

WEBINAR:

Data management and ethics

- Open Science & FAIR data principles
- Sensitive data management lifecycle
- Common breaches in handling sensitive data
- Semantic artefacts and data annotation:
 - metadata, ontologies and vocabularies
 - overview of ontologies for EMG/EEG data
 - data annotation examples
- Open access data repositories



Image generated with Stable Diffusion

Lecturers: Milan Ojsteršek, Matjaž Divjak
University of Maribor, Slovenia



Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders



HybridNeuro focuses on development of Hybrid Neural-machine Interfaces, which record cerebral and muscular signals, and aims to improve the objectivity, precision and personalisation of monitoring and rehabilitation of neuromuscular disorders, such as stroke.



Funded by
the European Union



UK Research
and Innovation



CHALMERS
UNIVERSITY OF TECHNOLOGY



This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement no. 101079392 and from the UK Research and Innovation (UKRI) government's Horizon Europe funding guarantee scheme under grant agreement no. 10052152.



Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders

Our mission:

- 1 Exploratory research project for development of Hybrid Neural-machine Interfaces
- 2 Summer schools
- 4 Workshops
- 8 Webinars
- 1 Biomedical Signals Data Repository
- 1 Massive open online course (MOOC) on Hybrid Neuroscience
- 1 International HybridNeuro Hub
- 12 National/international events

Visit us at:

<https://www.hybridneuro.feri.um.si>

<https://twitter.com/hybridneuro>



This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement no. 101079392 and from the UK Research and Innovation (UKRI) government's Horizon Europe funding guarantee scheme under grant agreement no. 10052152.



University of Maribor



Imperial College
London



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH



REGISTER NOW!



GA No. 10052152



GA No. 101079392



FREE REGISTRATION

Travel & accommodation costs
to be covered by participants

Summer school on Hybrid Neural Interfaces

July 8th-12th 2024, Maribor, Slovenia

- Surface & intramuscular HDEMG
- Identification of neural codes
- EEG & functional brain connectivity
- Corticomuscular coupling
- Movement augmentation
- Hybrid Neural Interfaces in practice
- Keynote lectures
- Practical examples
- Student 2 student explanations
- Present your project
- Ask top experts
- Active consultations

WEBINAR:

Data management and ethics

- Open Science & FAIR data principles
- Sensitive data management lifecycle
- Common breaches in handling sensitive data
- Semantic artefacts and data annotation:
 - metadata, ontologies and vocabularies
 - overview of ontologies for EMG/EEG data
 - data annotation examples
- Open access data repositories



Image generated with Stable Diffusion

Lecturers: Milan Ojsteršek, Matjaž Divjak
University of Maribor, Slovenia



Data management and ethics

Milan Ojsteršek

University of Maribor, Faculty of Electrical Engineering and Computer Science

milan.ojstersek@um.si



GA No. 101079392



GA No. 10052152



CHALMERS
UNIVERSITY OF TECHNOLOGY

Imperial College
London

Openscience

Michel Nielsen said „**Open Science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the Discovery process**”. **Scientific Knowledge of all kinds: journal articles, data, code, online software tools, questions, ideas, speculations, failures, ...and anything which can be considered knowledge.**”

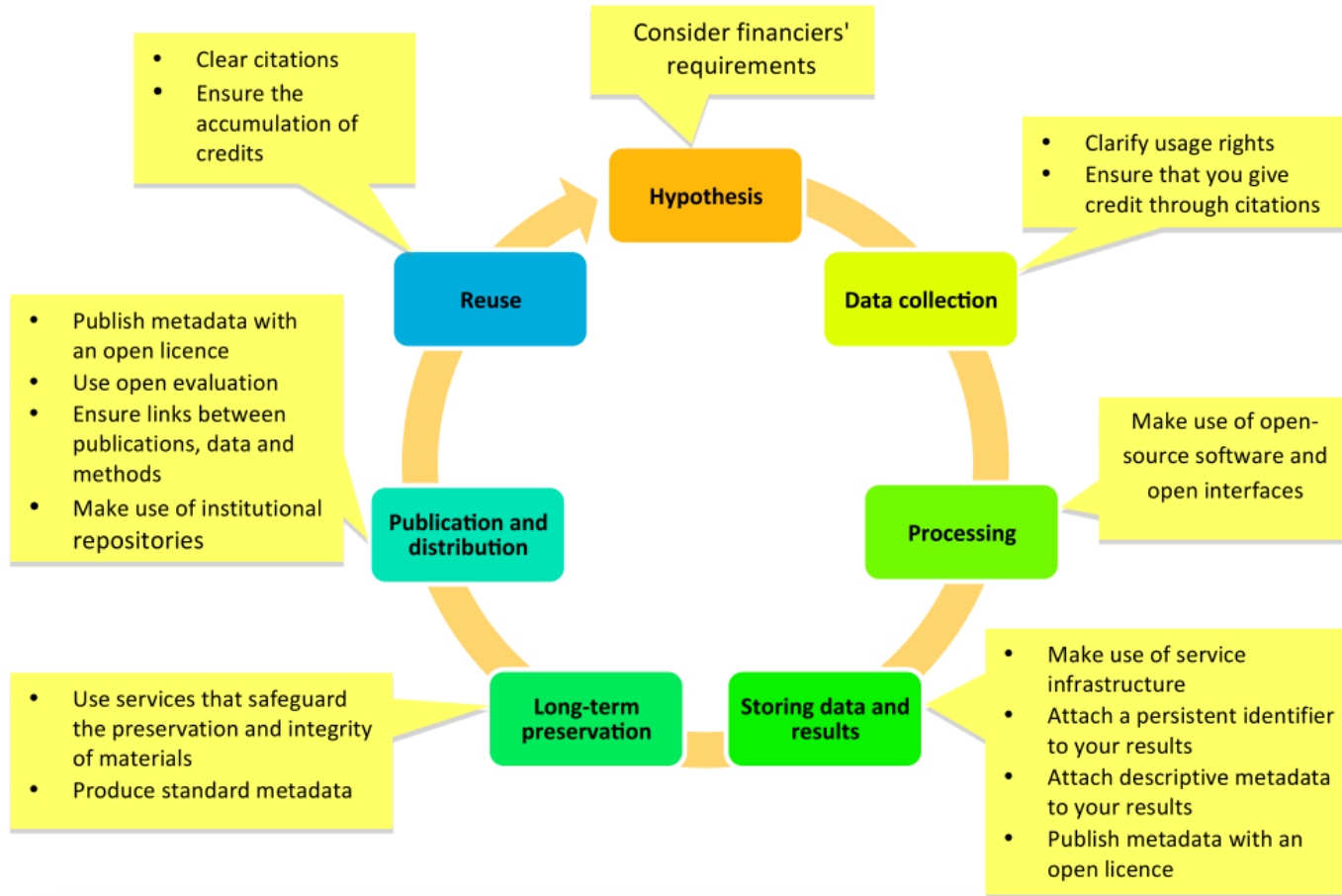
Open access to scientific publications
Open Data
Open Software
Open services and workflows
Open Research Infrastructure
Open peer review
Open Learning Materials
Open Research Methods
Open Lab Notes
Citizen Science



Reproducibility of research

Foster: What is Open Science? Introduction. Available at <https://www.fosteropenscience.eu/content/what-open-science-introduction>

Promoting openness at different stages of the research process



Foster: Open Science and Research Initiative (2014). *Open Science and Research Handbook*. [English version]. Available at <https://www.fosteropenscience.eu/sites/default/files/pdf/3986.pdf>

Open Reproducible Research



Open Reproducible Research is based on:

- **Irreproducibility Studies:** The act during which the results of a study or an experiment can be replicated and reproduced.
- **Open Lab/Notebooks:** Laboratory research records, diaries, journals, workbooks etc. offered online free of cost with terms that allow reuse and redistribution of the recorded material.
- **Open Science Workflows:** A sequence of processes scientists make to administer and disseminate convoluted scientific examinations offered online and free of cost allowing the reuse of the material.
- **Open Source in Open Science:** Software where the source code is available free of cost with terms that allow dissemination and adaptation.
- **Reproducibility Guidelines:** Ground rules to assist with the recreation of research experiments and studies.
- **Reproducibility Testing** refers to the process of validating that the reported research results can be obtained in an independent experiment

Open Science Tools

Refers to the tools that can assist in the process of delivering and building on Open Science.

Tools are:

- **Open archives** that host scientific literature, data, software and other research objects and make their content freely accessible to everyone in the world.
- **Open services** offered by organisations and institutions which is possible to use free of cost.
- **Open Workflow Tools** (apparatuses and services) that promote open scientific projects.

Open Data

*A piece of data or content is open if **anyone** is **free to use, reuse, and redistribute** it — subject only, at most, to the requirement to attribute and/or share-alike” -- opendefinition.org*

This means, according to the Open Knowledge Foundation:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and Redistribution:** the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, reuse and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed

Open knowledge foundation: What is open? <https://okfn.org/en/library/what-is-open/>

Research Data

Types of data:

- Raw / Cleaned or Filtered
- Field data
- Experimental data
- Derived data
- Qualitative / Quantitative
- Structured / Semi-structured/ Unstructured
- Tabular/ Hierarchical / Graphs
- Open access / Restricted access
- Linked data
- Metadata
- Big data

Data sources:

- Devices
- Instruments
- Sensors
- Software
- People

- Observations
- Experiments
- Simulations
- Emulations
- Surveys
- Interviews
- Text analysis
- Text mining

Data types:

- Numeric (tables, measurement results, counters, sensor data, etc.).
- Textual (books, notes, surveys...).
- Audio visual data (images, sound recordings, video, animation).
- Spatial.
- Specific to the scientific discipline.
- Specific to the measuring instrument or device.

Data management ethics



Ethics in data management refers to the principles, guidelines, and standards that govern the responsible collection, storage, processing, and usage of data.

Key aspects are:

- privacy of individuals whose data is being collected and analyzed.
- Transparency of data collection, data usage, and access to data.
- Accuracy and reliability of data.
- Establishing clear policies, procedures, and responsibilities for data management.
- Protecting data from unauthorized access, breaches, and cyberattacks.
- Adhering to relevant laws, regulations, and industry standards.
- Usage of data in ways that align with ethical principles.
- Data collection and analysis should be fair and avoid biases that can lead to discrimination.
- Clarifying ownership rights and obtaining explicit consent for data collection and usage.

Ethical data management is an ongoing process that requires continuous monitoring, evaluation, and improvement.

Why data ethics are important?



- **Protects individuals:** Our data is often personal and can be used to make decisions about us, like job applications or loan approvals. Ethical data management ensures our information is treated respectfully and used as intended.
- **Builds trust:** When organizations handle data ethically, it builds trust with users who are more likely to share their information.
- **Mitigates risks:** Unethical data practices can lead to security breaches, discrimination, and bias in AI algorithms.

Important legislation for data management ethics

- **General Data Protection Regulation (GDPR)**: The GDPR is a comprehensive data protection law in the European Union (EU) that governs the processing of personal data of EU residents. It sets strict requirements for the collection, storage, processing, and transfer of personal data and imposes significant penalties for non-compliance.
- **UK Data Protection Act 2018 (DPA 2018)**. The DPA 2018 was enacted to supplement the GDPR and ensure that data protection standards remain consistent in the UK post-Brexit.
- **European Health data space: Data Governance Act, primary use of data, secondary use of data, regulation to set up the European Health Data Space, harmonised rules on fair access to and use of data (Data Act), measures for a high common level of security of network and information systems across the EU.**
- **Health Insurance Portability and Accountability Act (HIPAA)**: HIPAA is a United States federal law that regulates the handling of protected health information (PHI) by healthcare providers, health plans, and their business associates. HIPAA establishes standards for the security and privacy of PHI to protect patients' confidentiality and prevent unauthorized access.

