without personal data.

To properly set up the publication of research data in open mode or their “FAIRification”, the different roles of the various parties in the institutional environment must be clarified in particular. Thus, it primarily involves a managerial decision.

At the institutional level, there needs to be a clear policy on who is authorized to enter into licensing and other agreements relating to the publication of/access to the research data. It must also be determined how the institution’s facilities will help this person, or the level of interaction between the individual and the local support must be established for this specific activity.

The most practical situation is if the primary person designated to enter into publication agreements at the institution is the person closest to the data – ideally, the data creator (researcher) or the research group leader who has access to the institution’s legal support. In the case of inter-institutional collaboration, it is advisable to establish the data handling regime at the very beginning (ideally with a contract).

R1.2 (Meta)data are associated with detailed provenance

Data provenance (also referred to as data lineage) is information about the entities, activities, and people involved in the creation of data that can be used to assess quality, reliability, or credibility. It is a part of academic research, and over the years, several models have been developed to cover this area. At its core, it is metadata associated with records that describe in detail the provenance, changes, and information supporting the credibility or validity of data. Data provenance is also important, among other things, for tracking errors in data and for reporting.

Simply put, data provenance helps answer the following questions:

- Why were the data created?
- How were the data created?
- Where were the data created?

Co-funded by the European Union
MINISTRY OF EDUCATION, YOUTH AND SPORTS

- When were the data created?
- Who created the data?
- Who cites the data and how?
- Does the data contain data from someone else that have been transformed or expanded?

This information should be described in a machine-readable format.[67]

The most general model is the W3C Provenance Data Model, which is applicable to many fields and areas. Detailed specifications are available on the official W3C group web pages: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

Currently, an ISO standard is being developed to deal with the provenance of information in the biotechnology environment based on the W3C Provenance Data Model. The first part of this ISO standard was published in 2023 and focuses on the provenance of biological material and associated data: ISO/TS 23494-1:2023 – Biotechnology – Provenance information model for biological material and data – Part 1: Design concepts and general requirements.

R1.3 (Meta)data meet domain-relevant community standards

There are several standards for the various types of data, ranging from general dataset descriptions (e.g. Dublin Core) to specific types of data. Thus, it is a good idea to think about their correct use from the very beginning of a project:

- Decide at the beginning of the project which databases and repositories will be used for the specific types of data;
- Differentiate between general repositories (e.g. Zenodo) and domain-specific repositories (e.g. CLARIN, ČSDA, Github) for a specific type of data;
- Visit the repository's web pages – searchable, for example, via www.re3data.org – and check the information about the required metadata;
- Go through the submission process to identify metadata requirements;
- Keep in mind that specific repositories for a certain type of data usually have validators for metadata.

If it is still not clear which repository should be used, the following should be done:

- Find out what the recommended minimum amount of information is, using e.g.
    - https://www.fged.org/projects/miame/
- Use the metadata required for our data type in our community or other metadata recommended in the following resources:
    - RDA https://rd-alliance.github.io/metadata-directory/standards/

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

- FAIRSHARING https://fairsharing.org/ in the sections "Standards" and "Collections"
- DCC https://www.dcc.ac.uk/guidance/standards/metadata/list

These may not always be official standards, since many communities/domains are still developing their standards. Thus, communities may have standards that are less formal, but that still increase the "fairness" of publishing (meta)data in a manner that increases the possibility of reuse for the community/domain.

In some situations, the data creator may have valid and specified reasons for deviating from standard good practice in a specific domain. A description of these deviations should also be part of the metadata about the respective dataset.

Semantic interoperability

The key to successful data exchange lies in the mutual understanding of the shared data. Semantic interoperability can be thought of as a state in which two people studying a particular artifact from the domain in which they work trigger the same cognitive processes, i.e. they understand the artifact as similarly as possible.

If a biologist talks to an actor about culture, one may think of the result of their last experiment with bacteria, and the other will think of developments in the world of theatre. Semantic interoperability should ensure that both understand what the other is thinking when they mention a protocol or growing a culture.

Semantic interoperability helps to describe data handling procedures and practices in specific domains or communities. Defining such a procedure can significantly streamline the entire process of data collection, evaluation, and publication, and can, for example, clarify or define the use of workflows or tools that effectively use or adhere to the FAIR Principles.

It is particularly important for semantic interoperability to be defined for interdisciplinary collaboration since misunderstandings could arise in the use of specific terms or procedures. In general, however, it can effectively contribute to the clarity and automation of data management procedures or even integration of data from various instruments/tools during data collection. In practice, it can help to ensure compatibility, e.g. when aggregating data.

