

Importance of metadata and FAIR

Improving data quality in biodiversity research



A Game of Trust

Do you trust publicly available reference genomes?

Why not?

What defines a well described genome?

Do you know?

Do all genomes deserve the label “reference”?

Why not?

Metadata matters!

Background

For biodiversity reference genome research, how do we...

... ensure high quality samples and data?

... increase FAIRness of data (including reusability of the genome)?

... maintain public/scientific trust in results?

FAIR?

FAIR

- To be useful for others data should be:
 - **FAIR** - Findable, Accessible, Interoperable, and Reusable
... for both Machines and Humans

Wilkinson, Mark et al. “The FAIR Guiding Principles for scientific data management and stewardship”.
Scientific Data 3, Article number: 160018 (2016) <http://dx.doi.org/10.1038/sdata.2016.18>



DOI: 10.1038/sdata.2016.18

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

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 Reusable:

- R1. (meta)data are richly described with 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 detailed provenance
- R1.3. (meta)data meet domain-relevant community standards



The FAIR guiding principles

*To be **Findable**:*

- F1.** (meta)data are assigned a globally unique and persistent identifier
- F2.** data are described with rich metadata (defined by R1 below)
- F3.** metadata clearly and explicitly include the identifier of the data it describes
- F4.** (meta)data are registered or indexed in a searchable resource

*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 **Reusable**:*

- R1.** meta(data) are richly described with 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 detailed provenance
 - R1.3.** (meta)data meet domain-relevant community standards

Findable, Accessible, Interoperable & Reusable

More value to publicly
funded research

Improve peer-review process

High research
integrity

Better research output

Background: 'FAIR Principles'
by Martínez-Lavanchy, et al (2019),
CC-BY 4.0. doi:10.11581/dtu:00000049



Findable, Accessible, Interoperable & Reusable

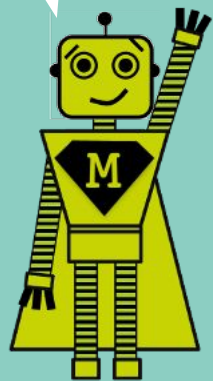
More value to publicly
funded research

Improve peer-review process

Works with
my software

High research
integrity

Better research output



Background: 'FAIR Principles'
by Martínez-Lavanchy, et al (2019),
CC-BY 4.0. doi:10.11581/dtu:00000049

FAIR in the context of other principles

- **Open Science** – barrier free, open access to research outputs and transparency in the research cycle
- **FAIR** – Findable, Accessible, Interoperable & Reusable, “*as open as possible, as closed as necessary*” (for people and for software)
- **Reproducibility** – results can be reproduced using the data, code, and documentation provided

Reusable for every
possible use case
now & in future

← Good enough!
(for now?)