Research misconducts regarding data publication

- Plagiarism
- Paper mills
- Gift authorship
- Data sets slicing
- Faking data
 - Data falsification (fitting data to a hypothesis).
 - Data fabrication (generating data without actual measurements). Examples: [Rajvir Dahiya](#), [George Laliotis](#), [Ming He](#)
 - Data amputation (elimination of inappropriate or outlier data, which not fitting to a hypothesis).
 - Data imputation (adding data in place of missing data)
- Falsification or manipulation of images ([Khalid Shah papers](#))
- AI generated images and texts.
- See more on [Retraction watch database](#) and [Elisabeth Bik blog](#)

Why some researchers might not follow research ethics



- **Pressure to Publish** (competition, funding).
- **Lack of awareness or understanding** (early career researchers, misinterpretation).
- **Desire for personal gain** (fame and recognition, financial gain, corruption).
- **Conflict of interest.**
- **Researchers don't publish of negative research results.**

Things that can help promote research ethics



- **Training:** Ensuring researchers receive proper training in research ethics is crucial.
- **Strong Oversight:** Research institutions need to have clear and well-enforced ethical guidelines and effective oversight mechanisms.
- **Open Communication:** Creating an environment where researchers feel comfortable raising concerns about potential ethical issues is important.
- **Focus on Quality over Quantity:** Shifting the emphasis from publishing frequently to publishing high-quality research conducted ethically can help reduce pressure on researchers.

By promoting a strong culture of research ethics, we can ensure that research is conducted responsibly and benefits society as a whole.

What's in the can?



Source:



Now you know!



Source:  DATA DOCUMENTATION INITIATIVE

What these numbers mean?

001	12	01	98
002	36	02	175
003	72	01	94
004	42	01	130
005	18	02	125

Source:  **DDI**[®]
DATA DOCUMENTATION INITIATIVE

Now you know!

Respondent ID	Age	Sex	Weight
001	12	01	98
002	36	02	175
003	72	01	94
004	42	01	130
005	18	02	125

Age is in years as of last birthday

Sex: 01 = Male, 02 = Female

Weight is in pounds

Metadata

Source:  **DDI**[®]
DATA DOCUMENTATION INITIATIVE

Metadata Standards - Simple

Dublin Core

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

DataCite

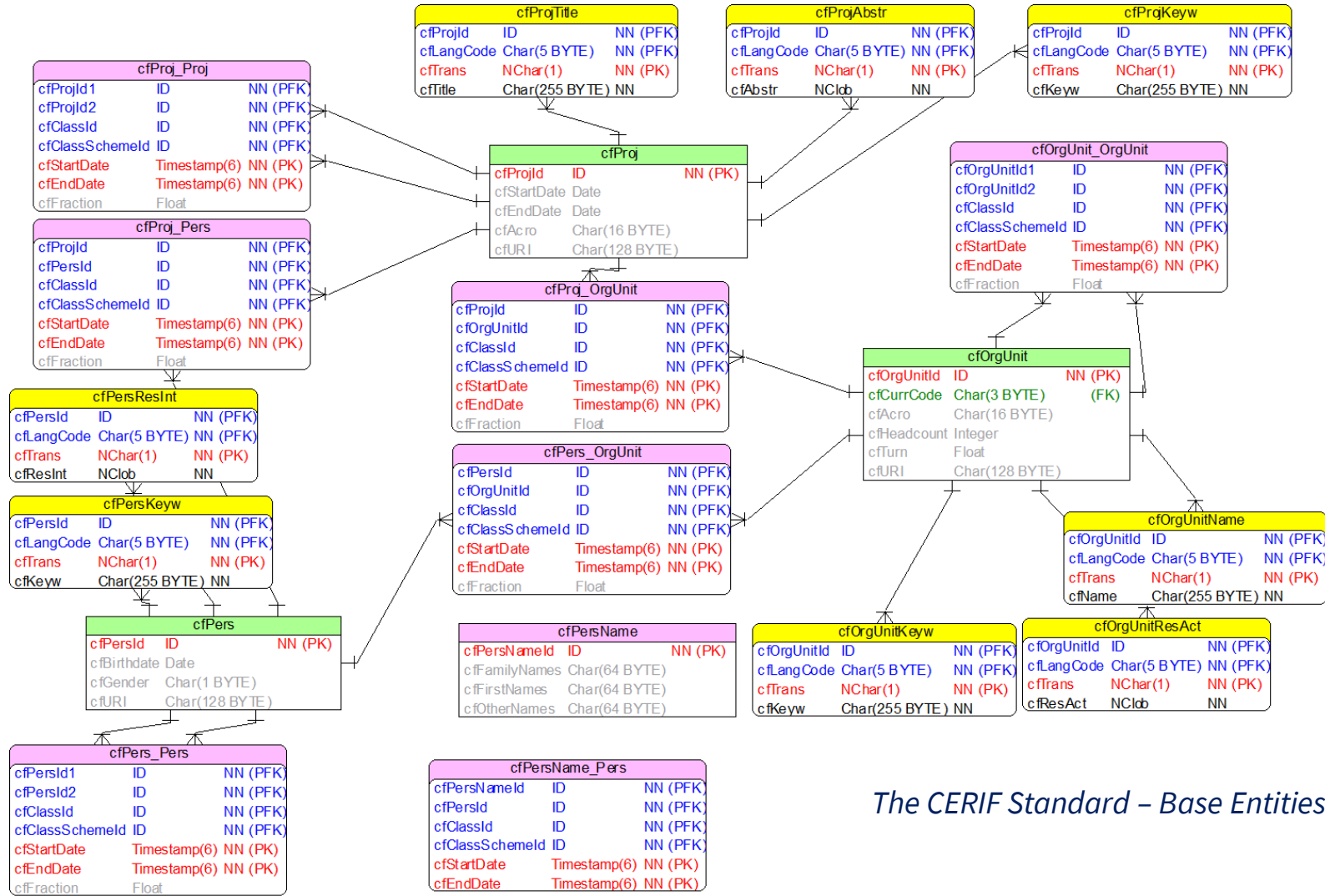
- Title
- Creator
- Publisher
- Identifier
- Publication Year
- Resource Type
- Subject
- Contributor
- Date
- Related identifier
- Description
- Geolocation
- Language
- *Alternate identifier*
- *Size*
- *Format*
- *Version*
- *Rights*
- *Funding Reference*

EDMI

- Name
- Description
- Identifier
- url
- Creator
- Date Created
- license
- Data Standard
- Date Modified
- Access URL
- Access Interface
- Structure
- Included In
- Measurement Technique
- Keywords
- Variable Measured
- Format
- Scientific Type
- Includes
- Content Type
- Size
- Authentications

- Version
- Metric
- Same as
- Spatial Coverage
- Temporal coverage
- Citation
- Reference citation
- compression

Metadata standards - complex



The CERIF Standard – Base Entities

Important sources for metadata



RDA [Guidelines for publishing structured metadata on the Web](#)
DCC list of metadata standards: <https://www.dcc.ac.uk/guidance/standards/metadata/list>
RDA metadata standards catalog: <http://rd-alliance.github.io/metadata-directory/standards/>
[FAIRsharing.org](#) - A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*. DCMI Metadata Terms: <https://dublincore.org/specifications/dublin-core/dcmi-terms/>
DataCite Metadata Schema 4.3: <https://schema.datacite.org/meta/kernel-4.3/>
DCAT 3.0: <https://www.w3.org/TR/2021/WD-vocab-dcat-3-20210504/>
DCAT Application profile for data portals in Europe 2.0.1: <https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe>
GeoDCAT-AP 1.0.1 <https://joinup.ec.europa.eu/solution/geodcat-application-profile-data-portals-europe/distribution/geodcat-ap-101-docx>
StatDCAT- AP 1.0.1: <https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/statdcat-application-profile-data-portals-europe/release/101>
DDI Codebook 2.5: https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html
Bioschemas: <https://bioschemas.org/>

Semantic artefacts

Semantic artefacts (Corcho et al., 2021) are defined as a machine-actionable and -readable formalization of a conceptualization enabling sharing and reuse by humans and machines. Semantic artefacts can take various forms and serve different purposes, including:

- Ontologies

- Taxonomies

- Vocabularies

- Controlled vocabularies

- Metadata schemas

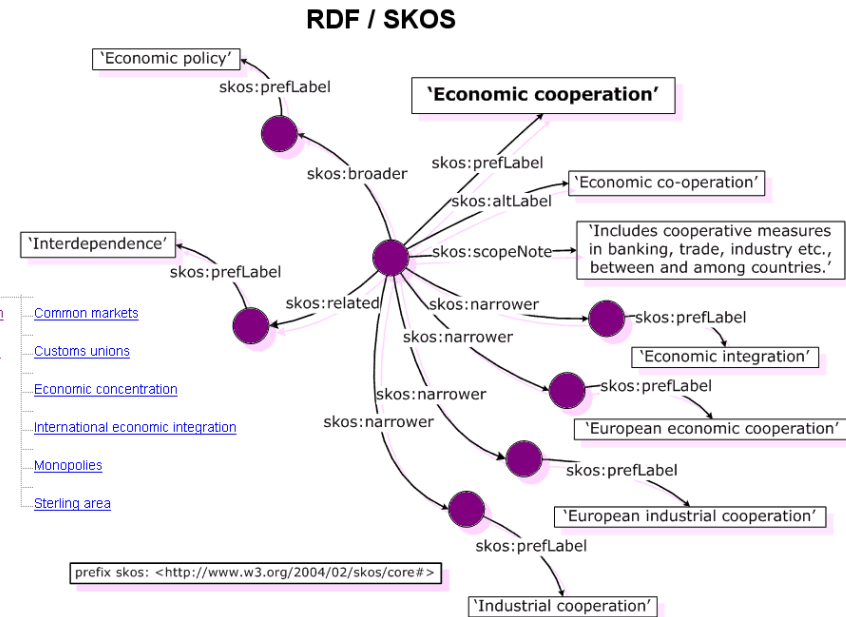
- Data dictionaries

- Semantic models

Corcho, O. et al., EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, <https://data.europa.eu/doi/10.2777/620649>

SKOS - Simple Knowledge Organization System

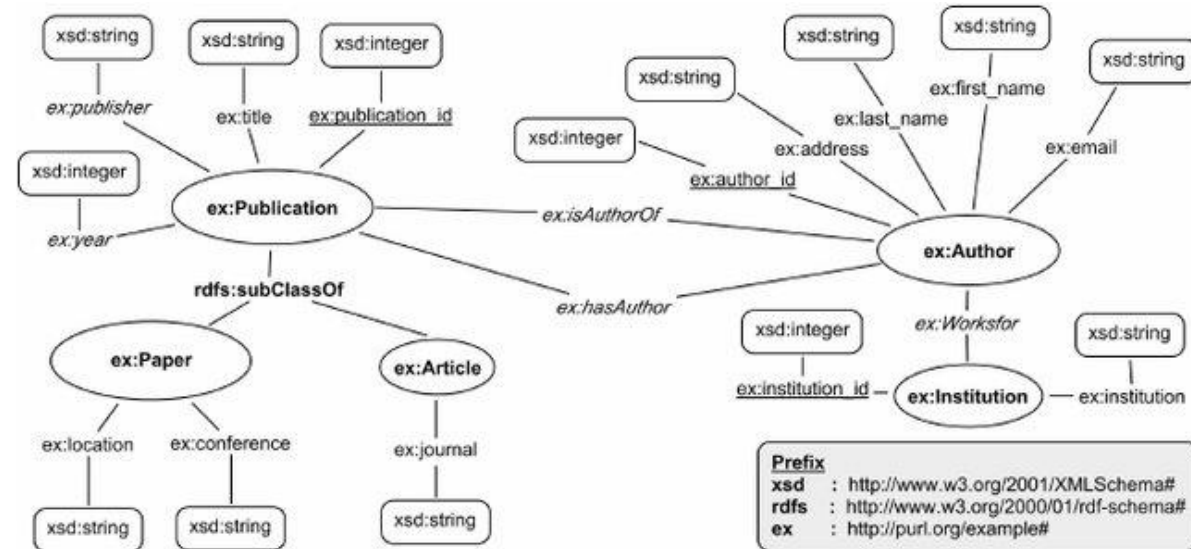
SKOS is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web.