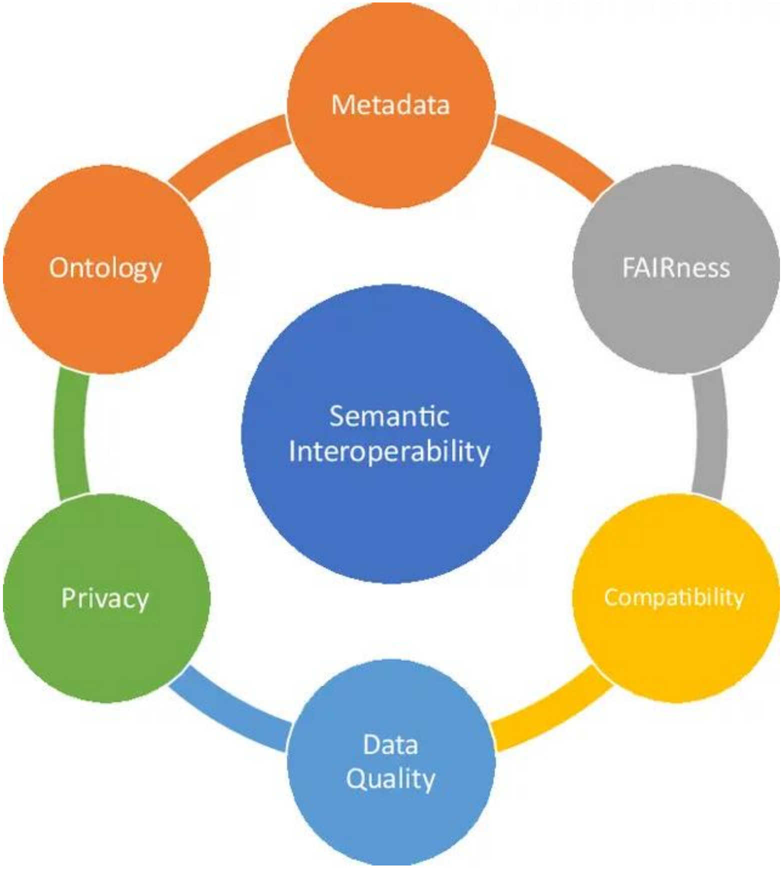


Figure 4: Semantic interoperability in data spaces is a complex problem involving multiple aspects.[68]

There is the SIP Wizard tool (https://sip-wizard.ds-wizard.org/wizard/) developed in cooperation with the *EOSC Semantic Interoperability Task Force* (details and instructions can be found here: https://osf.io/fn2wj/?view_only=) employing the Data Stewardship Wizard user interface and environment (https://ds-wizard.org/).

A semantic interoperability profile (SIP, Figure 6) can be defined in the tool, which is a list of declared implementation options focusing on the interoperability aspects of the FAIR Guiding Principles. It includes semantic artifacts and their support services selected by the community for a specific case study and data type.

The wizard depicts the SIP using a questionnaire that requires responses that explicitly profile a community's approach to semantic interoperability. SIPs are published by the SIP Wizard as FAIR (machine-readable) and open data (nanopublications[2]), **which can then serve as a reference for practical FAIR data management activities performed by members of that community**. SIP publishing also encourages the reuse and repurposing of SIPs by other communities, **saving time in "reinventing the wheel"** while promoting commonality in FAIR SIP implementation options, and the used FAIR Supporting Resources can be searched in FAIR CONNECT.[69]