Use once,
throw away



Towards a web of FAIR data and services

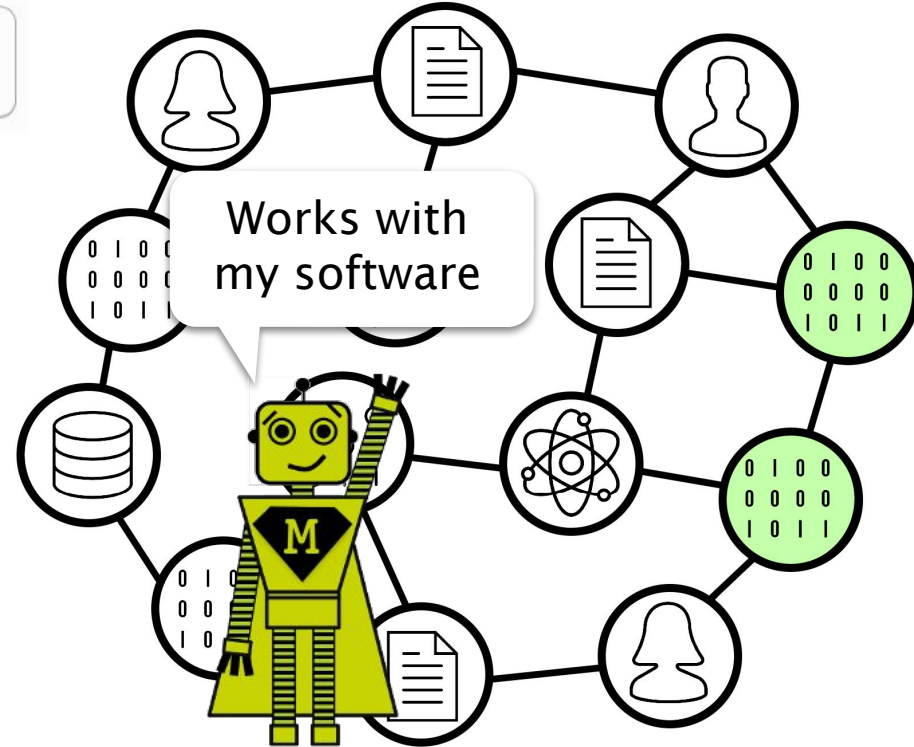
Search for Datasets



“From Gutenberg to Berners-Lee”

- Where will people be looking?
- Who should have access?
- Which standards/specifications?
- What else will be expected?
(Rich metadata!)

Graph: “PID Graph” from the FREYA project



Robot: MetaManMachine by Nikola Vasiljevic (2021),
CC BY-SA 4.0, doi:10.5281/zenodo.4471098

The FAIR guiding principles

To be Findable:

- F1.** (meta)data are assigned a globally unique and persistent identifier
- F2.** data are described with rich metadata (defined by R1 below)
- F3.** metadata clearly and explicitly include the identifier of the data it describes
- F4.** (meta)data are registered or indexed in a searchable resource

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

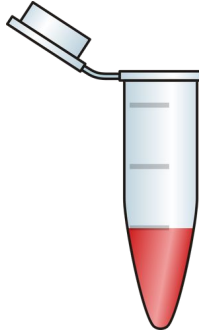
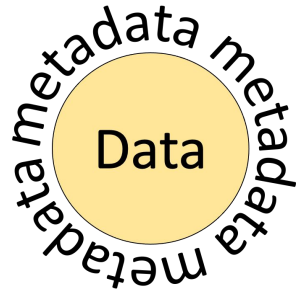
To be Reusable:

- R1.** meta(data) are richly described with 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 detailed provenance
 - R1.3.** (meta)data meet domain-relevant community standards

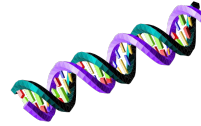
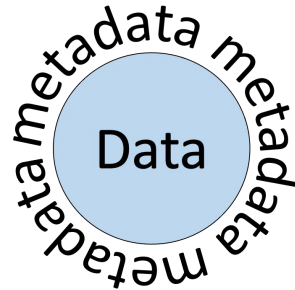
What is metadata?



Source: [Openclipart](#)

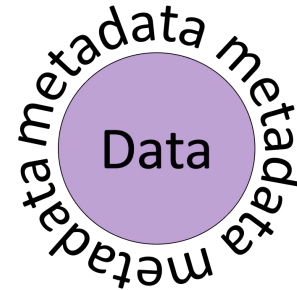


Source: [Openclipart](#)



.fastq

Source: [Openclipart](#)



Source: [Publicdomainpictures](#)

FAIR by design



Study & data
design

Sampling
& specimen
collection

Sample
preparation

Sample analysis
& data generation

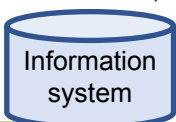
Data processing
to prepare inputs
for analysis

Data
analysis

Communicating
results

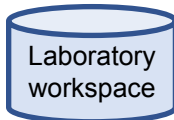
Procedures

data protection,
ethics permit,
infrastructure,
standards,
protocols,
data dictionaries,
data access, ...