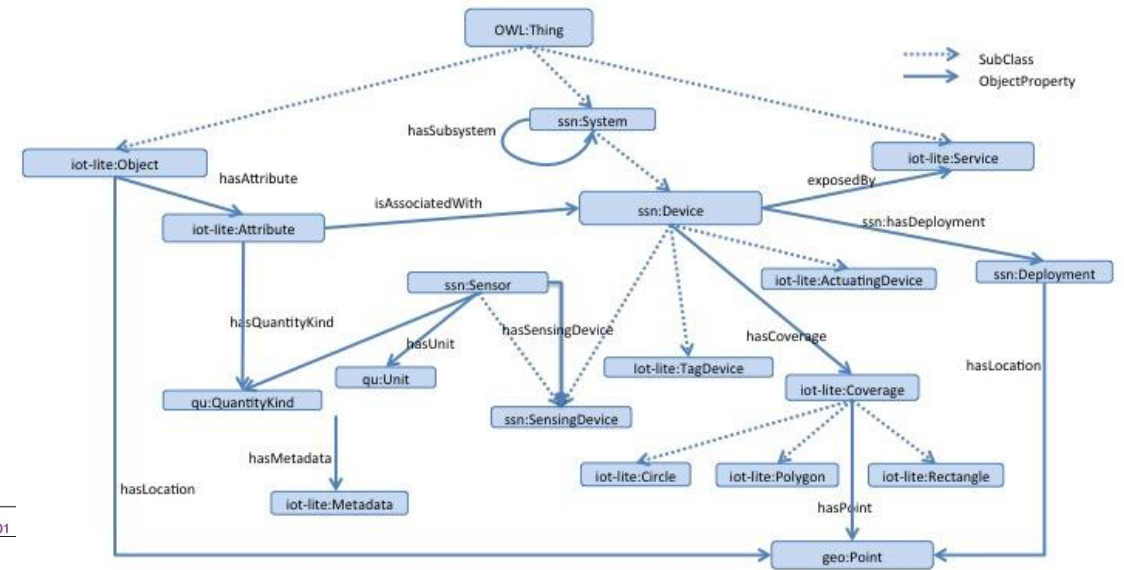
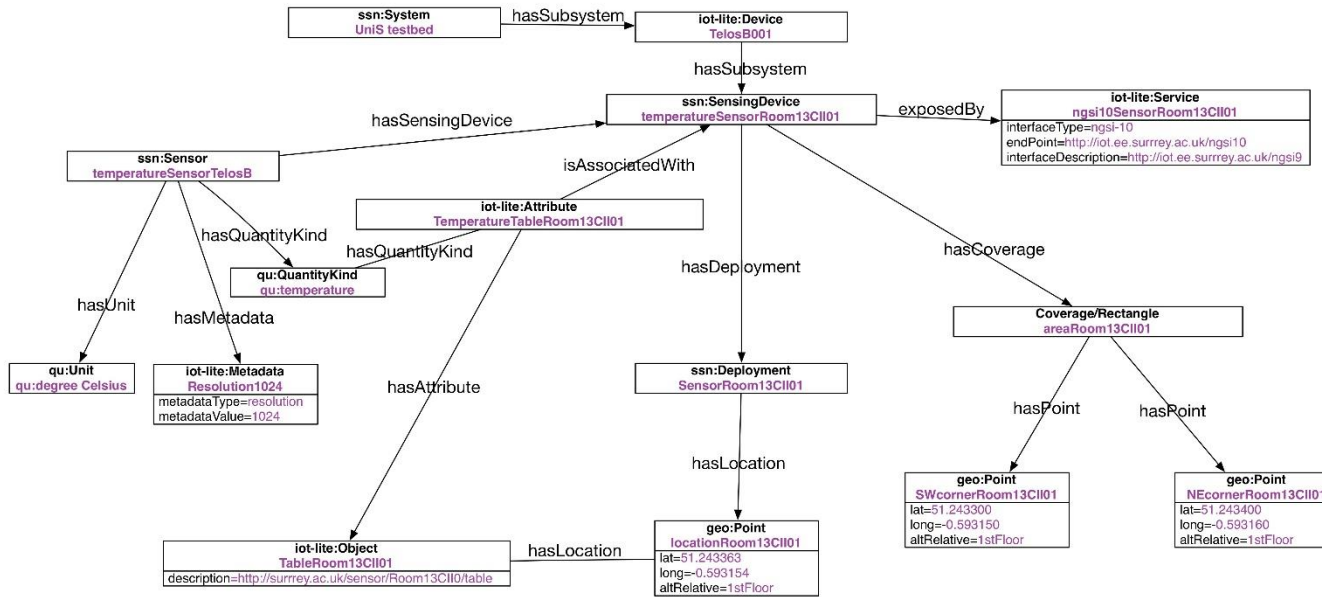
Ontology

Components of ontology are:

- **Individuals:** Instances or objects (the basic or "ground level" objects).
- **Classes:** Sets, collections, concepts, classes in programming, types of objects or kinds of things.
- **Attributes;** Aspects, properties, features, characteristics or parameters that objects (and classes) can have.
- **Relations:** Ways in which classes and individuals can be related to one another.
- **Function terms:** Complex structures formed from certain relations that can be used in place of an individual term in a statement.
- **Restrictions:** Formally stated descriptions of what must be true in order for some assertion to be accepted as input.
- **Rules:** Statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form.

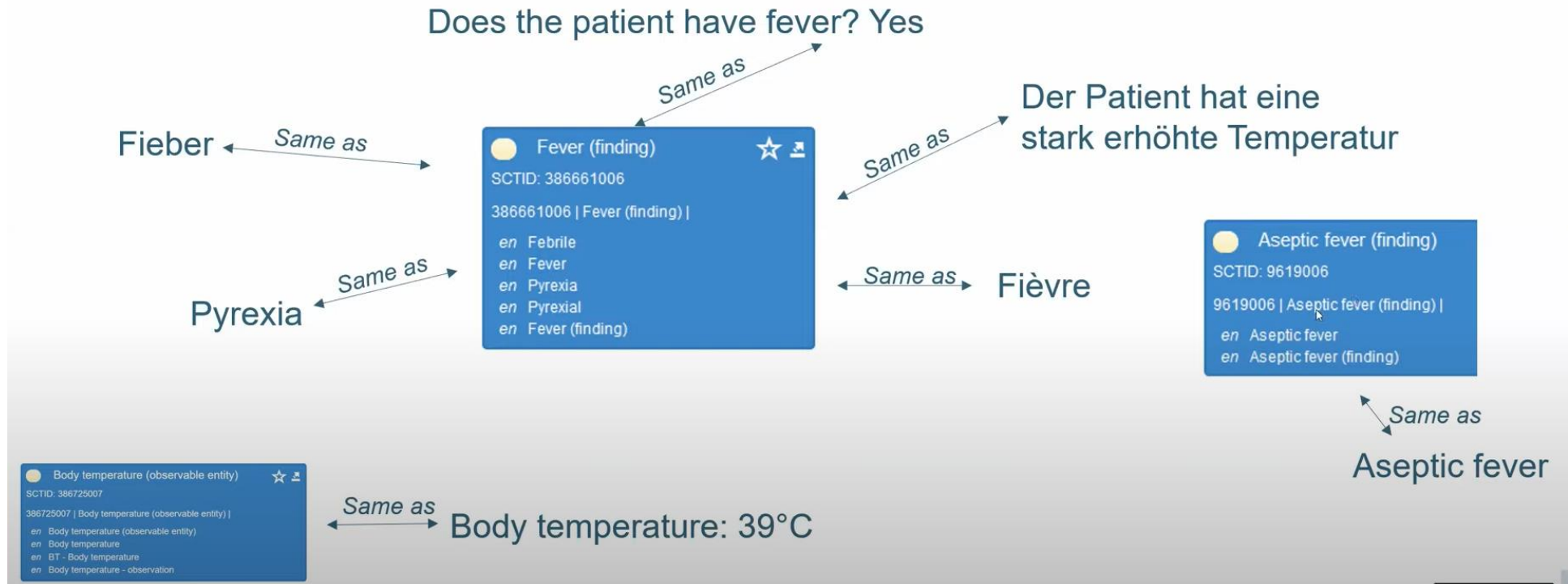


IoT Lite Ontology example



Source: IoT Lite Ontology: Available on <https://www.w3.org/Submission/iot-lite/>

Example of a semantic description of a patient's fever



Source: Sabine Osterle. SPHN Data Ecosystem for FAIR Data. Available on <https://www.youtube.com/watch?v=pqV0qp4oisM> [22. 5. 2022]

Semantic artefacts catalogs

- DDI Controlled Vocabularies: <https://ddialliance.org/controlled-vocabularies>
- CESSDA vocabularies: <https://vocabularies.cessda.eu/>
- COAR vocabularies: <https://www.coar-repositories.org/news-updates/what-we-do/controlled-vocabularies/>
- Research vocabularies Australia: <https://vocabs.arcd.edu.au/search#!?q=&pp=15&p=1&activeTab=vocabularies>
- Unified Medical Language System (UMLS): <https://www.nlm.nih.gov/research/umls/index.html>
- Getty's vocabularies: <https://www.getty.edu/research/tools/vocabularies/index.html>
- Data privacy vocabulary: <https://harshp.com/dpv/dpv/>
- BARTOC: <https://bartoc.org/>
- Backbone thesaurus: <https://www.backbonethesaurus.eu/>
- NERC: <https://vocab.nerc.ac.uk/>
- Finto: <https://finto.fi/en/>
- EU vocabularies: <https://op.europa.eu/en/web/eu-vocabularies>
- Linked open vocabularies: <https://lov.linkeddata.es/dataset/lov/>
- SPAR Ontologies (<http://www.sparontologies.net>)
- Open BioMedical Ontologies (OBO) - <https://obofoundry.org>
- [Bioportal](https://bioportal.lirmm.fr/) – a comprehensive repository of biomedical ontologies, <https://bioportal.lirmm.fr/>
- [AgroPortal](#) is an ontology portal/repository (with periodically updated versions) dedicated to the agronomic and plant domains.
- <https://www.ebi.ac.uk/ols4/ontologies>
- <https://www.nlm.nih.gov/research/umls/index.html>
- <https://medportal.bmicc.cn/ontologies>
- [OGC standards](#) – list of standards and ontologies on geospatial domain.
- [QUDT CATALOG](#) - Quantities, Units, Dimensions and Data Types Ontologies

Persistent identifiers



- <https://doi.org/10.13140/2.1.2889.8561>
- <https://arxiv.org/abs/2011.10574>
- <http://hdl.handle.net/11356/1244>
- PMID: [32389849](https://pubmed.ncbi.nlm.nih.gov/32389849/)
- PMCID: [PMC7204709](https://pubmed.ncbi.nlm.nih.gov/PMC7204709/)
- http://purl.org/coar/resource_type/c_5ce6
- [URN:NBN:SI:DOC-ZCQPLPGX](https://nbn-resolving.org/urn:nbn:si:doc-zcqplpgx)

- <https://orcid.org/0000-0003-1743-8300>
- <https://viaf.org/viaf/118892012/>
- <http://www.isni.org/isni/0000000114559647>
- <https://ror.org/01d5jce07>
- [grid.17063.33](https://grid.ac/identifiers/10.13140/2.1.2889.8561)

FAIR (Findable, Accessible, Interoperable, Reusable)

What is FAIR DATA?



Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE



Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE



Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE



Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Source:

<https://libereurope.eu/blog/2018/07/13/fairdataconsultation/liber-fair-data-2/>

To be Findable:

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

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

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

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

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

A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

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

I2. (meta)data use vocabularies that follow FAIR principles.

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

TO BE RE-USABLE:

R1. meta(data) have a plurality of accurate and relevant attributes.

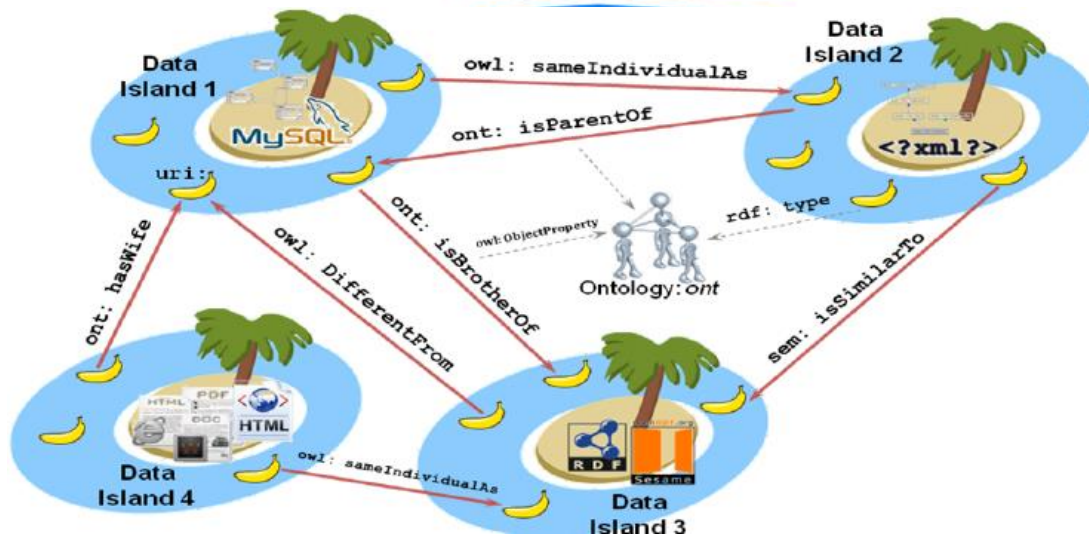
R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

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

Source: <https://www.force11.org/group/fairgroup/fairprinciples>

How to achieve interoperability between data islands?



(Meta)data Interoperability principles:

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (Meta)data use vocabularies that follow FAIR principles.
- (Meta)data include qualified references to other (meta)data.

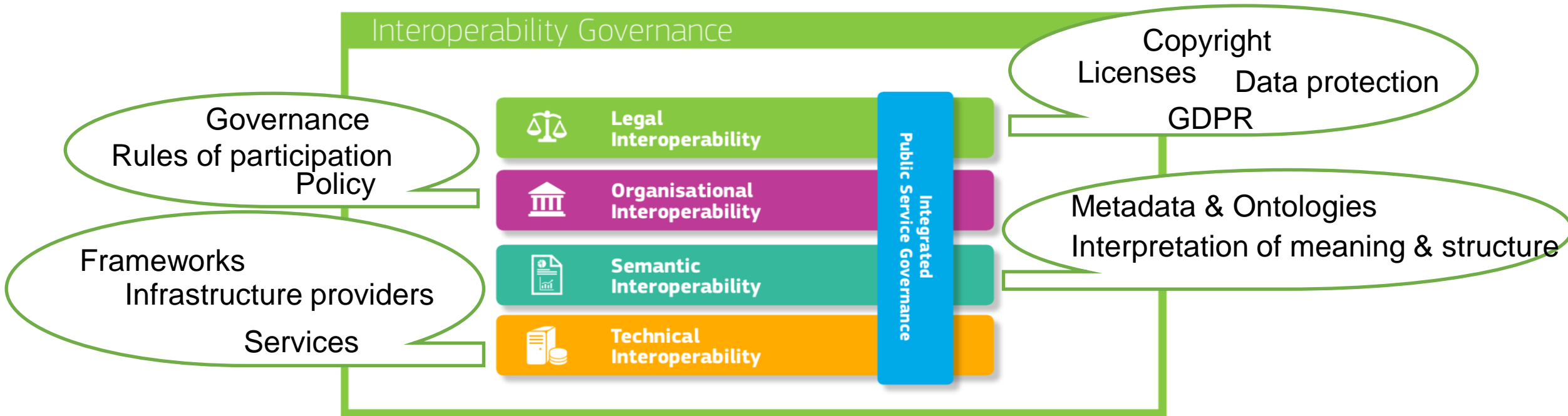
Source: https://commons.wikimedia.org/wiki/File:Islands_of_Data.svg

Source:

https://www.researchgate.net/publication/267692879_Towards_Executable_Reality_Business_Intelligence_on_Top_of_Linked_Data/figures?lo=1

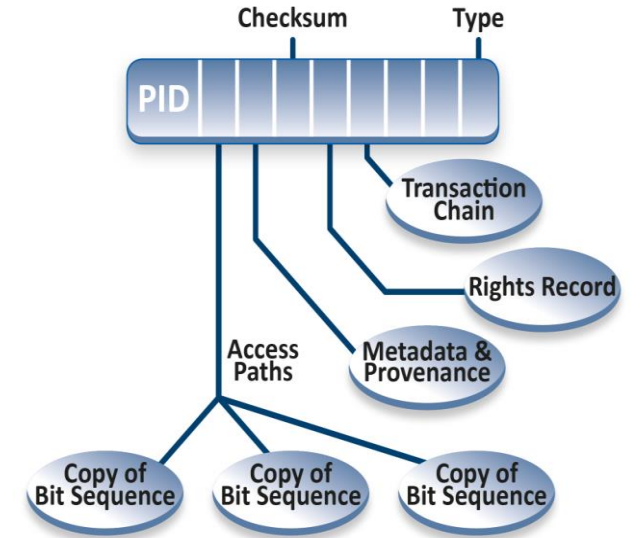
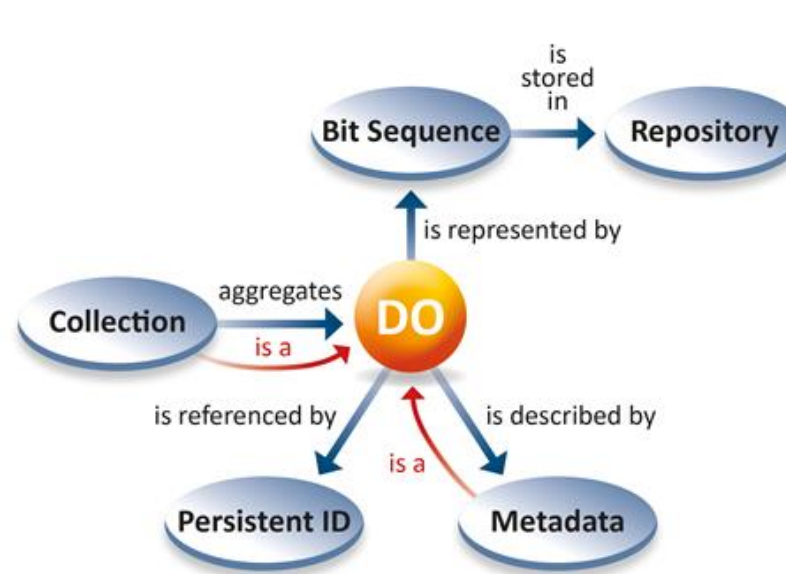
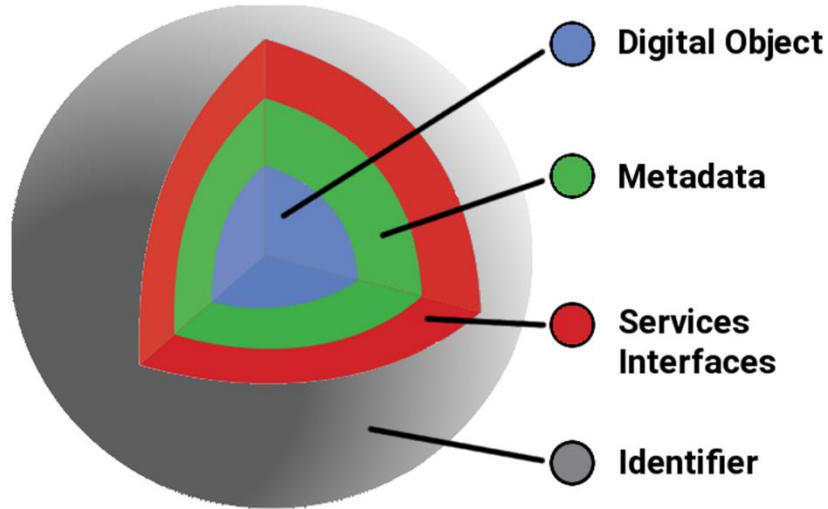
Source: Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi:10.1038/sdata.2016.18 (2016)

Layers of interoperability



Source: [The European Interoperability Framework four levels of interoperability](#)

FAIR digital object



[Digital Object Interface Protocol](#)

Source: RDA's Data Foundation & Terminology Group (DFT)
2014: Core Model

Source: Schwardmann, U., 2020. Digital Objects – FAIR Digital Objects: Which Services Are Required?. *Data Science Journal*, 19(1), p.15. DOI: <http://doi.org/10.5334/dsj-2020-015>

FDOs and semantic interoperability

1. FDOs should be described using standard, commonly accepted semantic artefacts, formats, and protocols. This allows for greater compatibility and ease of integration with other systems and tools.
2. FDOs should be stored in open, machine-readable formats that can be easily accessed and processed by both humans and machines.
3. Data should be made available through open application programming interfaces (APIs) that allow other systems and applications to access and integrate with it.
4. FDOs should be accompanied by detailed metadata that describes their content, structure, and context. Metadata should explain how they were created and can be reused and should also include information about any copyright or other restrictions on the use of the FDO. This helps to ensure that data can be easily understood and used by others.
5. FDOs should be assigned persistent, globally unique identifiers that can be used to reference and cite them. This enables FDOs to be easily discovered, cited, and reused by others.

Research data lifecycle

- Planning and searching for data sources.
- Collecting and creating.
- Processing and analysis.
- Publishing and sharing.
- Digital preservation.
- Re-use of data and other research results.

Planning and searching for data sources

- What will we research?
- What research data will we collect or generate?
- How will the research data be collected or generated?
 - Methods of collection, data quality, IPR of re-used data.
- Preparation of relevant documents:
 - Data management plan.
 - Preparation of informed consent form.
 - Application for ethical approval of the research study.

What the research data management plan should contain?

- Project basic metadata and description.
- What research data will you collect or generate?
- How will the research data be collected or generated?
- What documentation will adequately describe your research data?
- Choice of metadata and data standard.
- How will you address ethical issues?
- How will you deal with copyright and intellectual property rights (IP/IPR) issues?
- How will research data be stored and backed up during the research?
- How will you manage access and ensure the security of research data?

What the research data management plan should contain - continued?

- What research data has long-term value and should be stored, shared and/or preserved?
- Choosing an appropriate repository or data archive.
- How will you share the research data?
- Are there any restrictions on sharing research data?
- Who will be responsible for handling the research data?
- What resources will you need to implement your plan?
- How will potential users find your research data?
- How to make research data openly available?
- How to make your research data interoperable?
- How to ensure reuse of research data.

Sharing sensitive data

- Sensitive data that contain potentially identifying information -- whether it be human subject data or other types of sensitive data -- will likely need to be modified prior to sharing these data with the public. It is important that these modifications are made in order to protect participant confidentiality, the location of endangered wildlife, or for other relevant reasons. However, **these modifications may affect the data to the point where reproducibility or additional subsequent research by others is no longer possible.**
- You might consider retaining multiple versions of the data: one that is suitable for public release, and one that is suitable for further research but that is available on a highly restricted basis.