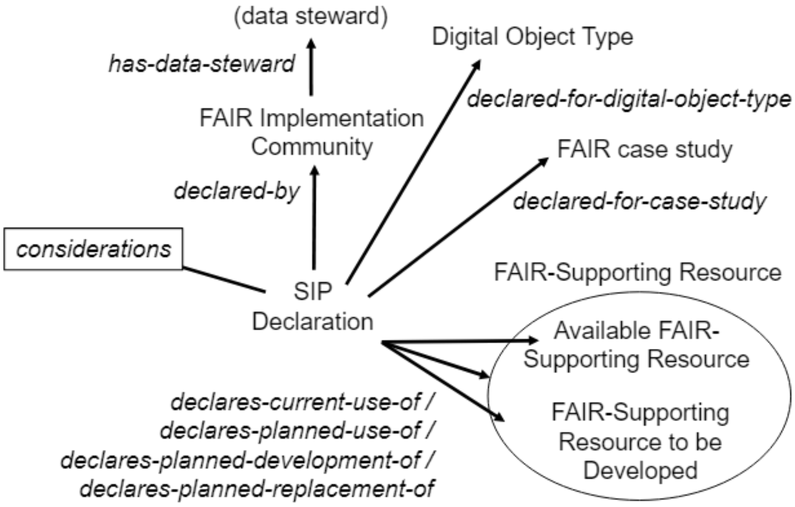

(data steward)
has-data-steward
Digital Object Type
declared-for-digital-object-type
FAIR Implementation Community
FAIR case study
declared-by
declared-for-case-study
considerations
FAIR-Supporting Resource
SIP Declaration
Available FAIR-Supporting Resource
FAIR-Supporting Resource to be Developed
declares-current-use-of /
declares-planned-use-of /
declares-planned-development-of /
declares-planned-replacement-of

Figure 5: Specifications for the semantic interoperability profile (SIP).[70]

[2] Nanopublications are expressed as a knowledge graph with metadata that are formal and machine interpretable. Since nanopublications are citable, they provide incentives for researchers to make their data accessible in standard formats that support data accessibility and interoperability.

Machine actionability

To understand what machine actionability is, we first need to define machine-readable data, which can be divided into two groups:

- Human-readable data that are tagged so that they can be read by machines (e.g. microformats, RDFa, HTML).
- Data file formats intended primarily for machine processing (CSV, RDF, XML, JSON). These formats are machine-readable only if the data they contain are formally structured; exporting a CSV file from a poorly structured table does not adhere to the definition.

Machine actionability refers to information that is consistently structured so that automatic processing or data extraction can be set up or programmed according to that structure. It has recently been emphasized in data management plans, metadata, and data analysis because tools must efficiently process and evaluate them.

Machine-actionable/-readable data must be structured.

Many research institutes have already introduced or are planning to introduce the use of an electronic lab notebook (ELN), a laboratory information management system (LIMS), or similar systems for managing samples, reagents, metadata, and data during a research project. The reason for this is that such software could structure information and make (meta)data “more” machine-actionable compared to traditional laboratory books or individual folders and files on a computer. The use of machine-actionable (meta)data allows for scalable solutions that can be applied over a project’s lifetime, increasing efficiency and ensuring that findings and contributions remain relevant within a grant or research group.

Similarly, funding providers and institutions ask researchers to make their (meta)data accessible in accordance with the FAIR Principles and in a machine-actionable manner. This means that (meta)data should be in databases that can expose them in such a way to allow search engines and harvesting servers to discover them, index them, and link them to other relevant contextual information, thereby vastly enhancing the likelihood of reusing the data.

Researchers benefit greatly from structuring metadata and data according to established standards in databases, which facilitates the simplified searching, filtering and reproducibility of experiments across different parameters and experimental conditions. This approach enables the easier integration of datasets, automation of manipulation tasks using common software tools, such as R and OpenRefine, and the use of visualization and research tools. In addition, the system allows for the seamless import, export, and exchange of metadata between platforms (while adhering to semantic interoperability in the community/domain). Moreover, machine-actionable metadata improve the findability of reference data and existing datasets through search engines and specialized data catalogues and portals, thereby increasing the efficiency of research and collaboration.

In addition to domain- or community-specific tools that can use metadata to structure, organize, or search, there are some general-purpose tools available.

Machine-Actionable DMP (maDMP)

The Data Stewardship Wizard is directly developed with the idea of using machine actionability to create a Data Management Plan – whether it involves a specific selection, highlighting important issues for a specific phase of a project, or exporting a completed knowledge model to a "paper" human-readable version of the DMP. The actionability of the whole maDMP will be expanded in the future through APIs, and a direct connection to both the SIP Wizard and other external services (CESNET) is planned. However, what every user can benefit from is, for example, the automatic evaluation of FAIR metrics, good DMP creation practices, and openness. These metrics are then displayed in graph form and provide an important form of feedback to the user.

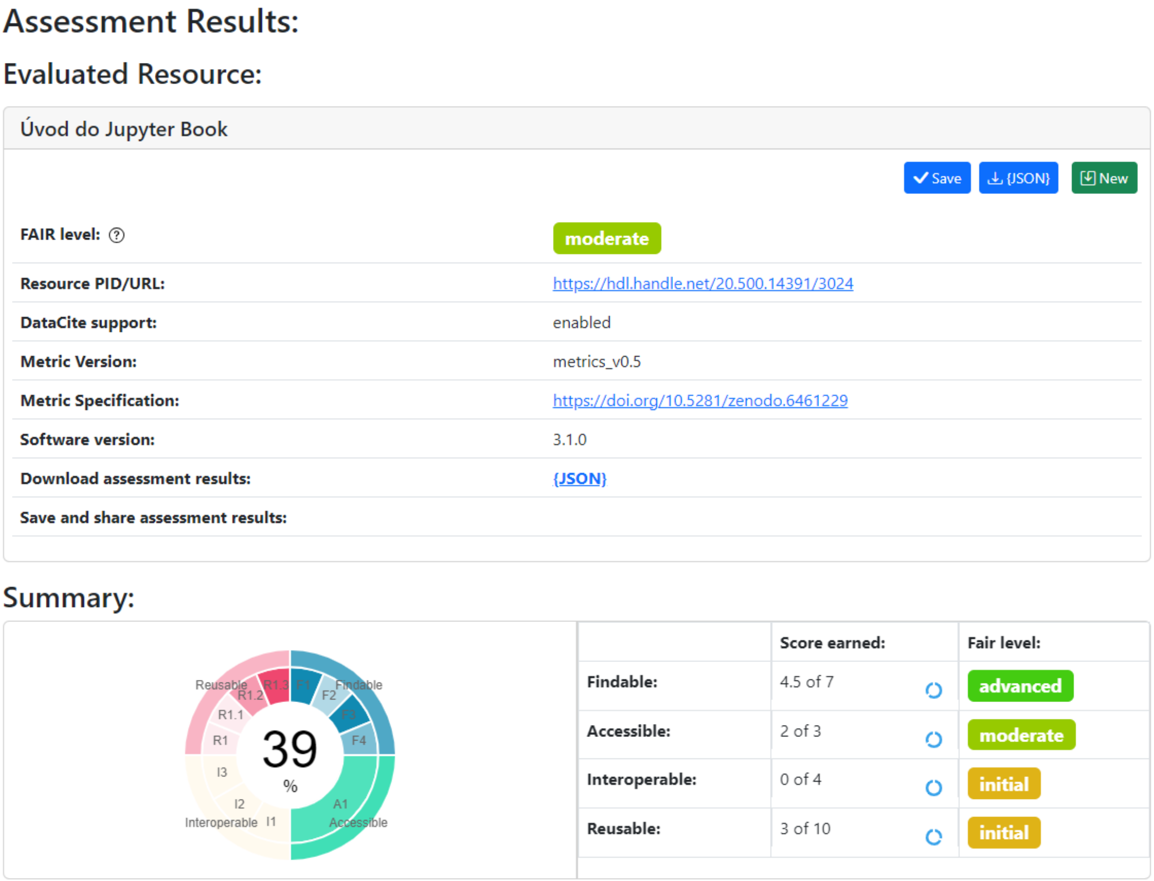
For a data steward who reviews/comments on the DMP, the metrics graph provides an overview of the consistency of completion or an overview of whether the metrics are consistent with the intent of a grant and the established internal rules for data management and publishing.

Machine-Actionable FAIR (maFAIR)

Another useful tool for using machine actionability is the automatic assessment of "fairness" through a persistent identifier for the data. This is the F-UJI tool https://www.f-uji.net/, which is a web service for the programmatic evaluation of the "fairness" of research datasets based on metrics developed in the FAIRsFAIR project. A web form can be used to enter the identifier (e.g. DOI, URL) of the dataset to be assessed. Optionally, the URI of a metadata service endpoint (OAI-PMH, SPARQL, CSW) can also be entered, which can be used by F-UJI to identify additional information.

This tool is especially useful for searching published data that we want to reuse. However, it will be most useful for datasets in domain-specific repositories where it is easier to find data relevant to the research community/domain.[71]

Co-funded by the European Union
MINISTRY OF EDUCATION, YOUTH AND SPORTS



Assessment Results:

Evaluated Resource:

Úvod do Jupyter Book

FAIR level: ⓘ	moderate
Resource PID/URL:	https://hdl.handle.net/20.500.14391/3024
DataCite support:	enabled
Metric Version:	metrics_v0.5
Metric Specification:	https://doi.org/10.5281/zenodo.6461229
Software version:	3.1.0
Download assessment results:	{JSON}
Save and share assessment results:	

Summary:

	Score earned:		Fair level:
Findable:	4.5 of 7	○	advanced
Accessible:	2 of 3	○	moderate
Interoperable:	0 of 4	○	initial
Reusable:	3 of 10	○	initial

Figure 6: Assessment results of datasets using the F-UJI tool.
Source: https://www.f-uji.net/