Types of identifying information

- **Direct identifiers:** These data point directly to an individual and are typically removed from data sets before sharing with the public. These may include: name, initials, mailing address, phone number, email address, unique identifying numbers, like Social Security numbers or driver's license numbers, vehicle identifiers, medical device identifiers, web or IP addresses, biometric data, photographs of the person, audio recordings, names of relatives, dates specific to individual, like date of birth, marriage, etc.
- **Indirect identifiers:** These may seem harmless on their own, but can point to an individual when combined with other data. It has been recommended that datasets containing three or more indirect identifiers should be reviewed by an independent researcher or ethics committee to evaluate identification risk. Any indirect information not needed for the analysis should be removed. It may be reasonable to supply some of these types of data in aggregated form (like ranges of annual incomes instead of exact numbers). Indirect identifiers may include: place of medical treatment or doctor's name, gender, rare disease or treatment, sensitive data like illicit drug use or other "risky behaviors,,, place of birth, socioeconomic data, like workplace, occupation, annual income, education, etc, general geographic indicators, like postal code of residence, household and family composition, ethnicity. birth year or age, verbatim responses or transcripts

Techniques for data desensitization

- **Data Masking:** This technique involves hiding sensitive data by replacing it with non-sensitive information, such as a random string of characters. This can be done for specific fields or for an entire dataset. For example, credit card numbers can be masked by showing only the last four digits.
- **Access Control:** Access control involves limiting who can view or access sensitive data. This can be done through user authentication, permissions management, or other security measures.
- **Data Encryption:** This technique involves converting data into a code that can only be read with a key or password. This makes the data unreadable to anyone without the proper authorization. Encryption can be applied to entire datasets or specific fields containing sensitive information.
- **Data Perturbation:** This technique involves modifying the data to make it less sensitive while still retaining its statistical characteristics. This can be done by adding random noise to the data or by rounding values. For example, a patient's birthdate can be perturbed by adding or subtracting a few days.
- **Data Aggregation:** This technique involves combining data from multiple sources to obscure specific details while still providing useful insights. For example, instead of providing individual sales figures for each store, the data can be aggregated to show total sales by region.
- **Pseudonymization:** Pseudonymization is the process of replacing personally identifiable information with a pseudonym, or a unique identifier that cannot be used to directly identify the individual.
- **Data Obfuscation:** This technique involves replacing sensitive data with nonsensitive data that still preserves the structure and format of the original data. For example, a person's name can be obfuscated by replacing it with a random name that has the same number of characters.
- **Data Redaction:** Redaction involves permanently removing sensitive information from a document or record, either by blacking it out or completely deleting it. This is often used in legal or regulatory contexts to protect privacy or security.

Sharing Sensitive Data with Confidence: The Datatags System

Tag Type	Description	Security Features	Access Credentials	
Blue	Public	Clear storage, Clear transmit	Open	Non-confidential information
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration	Non-confidential information
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA	Potentially harmful personal information
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA	Sensitive personal information
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA	Very sensitive personal information
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA	Maximum sensitive personal information

Informed consent

- Description of the study, methods of data collection, and description of the data collected.
- Obligations of the research participant.
- Scope of commitment and compensation for participation.
- Risks and dangers of participation.
- Procedure for withdrawing from participation in the study.
- Benefits of participation.
- Voluntary nature of participation.
- Protection of personal data, publication of results, archiving and sharing of data.
- Contact details of the researcher and funding sources.
- Additional consents and clarifications on GDPR if personal data of research participants are collected (protection of personal data, retention period of personal data, anonymisation, confidentiality).

Informed consent - continued

The informed consent form must be written in easy understandable language. **The research participant must be given a clear opportunity to make the following points:**

- that he/she has read and understood the information about the project,
- that he/she has had the opportunity to obtain additional information,
- that he/she voluntarily agrees to participate in the project,
- that he/she understands that he/she can stop participating without giving a reason and without penalty,
- that confidentiality procedures (use of names, pseudonyms, anonymisation of data, etc.) have been explained; and
- that the ways in which research data can be used for publication, sharing and archiving are explained.

The form should include signatures and signature dates for both the participant and the researchers.

The research participant should be informed that **after the project is completed, the research data will be stored and made available to other researchers for secondary analysis for research and educational purposes.**

Application for ethical approval

- Applicant's contact details.
- Description of the research (purpose, duration, description of the data collected, methods of collection, ethical aspects, benefits, risks and burdens of the participants involved, information on the financial compensation of the participants, safety of the participants, insurance in case of possible damage to health).
- The funder and the costs of the research.
- Statements by the principal investigator, and his bosses.
- Informed consent form for the participant.
- In the case of children and other vulnerable groups, a consent form from parents/guardians.
- Information regarding the collection, security and storage of personal data.
- Signature of the applicant.

Ethics and research data management



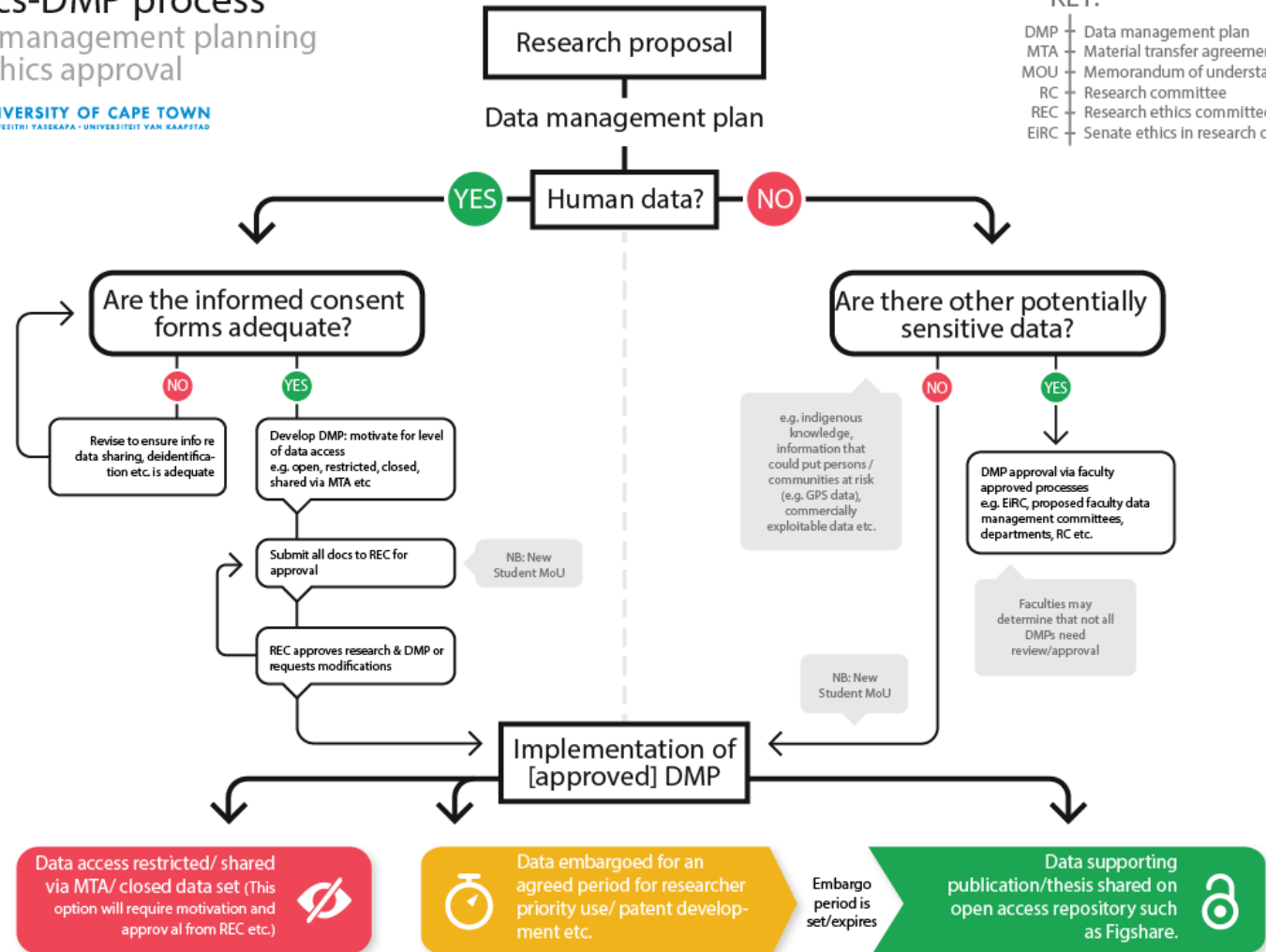
Ethics-DMP process Data management planning for ethics approval



KEY:

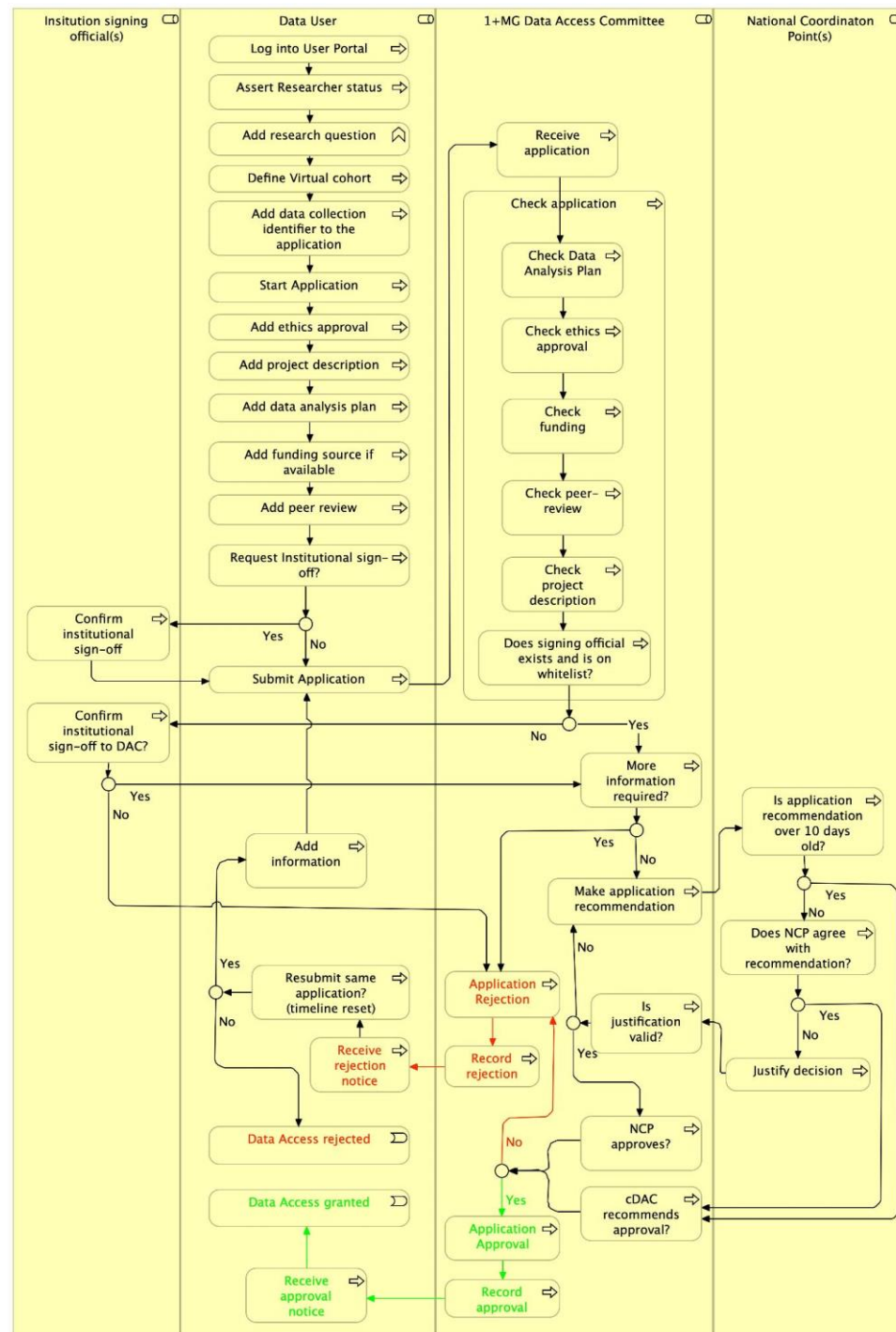
- DMP | Data management plan
- MTA | Material transfer agreement
- MOU | Memorandum of understanding
- RC | Research committee
- REC | Research ethics committee
- EIRC | Senate ethics in research committee

Source: Ethics and research data management.
Available at:
<http://www.researchsupport.uct.ac.za/ethics-and-research-data-management>





Example of a workflow for using genomic data in a Genomics Data Infrastructure project



Data processing and analysis

- Data insertion, digitisation, transcription, translation.
- Verification, validation, cleaning, anonymisation of data where necessary.
- Data description, structuring and documentation.
- Data analysis.
- Interpretation of results.
- Data management and storage (backups: [3-2-1 backup rule](#)).