Useful links

UNESCO Recommendation on Open Science	UNESCO Recommendation on Open Science
Act no. 130/2002 Sb.	Act on the Support of Research and Development from Public Funds and on Amendments to Certain Related Acts (the Research and Development Support Act)
A FAIRy tale	A guide to the FAIR Principles for research data
AMNESIA	A tool for anonymizing data
ARGOS	A DMP tool
COAR	A repository search engine
CoreTrustSeal	Certification of trusted repositories
Creative Commons	Open licences for research objects
DCC DMP checklist	A DMP checklist
DMP online	A DMP tool
DMP Tool	A DMP tool
DSW	A DMP tool
ELIXIR	An online resource
EOSC	European Open Science Cloud
EOSC-CZ	EOSC Czech initiative
EU Open Science	European Commission regulations and resources
FAIR Cookbook	An online resource
FAIR Data Self Assessment Tool	A tool for assessing the FAIR level of research data
FAIR DataSet Maturity	A tool for assessing the FAIR level of research data
FAIR checker	A tool for assessing the FAIR level of research data
FAIRAware	A decision-making tool for FAIR data
FAIRification framework	An online resource
FAIRification Process	An online resource
FAIRplus	An online resource
FAIRsFAIR	Practical solution for using the FAIR Principles
FAIRsharing	A registry of terminological artifacts, models/formats, reporting guidelines, and identifier schemas
FAIRToolkit	An online resource
FOSTER	Training materials for Open Science
F-UJI	A tool for assessing the FAIR level of research data
GDPR decision tree	GDPR decision tree
GO FAIR	A community working to implement the FAIR Guiding Principles
HOW TO FAIR	Putting the FAIR Principles into practice, helping to develop a data management plan, and disseminating the results of a research project

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

National Repository Platform
for Research Data (NRP)
CESNET, zájmové sdružení právnických osob
Mailing address: Generala Piky 430/26, 160 00 Prague
info@eosc.cz; www.eosc.cz

Choose an open source license	A tool for selecting an appropriate licence
IDENTIFIKATORY.CZ	Czech web pages on persistent identifiers
Národní politika výzkumu, vývoje a inovací 2021+	National Research, Development and Innovation Policy of the Czech Republic 2021+
OA checker	A tool for checking Open Access journals
OpenAIRE	Infrastructure for open access to research (tools and services)
OpenDOAR	A directory of Open Access repositories
RDM/data steward training	An education portal
Re3data	A repository search engine
Science Europe guidance document	DMP guidelines for individual domains
Sherpa Romeo	A research engine for Open Access journals
Směrnice Evropského parlamentu a Rady (EU) 2019/1024	Directive (EU) 2019/1024 of the European Parliament and of the Council
Turning FAIR into reality	Final Report and Action Plan from the European Commission Expert Group on FAIR Data

References

[1] WILKINSON, Mark D., Michel DUMONTIER, IJsbrand Jan AALBERSBERG et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [online]. 2016, **3**(1), 160018. ISSN 2052-4463. Available at: doi:10.1038/sdata.2016.18

[2] FRIESIKE, Sascha and Thomas SCHILDHAUER. Open Science: Many Good Resolutions, Very Few Incentives, Yet. In: Isabell M. WELPE, Jutta WOLLERSHEIM, Stefanie RINGELHAN and Margit OSTERLOH, ed. *Incentives and Performance: Governance of Research Organizations* [online]. Cham: Springer International Publishing, 2015 [accessed on 2024-06-27], pp. 277–289. ISBN 978-3-319-09785-5. Available at: doi:10.1007/978-3-319-09785-5_17

[3] Open Science. *European Commission* [online]. 3 November 2024 [accessed on 2024-11-03]. Available at: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

[4] Open innovation, open science, open to the world. Shaping Europe's digital future. *European Commission* [online]. 16 June 2016 [accessed on 2024-06-27]. Available at: https://digital-strategy.ec.europa.eu/en/library/open-innovation-open-science-open-world

[5] HNÁTKOVÁ, Eva, Eva DIBUSZOVÁ and Martin SVOBODA. *Otevřená věda: Analýza mezinárodního prostředí* [online]. 2022 [accessed on 2024-06-27]. Available at: https://hdl.handle.net/20.500.14391/2894

[6] PONTIKA, Nancy, Petr KNOTH, Matteo CANCELLIERI and Samuel PEARCE. Fostering open science to research using a taxonomy and an eLearning portal. In: *i-KNOW '15: 15th International Conference on Knowledge Technologies and Data-Driven Business*: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business* [online]. Graz: ACM, 2015, pp. 1–8. ISBN 978-1-4503-3721-2. Available at: doi:10.1145/2809563.2809571