Clear license information is important because...

It **tells users** and **reusers** exactly **what they can do** with your data and metadata.

It **encourages the use and reuse** of your data and metadata the way you want them to be used and reused.

It **creates visibility** of your efforts downstream (if you ask for attribution).


*If **no explicit licence** is provided, a user does not know what can be done with the data/metadata – **the default legal position is that nothing can be done without contacting the owner on a case-by-case basis.***


Four key tips when publishing information about the licence


1. Make sure your licensing information is easy to find.
2. Include information about the license in the metadata of each data set.
3. Use simple licenses to ensure they are easy to understand by re-users.
4. Check spelling, typos and spacing to ensure consistency in the names of the licenses used.

Clear licence information - example

Title: Short example of Biceps Brachii muscle sEMG decomposition using the DEMUSE Tool-TEST

Authors:  [Holoobar, Aleš](#), Faculty of Electrical Engineering and Computer Science, University of Maribor (Author)

Files:  [sEMG_SynthSig_BicepsBrachii_Holoobarv_1.0_dataset_overview.pdf](#) (60.07 KB)
 MD5: FCBB8E622B149D416191F81B7531C463

 This document has files, available only to logged in users with granted permissions. [Request access](#)

Language: English


Work type: Not categorized

Typology: 2.20 - Complete scientific database of research data

Organization: FERI - Faculty of Electrical Engineering and Computer Science

Abstract: This dataset contains 4 examples of simulated multichannel surface EMG signals of the Biceps Brachii muscle and results of their decomposition into separate motor unit activity. It is intended as a demonstration of the DEMUSE Tool software for sEMG decomposition and as a basis for practical example of dataset preparation for the HybridNeuro project webinar on Data management and ethics. Two sets of data are included: the raw simulated sEMG signals and the results of decomposition of those signals with the DEMUSE Tool.

Keywords: [surface EMG](#), [decomposition](#), [motor unit](#), [DEMUSE](#), [biceps brachii](#), [dataset](#)





PID: [20.500.12556/DKUM-88038](#) 

Data col. methods: Simulation


Publication date in DKUM: 05.04.2024

Views: 88

Downloads: 5

Metadata:    

Categories: [Misc.](#)

Cite this work : HOLOBAR, Aleš, no date, *Short example of Biceps Brachii muscle sEMG decomposition using the DEMUSE Tool-TEST* [online]. Complete scientific database of research data. [Accessed 8 April 2024]. Retrieved from: <https://dk.um.si/lzpisGradiva.php?lang=eng&id=88038>

[Copy citation](#)

Document is financed by a project

Funder: EC - European Commission
 Funding programme: HE
 Project number: 101079392
 Name: Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders
 Acronym: HybridNeuro

Funder: UKRI - UK Research and Innovation
 Funding programme: Innovate UK
 Project number: 10052152
 Name: Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders (HybridNeuro)

Licences

License: CC0 1.0, Creative Commons CC0 1.0 Universal



Link: <https://creativecommons.org/publicdomain/zero/1.0/deed.en>

Description: CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Licensing start date: 05.04.2024

Different data have different licensing needs

- Some data(sets) may be required to be **openly available** (e.g. subject to a Constitution of the Republic of Slovenia).
- Some data(sets) may be **subject to restrictions** (e.g. privacy, national security, third party rights).
- Some data(sets) may be **available for reuse but not for modification** (e.g. legal texts, public budgets (if modifications are made, it must be made clear that the data is not the actual authentic version)).
- Some data(sets) may be published **allowing derivations** with attribution of authoritative source (e.g. legal commentary, translations).

Licensing approaches: Creative Commons (1)



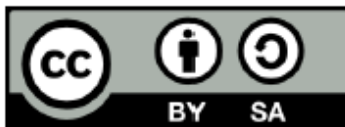
Public domain - No rights reserved – allows licensors to waive all rights and place a work in the public domain. others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.



Public Domain Mark – “No know copyright”– allows licensors to waive all rights and place a work in the public domain. It indicates that a work is no longer restricted by copyright and can be freely used by others.



Attribution – Others can distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation.



Attribution-ShareAlike – Others can remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms.

Licensing approaches: Creative Commons (2)



Attribution-NoDerivs – Allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.



Attribution-NonCommercial – Others can remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.



Attribution-NonCommercial-ShareAlike – Others can remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.



Attribution-NonCommercial-NoDerivs – Only allows others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially.

Good practices for licensing your data

Good practices:

If the original data is in the public domain (e.g. by law), keep it there – use for example the Creative Commons Zero Public Domain Dedication or the Open Data Commons Public Domain Dedication and License (PDDL)

For some documentation integrity needs to be protected – use a No- Derivatives licence, for example Creative Commons Attribution- NoDerivs, but only if really necessary

Avoid Non-Commercial licences if at all possible, as these seriously restrict reuse.

Licenses for data should provide appropriate security and control (but not more than that).

Using an open and unrestricted license for your data



Whenever data is licensed for open and unrestricted access, reusers can create new knowledge from combining it.

For example:

Cross-referencing public spending with geographic data to visualise which regions are better funded.

Matching public transport timetables with GPS data to be able to give real time information on delays.

Measuring performance of public services based on transaction counters and waiting times.

Deriving recommendations for prevention policies relating accident statistics with weather data and road maps.

Good practices for licensing your metadata

What you need to think about:

- Metadata helps people to discover your data.
- The wider your metadata is distributed, the higher your visibility is.
- Others may want to add to it, enhance it, link to other resources.

Good practices:

- Licences for metadata should be as open as possible.
- A public domain licence allows the widest reuse.
- An attribution licence ensures you get credit downstream, but may cause problems if data is shared multiple times (attribution stacking).

Publishing and sharing data

- Resolution of copyright,
- preparation of metadata for »searching« and documentation for users,
- publication and sharing of data (possibly in a data archive),
- control over data access,
- promotion of data.

- Which data has long-term value and should be kept?
- Transfer data into a format that is independent of current technology.
- Ensuring data archival according to the [3-2-1 backup rule](#).
- What data needs to be stored or destroyed due to contractual relationships with data providers?
- How long should the data be stored or retained?
- How will access and security be ensured?

Citing data

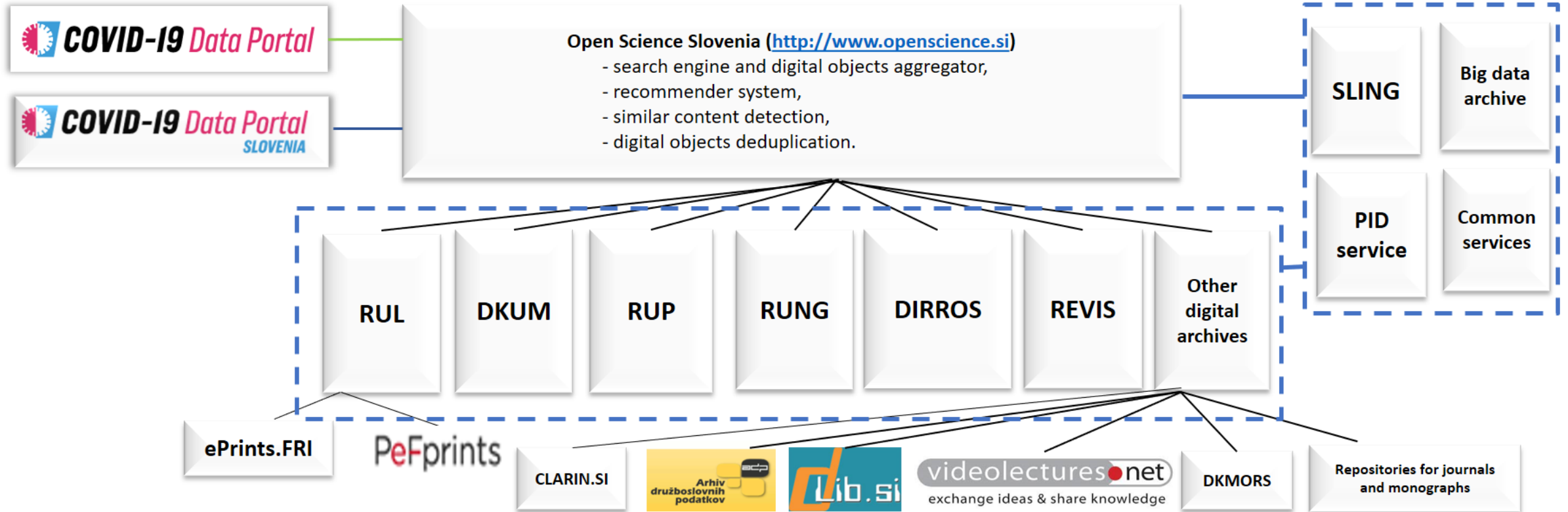
The citation of the data source in the publication should generally include the following information:

- the author(s) of the research data,
- the name of the dataset,
- the producer (the institution under whose auspices the data were produced),
- the place and year of production of the dataset,
- the version of the dataset,
- distribution,
- series,
- place and year of publication of the research data in the data repository,
- the data repository where the data can be accessed,
- persistent identifier.

Links related to the research data management

- [CESSDA Data Management Expert Guide](#)
- [Data Management Skillbuilding Hub](#)
- [UK data service YouTube channel](#)
- [Funders' data plan requirements](#)
- [DMPOnline](#)
- [Checklist for a Data Management Plan](#)
- [Consent for data sharing](#)
- [DMP Assistant](#)

Structure diagram of Slovenian open access infrastructure



Slovenian COVID 19 national portal is available on
<http://covid19dataportal.si/>

Establishing processes to support the handling of research data in the national open access infrastructure

- Pre-publication activities.
- Publication in the repository or data archive.
- Digital preservation.

Phase before publication of research data and required documentation



Phase before publication of research data:

- Planning and finding data sources.
- Preparation of a research data management plan, applications for the ethics commission and proposals for informed consent, proposals for declarations by data providers.
- Obtaining relevant statements and opinions.
- Data collection and creation.
- Data Processing and analysis.
- Preparation of files in appropriate formats.
- Preparation of documentation.

Before a researcher applies for the publication of a research dataset in the national open access infrastructure, he must have:

- a data management plan (if requested by the funder or the organization in which he is employed),
- metadata about the research dataset,
- documentation that is necessary for understanding and using the data,
- data files in appropriate formats,
- ethical approval if the research study involves humans, animals or environmental data,
- statements of data providers and signed informed consents of research participants,
- defined licenses for the use of research data,
- the software, containers, workflows that was used to generate or process the data, if he created it himself,
- research notes and other research results, if any.

Publication phase

- The researcher inserts the research data set and other research results into the repository or data archive himself or his librarian inserts them .
- The librarian checks the adequacy of the metadata and whether the appropriate documentation is available.
- The librarian informs the appropriate authority within the institution, which is in charge of checking the appropriateness of data publication and other research results, that the data set and other research results have been uploaded. They are accessible in closed access and are only available via a link that requires a password provided by the librarian.
- The appropriate body within the institution, which is in charge of checking the adequacy of the data publication, checks the adequacy of the content of the data set and other research results. If the content is appropriate, inform the librarian that the data set and other research results can be published.
- The librarian, after a positive response from the body within the institution, which is in charge of checking the appropriateness of data publication, publishes the data set and other research results in the repository and performs cataloging in COBISS.
- Central specialised information centre of the scientific field, established by Slovenian research and innovation agency checks the adequacy of the typology, metadata and documentation of the research data set and other research results.

Digital preservation phase

Data can be stored in different formats and in several versions. For digital preservation of research data, we must ensure the independence of the data from the technology. We will establish processes for digital preservation according to the OAIS reference model ([ISO 14721](#)) using metadata according to [the PREMIS](#) meta data standard.



Practical tips for FAIR dataset preparation

Matjaž Divjak

University of Maribor, Faculty of Electrical Engineering and Computer Science

matjaz.divjak@um.si



GA No. 101079392



GA No. 10052152



University of Maribor



CHALMERS
UNIVERSITY OF TECHNOLOGY

Imperial College
London

Elements of a dataset

- Data:
 - in original / RAW form
 - the processed data: possibly several versions
 - arranged in folders / collections and clearly labeled
- README file: a plain text file with detailed description of everything related to the dataset
- DOI: unique identifier, included in README
- Licence info: included in README
- Metadata: structured metadata description, usually entered during upload to repository or supplied in a separate XML/JSON file

Elements of a dataset: our example

- Example dataset:
 - synthetic surface EMG of biceps brachii muscle: 10 x 9 electrode grid, 500 MUs
 - 4 x RAW sEMG at 10, 30, 50, 70 % MVC
 - 4 x results of decomposition with DEMUSE Tool + manual cleaning

Name
└─ sEMG_SynthSig_BicepsBrachii_Holobar_v1.0.0
├─ decomposed
├─ raw_signals
└─ metadata.xml
└─ README.txt

Name	Size
MAT_file_variables_rawSignals.png	33,5 KiB
sEMG_SynthSig_BicepsBrachii_F10MVC_Len20_SNRInf_03-Apr-2024_rawSignals.mat	57,5 MiB
sEMG_SynthSig_BicepsBrachii_F30MVC_Len20_SNRInf_03-Apr-2024_rawSignals.mat	57,8 MiB
sEMG_SynthSig_BicepsBrachii_F50MVC_Len20_SNRInf_03-Apr-2024_rawSignals.mat	57,9 MiB
sEMG_SynthSig_BicepsBrachii_F70MVC_Len20_SNRInf_03-Apr-2024_rawSignals.mat	58,0 MiB

Name	Size
DEMUSE_Tool_parameters.png	71,8 KiB
MAT_file_variables_DEMUSE_edited.png	89,0 KiB
sEMG_SynthSig_BicepsBrachii_F10MVC_Len20_SNRInf_03-Apr-2024_DEMUSE_edited.mat	43,9 MiB
sEMG_SynthSig_BicepsBrachii_F30MVC_Len20_SNRInf_03-Apr-2024_DEMUSE_edited.mat	37,9 MiB
sEMG_SynthSig_BicepsBrachii_F50MVC_Len20_SNRInf_03-Apr-2024_DEMUSE_edited.mat	34,8 MiB
sEMG_SynthSig_BicepsBrachii_F70MVC_Len20_SNRInf_03-Apr-2024_DEMUSE_edited.mat	34,7 MiB

Suggested file formats

- Use open, public formats as much as possible
- Good guide: <https://guides.library.cornell.edu/ecommons/formats>

Data type	Recommended format	Acceptable format
Text	.txt, .rtf, .pdf, .html, .md, .odt	.doc, .xml
Table	.csv, .tab, .par, .xml, .hdf	.txt, .xls, .dbf, .ods, .sav, .dta
Images	.tif, .jpg, .png	
Sound	.flac	.wav, .mp3, .aif
Video	.mp4, .ogv, .ogg, .mj2	.avchd

Data in proprietary formats

- Describe data storage format in README
- If possible, include code / links for reading / writing files
- Our example:
 - .TXT for README
 - .MAT for signals + links to reader/writer software (Octave, Python, C)
 - .PNG for images
 - .XML for metadata

File naming conventions

- Include most important parameters in filenames
- Use underscores instead of spaces
- Our example:

```
sEMG_SynthSig_BicepsBrachii_F10MVC_Len20_SNRInf_ ...  
..._03-Apr-2024_rawSignal.mat
```

README.TXT file

- Should contain:
 - general information: authors, affiliation, date, license, funding info, ...
 - access information: DOI, links to repositories, citation info
 - dataset contents overview: short description, list of files, data size, ...
 - methods description: how data was generated / collected, parameters used, ...
 - description and links to used software
 - detailed description of data formats, variables, ...
- Good example: <https://cornell.app.box.com/v/ReadmeTemplate>
- Our example: README.txt

Data sharing licenses



- Use **CC0** (Public Domain Dedication) or **CC BY** (attribution license)
- For datasets, try to avoid attribution license, because it can lead to attribution creep
- You can still ask for attribution, not as a legal requirement but as “please attribute my data” in line with scientific norms
- Provide a citation for the dataset that others can copy and paste with ease
- Guide: <https://www.openaire.eu/how-do-i-license-my-research-data>
- Our example: CC0 Public Domain license, stated in the README + selected in the repository

Unique identifiers - DOI



- Usually generated by the repository, before the final submission
- Should be included in the README file
- Several different identifiers can be used: DOI, Orcid, Arxiv.org, ...
- Our example: **doi:10.5281/zenodo.10936952**

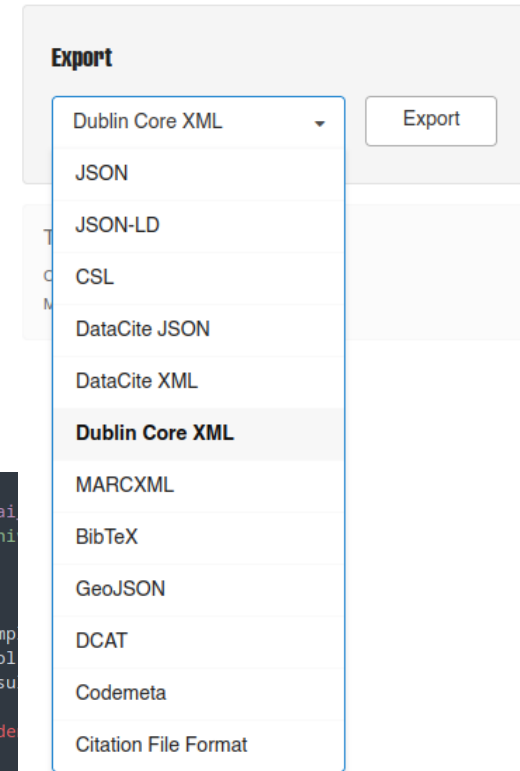
Basic metadata standards

- General purpose:
 - Dublin Core: general metadata standard for describing wide range of resources
 - schema.org: semantic vocabulary of tags for web pages
- Repositories often use their own standard, generally a subset of the Dublin Core metadata
- Domain Specific standards:
 - DCC list: <https://www.dcc.ac.uk/guidance/standards/metadata/list?page=2>
 - Carnegie Mellon Metadata guide: <https://guides.library.cmu.edu/c.php?g=472661&p=9230176>
 - No specific standard for EMG / EEG data?
- Our example: use repository standards and Dublin Core

Dublin Core (DC) metadata standard



- Most used general standard for describing basic dataset info
- Most repositories require entering a subset of DC elements
- Can usually be generated automatically by the repository from the entered data
- If your repository doesn't require metadata, include DC descriptions in a separate .XML file
- Our example: metadata entered during submission + included in metadata.xml



```
<?xml version='1.0' encoding='utf-8'?>
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:oai
http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchi
<dc:contributor>Divjak, Matjaž</dc:contributor>
<dc:creator>Holobar, Aleš</dc:creator>
<dc:date>2024-04-06</dc:date>
<dc:description>&amp;lt;p&amp;gt;This dataset contains 4 examp
activity. It is intended as a demonstration of the DEMUSE Tool
management and ethics (https://www.hybridneuro.feri.um.si/resu
DEMUSE Tool.&amp;lt;p&amp;gt;</dc:description>
<dc:identifier>https://doi.org/10.5281/zenodo.10936952</dc:ide
<dc:language>eng</dc:language>
<dc:publisher>Zenodo</dc:publisher>
<dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
<dc:rights>Creative Commons Zero v1.0 Universal</dc:rights>
<dc:rights>https://creativecommons.org/publicdomain/zero/1.0/legalcode</dc:rights>
<dc:subject>surface high density electromyogram</dc:subject>
<dc:subject>HDEMG</dc:subject>
<dc:subject>decomposition</dc:subject>
<dc:subject>motor unit</dc:subject>
<dc:subject>DEMUSE</dc:subject>
<dc:subject>simulation</dc:subject>
<dc:subject>biceps brachii</dc:subject>
<dc:subject>dataset</dc:subject>
<dc:title>Short example of Biceps Brachii muscle surface HDEMG decomposition using the DEMUSE Tool</dc:title>
<dc:type>info:eu-repo/semantics/other</dc:type>
</oai_dc:dc>
```


Dublin Core metadata standard



- Overview of the main 15 elements:

Title	Main title of the data	Type	Nature or genre of the data: text, image...
Creator	Creator(s) of the data	Format	File format, medium, physical size
Subject	Topic of the data, keywords	Rights	Usage rights or license
Description	Short description of the data	Source	Related resource from which the data is derived
Publisher	Entity responsible for making data available	Language	Language of the data
Contributor	Other contributors to the data	Relation	Related resources
Coverage	Location and time covered by the data	Identifier	Unique identifier, DOI
Date	Time period associated with data		

Usage guide:

<https://www.dublincore.org/specifications/dublin-core/usageguide/elements/>

Metadata for YouTube videos

- Important for better visibility / findability of your videos
- Available fields:
 - title: 100 chars
 - description: 5000 chars
 - tags / keywords: 500 chars
 - various other details: thumbnail, language, date, for children, age restriction, promotion, ...
- Suggestion: include additional info in Description: list of keywords, links to projects, funding info, ...

Video details

Title (required) ?
Hybrid Neuro project promotional video

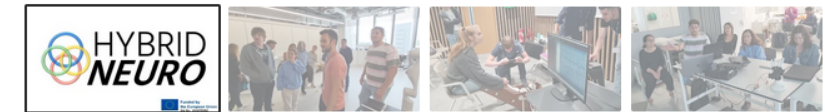
Description ?
A short promotional video for the HybridNeuro project that presents several examples of our various activities.

The HybridNeuro project combines the expertise of leading European partners in the field of Neural Interfaces to set up new pathways of analyzing human motor system and human movements and transfer the academic research into clinical and industrial practice. Link: <https://www.hybridneuro.feri.um.si/>

This project has received funding from the Horizon Europe Research and Innovation Programme under GA No. 101079392, as well as UK Research and Innovation organisation (GA No. 10052152).