[7] MENDEZ, Eva, Rebecca LAWRENCE, Catriona J. MACCALLUM, Eva Moar et al. *Progress on open science: towards a shared research knowledge system: final report of the open science policy platform* [online]. Luxembourg: Publications Office of the European Union, 2020 [accessed on 2025-02-07]. ISBN 978-92-76-18882-7. Available at: https://data.europa.eu/doi/10.2777/00139

[8] DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION (EUROPEAN COMMISSION). *OSPP-REC: Open Science Policy Platform Recommendations* [online]. Luxembourg: Publications Office of the European Union, 2018 [accessed on 2024-06-27]. ISBN 978-92-79-88333-0. Available at: https://doi.org/10.2777/958647

[9] BERTRAM, Michael G., Josefin SUNDIN, Dominique G. ROCHE, Alfredo SÁNCHEZ-TÓJAR, Eli S. J. THORÉ and Tomas BRODIN. Open science. *Current Biology* [online]. 2023, **33**(15), R792–R797. ISSN 09609822. Available at: doi:10.1016/j.cub.2023.05.036

[10] *Open Knowledge. Open Definition* [online]. [accessed on 2024-08-24]. Available at: https://opendefinition.org/

[11] *Doporučení Komise (EU) 2018/790 ze dne 25. dubna 2018 o přístupu k vědeckým informacím a jejich uchovávání* [online]. [accessed on 2024-08-24]. Available at: https://eur-lex.europa.eu/eli/reco/2018/790/oj/eng

[12] *How To FAIR* [online]. [accessed on 2024-06-28]. Available at: https://howtofair.dk/

[13] ENGELHARDT, Claudia, Katarzyna BIERNACKA, Aoife COFFEY et al. D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions. *Zenodo* [online]. 21 December 2021 [accessed on 2024-06-27]. Available at: doi:10.5281/zenodo.5905866

[14] MONS, Barend, Cameron NEYLON, Jan VELTEROP et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use* [online]. 2017, **37**(1), 49–56. ISSN 0167-5265. Available at: doi:10.3233/ISU-170824

[15] HIGMAN, Rosie, Daniel BANGERT and Sarah JONES. Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights* [online]. 2019, **32**(1) [accessed on 2024-06-27]. ISSN 2048-7754. Available at: doi:10.1629/uksg.468

[16] HERINGOVÁ, Hana and Petra ČERNOHLÁVKOVÁ. *Úvod do problematiky perzistentních identifikátorů* [online]. 22 March 2024 [accessed on 2024-07-17]. Available at: https://hdl.handle.net/20.500.14391/3002

[17] FAIR Principles. *GO FAIR* [online]. [accessed on 2024-07-19]. Available at: https://www.go-fair.org/fair-principles/

[18] FENNER, Martin and ARYANI, AMIR. Introducing the PID Graph. *DataCite* [online]. 28 March 2019 [accessed on 2024-07-30]. Available at: https://doi.org/10.5438/jwvf-8a66

[19] MIERZ, Sandra. Project TAPIR: Harvesting the power of PIDs. *TIB-Blog* [online]. 1 March 2022 [accessed on 2024-07-19]. Available at: https://blog.tib.eu/2022/03/01/project-tapir-harvesting-the-power-of-pids/

[20] ČERNOHLÁVKOVÁ, Petra. Národní podpora implementace perzistentních identifikátorů [online]. 2022 [accessed on 2024-07-19]. Available at: https://doi.org/10.48813/k2xs-y923

[21] Persistent identifiers. *Digital Preservation Handbook* [online]. [accessed on 2024-06-28]. Available at: https://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers

[22] VALE, Patrick. Creating a landing page. *Crossref* [online]. [accessed on 2024-06-28]. Available at: https://www.crossref.org/documentation/member-setup/creating-a-landing-page/

[23] JAKUBOVÁ, Veronika. FTP, SFTP, SMB a další protokoly pro přenos souborů: který vybrat? *MasterDC* [online]. [accessed on 2024-06-28]. Available at: https://www.master.cz/blog/ftp-sftp-smb-protokoly-pro-prenos-souboru-ktery-vybrat/

[24] *What is SAML vs OAuth? Find out what's different* [online]. [accessed on 2024-06-28]. Available at: https://auth0.com/intro-to-iam/saml-vs-oauth

[25] A1: (Meta)data are retrievable by their identifier using a standardised communication protocol. *GO FAIR* [online]. [accessed on 2024-06-28]. Available at: https://www.go-fair.org/fair-principles/metadata-retrievable-identifier-standardised-communication-protocol/