Thumbnail

Select or upload a picture that shows what's in your video. A good thumbnail stands out and draws viewers' attention. [Learn more](#)



Playlists

Add your video to one or more playlists to organize your content for viewers. [Learn more](#)

Select ▼

Audience

This video is set to not made for kids Set by you

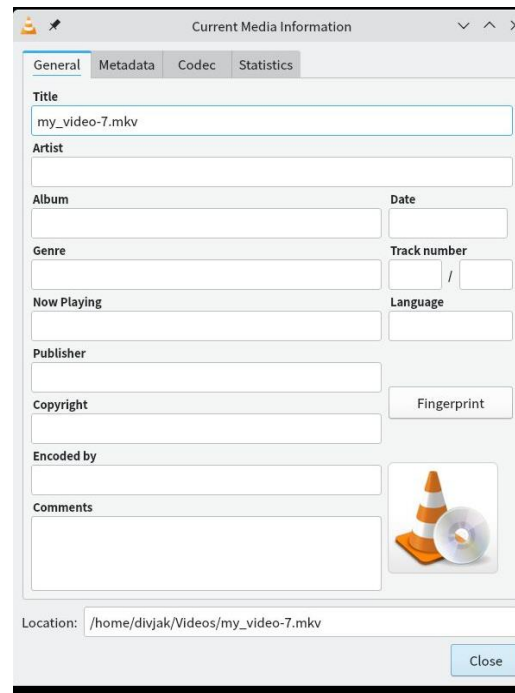
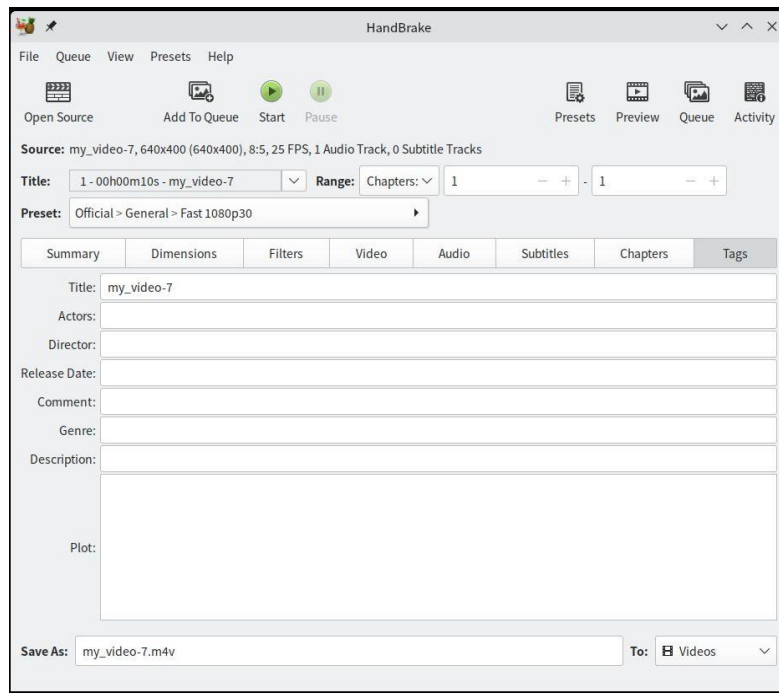
Regardless of your location, you're legally required to comply with the Children's Online Privacy Protection Act (COPPA) and/or other laws. You're required to tell us whether your videos are made for kids. [What's content made for kids?](#)

i Features like personalized ads and notifications won't be available on videos made for kids. Videos that are set as made for kids by you are more likely to be recommended alongside other kids' videos. [Learn more](#)

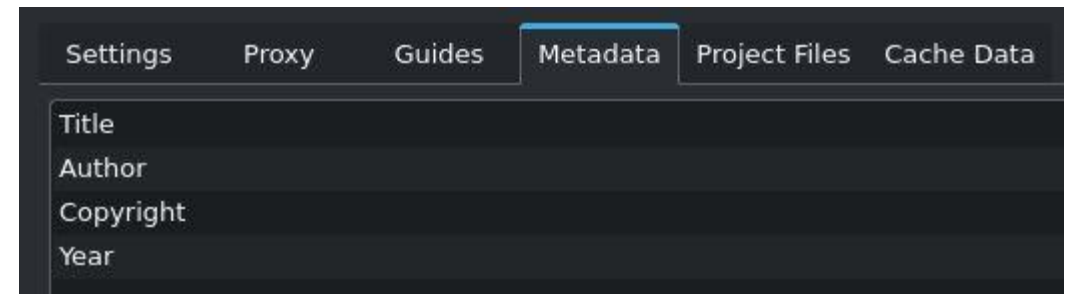
- Yes, it's made for kids
- No, it's not made for kids

Metadata for non-YouTube videos

- Many video formats (including MP4) support basic metadata:
 - added during video creation using the video editor
 - added to any video using tools such as Handbrake, VLC, ...



VLC video player



Kdenlive video editor

More complex metadata for EEG/EMG



- More thorough description of data: using ontologies / semantic artefacts
- Most repositories don't support ontological metadata
- Could be interesting and useful in the future
- Searching through ontologies:
 - BioPortal: repository of biomedical ontologies: <https://bioportal.bioontology.org/>
 - OntoBee: linked data server for ontologies: <https://ontobee.org/>
 - Ontology Lookup Service (OLS): <https://www.ebi.ac.uk/ols4/>
- A few interesting vocabularies:
 - SNOMEDCT: <https://www.nlm.nih.gov/healthit/snomedct/index.html>: includes concepts: muscle, motor unit, motor unit firing rate, muscle motor unit potential, alpha wave, millivolt, ...
 - Ontology of Consumer Health Vocabulary (OCHV): <https://bioportal.bioontology.org/ontologies/OCHV>: includes concepts: muscle, motor unit, action potential, brain wave, millivolt, ...
 - Unified Medical Language System (UMLS): <https://www.nlm.nih.gov/research/umls/index.html>

Data paper



- Another option for detailed description of a dataset: data paper published in a data journal
- Open Data Journals:
<https://www.fosteropenscience.eu/foster-taxonomy/open-data-journals>
- Good paper on data journals:
<https://insights.uksg.org/articles/10.1629/uksg.510>

Description: The lichen herbarium was started in 1985 and currently includes ca. 17,000 samples, collected mainly by D. Puntillo, in various parts of Italy, especially in Calabria. Several groups have been revised by specialists.

Column label	Column description
occurrenceID	A unique identifier for each occurrence record in the dataset.
type	The nature of the occurrence record.
language	Language used for the resource.
licence	Terms under which the dataset is made available.
institutionID	Unique identifier for the institution holding the specimens.
institutionCode	Acronym representing the institution.
datasetName	Title of the dataset.
basisOfRecord	The basis on which the record is made.
recordedBy	Individuals responsible for creating the occurrence record.
eventDate	Date on which the occurrence was recorded.
continent	Name of the continent where the occurrence was recorded.
country	Name of the country where the occurrence was recorded.
countryCode	Standardised code representing the country.
locality	Locality where the occurrence was recorded.

<https://bdj.pensoft.net/article/116965/>

Sampling methods

Description: Most specimens (61% of the total) were collected in Calabria, although there are also several specimens from other parts of Italy and exsiccata from international herbaria.

Sampling description: Specimen labels were digitised in a spreadsheet and standardised to comply with the Darwin Core (Wieczorek et al. 2012). Subsequently, the data were imported into a MySQL database and published on ITALIC (Nimis 2023) and GBIF (2023).

Due to the absence of geographic coordinates in the specimen labels, localities were georeferenced a posteriori (only Italian localities) combining Google Maps and QGIS (2023). The point-radius method was employed to determine both the coordinates and the associated uncertainty, adhering to the best georeferencing practices by Chapman and Wieczorek (2020). To enhance the precision of the georeferencing in Calabria, where the majority of samples have been collected, regional maps sourced from the Geoportale della regione Calabria (2023) have been consulted.

Quality control: The dataset includes specimens from taxonomically critical groups. To ensure the quality of the data, specimens were sent to specialists who revised the identification. The scientific names originally written on the specimen labels have been transcribed in the verbatimIdentification field. The currently accepted names, aligned with the most recent version of the Checklist of the Lichens of Italy (Nimis 2016) using the name match tool in ITALIC (Martellos et al. 2023), were reported in the scientificName field.

Geographic coverage

Description: The dataset contains 15219 georeferenced records (90% of the total) (Puntillo et al. 2023). Of these records, 10254 specimens were collected in Calabria while 4965 in other administrative regions of Italy: Campania (1959), Sicilia (829), Basilicata (582), Toscana (267), Friuli Venezia Giulia (263), Sardegna (249), Lombardia (189), Puglia (177), Lazio (124), Umbria (86), Valle d'Aosta (71), Veneto (51), Trentino-Alto Adige (31), Emilia-Romagna (26), Piemonte (18), Abruzzo (17), Marche (13) and Molise (13).

Trustworthy Digital Repositories



- In order of preference:
 - disciplinary / domain-specific: should be first choice, if they exist
 - institutional / national: usually only available to local organization members
 - general purpose: accept the widest range of data, no standard metadata scheme
- Search for repository:
 - Registry of research data repositories: <https://www.re3data.org/>
 - OpenDOAR: <https://www.jisc.ac.uk/opensoar>
- OpenAIRE guide: <https://www.openaire.eu/find-trustworthy-data-repository>
- Our example:
 - No domain-specific repository?
 - Institutional: Digital Library of University of Maribor (DKUM): <https://dk.um.si/>
 - General purpose: ZENODO: <https://zenodo.org/>

- Certified trustworthy national repository
- Available only to University of Maribor members



The screenshot shows the digital library interface with a navigation bar (INTRODUCTION, SEARCH, BROWSING, UPLOAD DOCUMENT, STATISTICS, LOGIN) and a search bar. The main content area displays document details for 'Short example of Biceps Brachii muscle sEMG decomposition using the DEMUSE Tool-TEST' by Holoobar, Aleš. It includes file information (sEMG_SynthSig_BicepsBrachii_Holoobarv_1.0_dataset_overview.w.pdf), abstract, keywords, and a citation box.

Document is financed by a project

Funder: EC - European Commission
Funding programme: HE
Project number: 101079392
Name: Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders
Acronym: HybridNeuro

Funder: UKRI - UK Research and Innovation
Funding programme: Innovate UK
Project number: 10052152
Name: Hybrid neuroscience based on cerebral and muscular information for motor rehabilitation and neuromuscular disorders (HybridNeuro)

Licences

License: CC0 1.0, Creative Commons CC0 1.0 Universal



Link: <https://creativecommons.org/publicdomain/zero/1.0/deed.en>

Description: CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Licensing start date: 05.04.2024

ZENODO repository



- General purpose
- Long standing repository with large user base, maintained by CERN
- No certification, but use encouraged by the European Commission
- Uses DataCite metadata
- Our example:
<https://zenodo.org/records/10936952>

A screenshot of the Zenodo repository interface. The top navigation bar is blue with the Zenodo logo, a search bar, and links for "Communities" and "My dashboard". Below the navigation bar, there is a prompt to "Select the community where you want to submit your record." The main content area shows a "Files" section with a table of files. The table has columns for "Preview", "Filename", "Size", and "Progress". The files listed include MAT files, sEMG files, a README.txt, and a metadata.xml. To the right of the file list is a "Draft" panel with buttons for "Save draft", "Preview", "Publish", and "Delete". There is also a "Visibility" section with "Public" and "Restricted" options, and an "Options" section with an "Apply an embargo" button.

Webinar materials



- All materials used in this webinar will be made publicly available:
 - video recording: YouTube
 - presentation slides: DKUM and Zenodo
 - dataset: DKUM and Zenodo
- More details on the HybridNeuro homepage:
<https://www.hybridneuro.feri.um.si/>
- QUESTIONS?