[26] *Zenodo – Research. Shared* [online]. [accessed on 2025-02-07]. Available at: https://about.zenodo.org/policies/

[27] DataverseNO Preservation Policy. *info: DataverseNO* [online]. [accessed on 2025-02-07]. Available at: https://site.uit.no/dataverseno/about/policy-framework/preservation-policy/

[28] *Interoperability* [online]. [accessed on 2024-05-15]. Available at: https://fairplus.github.io/the-fair-cookbook/content/recipes/interoperability.html

[29] HANZLÍKOVÁ, Dagmar. Jak FAIR jsou vaše výzkumná data? *Zenodo* [online]. 3 April 2020 [accessed on 2024-05-15]. Available at: http://doi.org/10.5281/zenodo.3739188

[30] Standardy pro metadata. *Národní digitální knihovna* [online]. [accessed on 2024-05-15]. Available at: https://standardy.ndk.cz/ndk/standardy-digitalizace/metadata

[31] KRANTZ, Peter. A lightweight semantic interoperability framework for countries and large organizations (and small ones). *Peter Krantz* [online]. 15 December 2010 [accessed on 2024-05-15]. Available at: https://www.peterkrantz.com/2010/a-lightweight-semantic-interoperability-framework-for-countries-and-large-organizations-and-small-ones/

[32] DIRECTORATE-GENERAL FOR DIGITAL SERVICES (EUROPEAN COMMISSION). *New European interoperability framework: promoting seamless services and data flows for European public administrations* [online]. Luxembourg: Publications Office of the European Union, 2017 [accessed on 2025-02-07]. ISBN 978-92-79-63756-8. Available at: https://data.europa.eu/doi/10.2799/78681

[33] MARÍN-ARRAIZA, Paloma. Interoperability and data reuse Module 3: FAIR Research Data in the Life Cycle. University of Vienna. 2023.

[34] Semantic Interoperability. Joinup. *European Commission* [online]. [accessed on 2024-05-15]. Available at: https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/glossary/term/semantic-interoperability

[35] EOSC EXECUTIVE BOARD, GENERÁLNÍ ŘEDITELSTVÍ PRO VÝZKUM A INOVACE (EVROPSKÁ KOMISE), Oscar CORCHO, Magnus ERIKSSON, Krzysztof KUROWSKI et al. *EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture* [online]. Luxembourg: Publications Office of the European Union, 2021 [accessed on 2025-02-07]. ISBN 978-92-76-28949-4. Available at: https://data.europa.eu/doi/10.2777/620649

[36] Cross-Domain Interoperability Framework (CDIF). *The WorldFAIR Project* [online]. 8 September 2022 [accessed on 2025-02-07]. Available at: https://worldfair-project.eu/cross-domain-interoperability-framework/

[37] *CSV, Comma Separated Values (RFC 4180)* [online]. 7 May 2024 [accessed on 2025-02-07]. Available at: https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml

[38] ISO/IEC 21778:2017. *ISO* [online]. [accessed on 2025-02-07]. Available at: https://www.iso.org/standard/71616.html

[39] OASIS Open Document Format for Office Applications (OpenDocument) TC. *OASIS* [online]. [accessed on 2025-02-07]. Available at: https://groups.oasis-open.org/communities/tc-community-home2?CommunityKey=4bf06d41-79ad-4c58-9e8e-018dc7d05da8

[40] I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. *GO FAIR* [online]. [accessed on 2024-05-15]. Available at: https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation/

[41] Otevřená data. *Architektura eGovernmentu ČR* [online]. [accessed on 2024-05-15]. Available at: https://archi.gov.cz/nap:otevrena_data

[42] *CSV Schema* [online]. [accessed on 2024-05-15]. Available at: https://digital-preservation.github.io/csv-schema/

[43] TENNISON, Jeni. *CSV on the Web: A Primer* [online]. 25 February 2016 [accessed on 2024-05-15]. Available at: https://www.w3.org/TR/tabular-data-primer/

[44] *XML Schema* [online]. [accessed on 2024-05-15]. Available at: https://www.w3schools.com/xml/xml_schema.asp#:~:text=An%20XML%20Schema%20describes%20the,Formed%22%20and%20%22Valid%22

[45] *JSON Schema* [online]. [accessed on 2024-08-01]. Available at: https://json-schema.org/

[46] *RDF Schema 1.1* [online]. [accessed on 2024-05-15]. Available at: https://www.w3.org/TR/rdf-schema/

[47] OWL. *Semantic Web Standards* [online]. [accessed on 2024-08-01]. Available at: https://www.w3.org/OWL/

[48] *Shapes Constraint Language (SHACL)* [online]. 20 July 2017 [accessed on 2024-05-15]. Available at: https://www.w3.org/TR/shacl/

[49] *FAIRsharing* [online]. [accessed on 2024-05-15]. Available at: https://fairsharing.org/

[50] *Linked Open Vocabularies (LOV)* [online]. [accessed on 2024-05-15]. Available at: https://lov.linkeddata.es/dataset/lov

[51] *BARTOC.org* [online]. [accessed on 2024-05-15]. Available at: https://bartoc.org/

[52] AVANÇO, Karla. FAIR Principles and Linked Open Data. *The road to FAIR* [online]. 30 July 2021 [accessed on 2022-04-28]. Available at: doi:10.58079/trrt

[53] BLANEY, Jonathan. Introduction to the Principles of Linked Open Data. *Programming Historian* [online]. 7 May 2017 [accessed on 2024-05-15]. Available at: https://programminghistorian.org/en/lessons/intro-to-linked-data

[54] *5hvězdičková otevřená data* [online]. [accessed on 2024-08-01]. Available at: http://5stardata.info/cs/

[55] BERNERS-LEE, Tim. *Linked Data – Design Issues* [online]. 27 July 2006 [accessed on 2024-05-15]. Available at: https://www.w3.org/DesignIssues/LinkedData.html

[56] Stanovení podmínek užití otevřených dat. *Otevřená data* [online]. [accessed on 2022-04-28]. Available at: https://opendata.gov.cz/cinnost:stanoveni-podminek-uziti

[57] Licenses. *Open Source Initiative* [online]. 16 September 2022 [accessed on 2024-08-01]. Available at: https://opensource.org/licenses

[58] R1.1: (Meta)data are released with a clear and accessible data usage license. *GO FAIR* [online]. [accessed on 2024-06-21]. Available at: https://www.go-fair.org/fair-principles/r1-1-metadata-released-clear-accessible-data-usage-license/

[59] R1.2: (Meta)data are associated with detailed provenance. *GO FAIR* [online]. [accessed on 2024-06-21]. Available at: https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/

[60] R1.3: (Meta)data meet domain-relevant community standards. *GO FAIR* [online]. [accessed on 2024-06-21]. Available at: https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/

[61] FAIRSHARING TEAM. *FAIRsharing record for: Brain Imaging Data Structure* [online]. FAIRsharing. 2018 [accessed on 2024-06-21]. Available at: doi:10.25504/FAIRSHARING.RD1J6T

[62] *Metadata review guidelines* [online]. [accessed on 2024-08-01]. Available at: https://data.4tu.nl/s/documents/Metadata_review_guidelines_June_2021.pdf

[63] *Data licenses* [online]. [accessed on 2024-08-01]. Available at: https://fairplus.github.io/the-fair-cookbook/content/recipes/reusability/ATI_licensing_data.html

[64] EL EMAM, Khaled and Luk ARBUCKLE. *Anonymizing Health Data*. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-6307-9.

[65] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*. 2016.

[66] EL EMAM, Khaled. *Guide to the de-identification of personal health information*. Boca Raton: Taylor & Francis, 2013. ISBN 978-1-4665-7906-4.

[67] FAIR Data Principles. *NFDI4Chem Knowledge Base* [online]. [accessed on 2024-06-21]. Available at: https://knowledgebase.nfdi4chem.de/knowledge_base/docs/fair/

[68] BOUKHERS, Zeyd, Christoph LANGE and Oya BEYAN. Enhancing Data Space Semantic Interoperability through Machine Learning: a Visionary Perspective. *arXiv* [online]. 15 March 2023 [accessed on 2024-05-15]. Available at: doi:10.48550/arXiv.2303.08932.

[69] GONZALEZ, Esteban and Anne-Sofie FINK. *Synchronisation Workshop 2023*.

[70] MAGAGNA, Barbara, Erik SCHULTES and Tobias KUHN. *FAIR Implementation Profile (FIP) Ontology* [online]. [accessed on 2024-05-15]. Available at: https://peta-pico.github.io/FAIR-nanopubs/fip/index-en.html

[71] Machine-Actionability. *Data Stewardship Wizard* [online]. [accessed on 2024-05-15]. Available at: https://ds-wizard.org/machine-actionability

[72] *Open data, FAIR data and RDM: the ugly duckling* [online]. [accessed on 2025-03-14]. Available at: https://www.uhasselt.be/en/university-library/research/research-data-management/fair