Biosamples and instruments

populations (statistical) and inclusion criteria,
physical processing steps,
working storage conditions,
long-term storage location,
sample quality assessment,
sample annotations,
reagents, ...



Data and computational workflows

digital processing steps,
working storage conditions,
long-term storage location,
data quality assessment,
sample/data annotations,
reference data, ...



Outputs

publications,
data,
tools,
workflows,
reports,
dashboards, ...



Making sense of your group's data

“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”



How about another group's data?

Making sense of another group's data

“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

“ data should be assessable so that judgments can be made about their reliability and the competence of those who created them”.

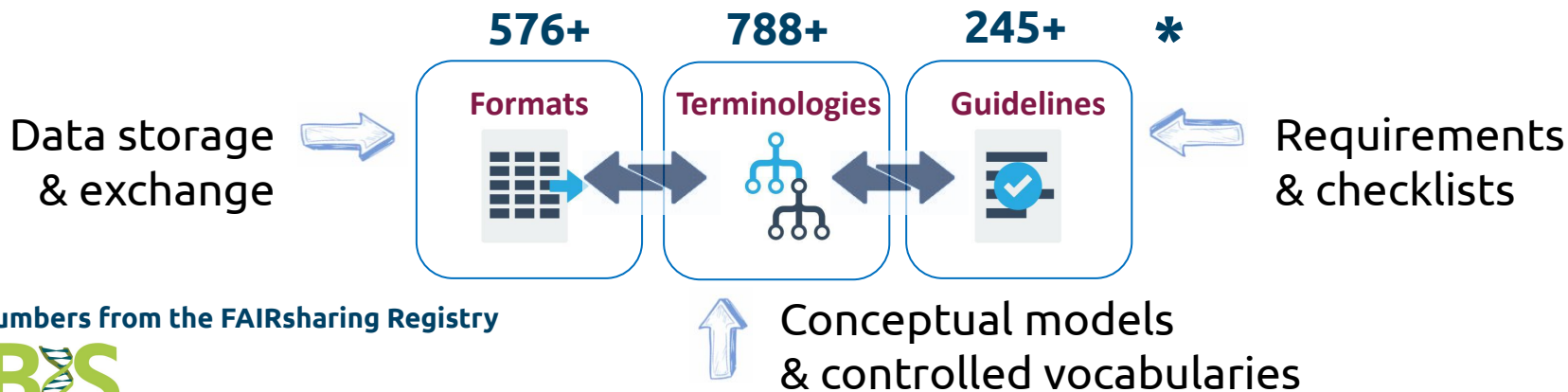


How about FAIR research data?

Making sense of FAIR research data

“ data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”

“ data should be assessable so that judgments can be made about their reliability and the competence of those who created them”.



The beauty of formalised metadata

Ok, but what is that?

- ✓ Describe the same thing! (i.e. Contextual terms)
- ✓ Use the same words! (i.e. Recommended terminologies)
- ✓ With good levels of detail! (i.e. Metadata templates)



Source: [Openclipart](#)

Repositories are your friends!



Controlled Vocabularies & Ontologies

Controlled Vocabularies

- A predefined, organised list of authorised **terms**
- Ensures consistency in data description
- *Examples:* Species taxonomies, Medical Subject Headings (MESH)

Ontologies

- A structured framework showing relationships **between concepts/terms**
- Supports complex data integration and interoperability

Benefits as Metadata Descriptors

- *Improved Data Consistency:* Reduces ambiguity and enhances searchability
- *Enhanced Interoperability:* Facilitates data sharing across systems
- *Efficient Data Management:* Streamlines classification and retrieval processes

Biodiversity

The variety of all native living organisms and their various forms and interrelationships.

Year introduced: 2004

Date introduced: July 9, 2003

PubMed search builder options

- ☐ Restrict to MeSH Major Topic.
- ☐ Do not include MeSH terms found below this term in the MeSH hierarchy.

Tree Number(s): G16.500.275.157.049, N06.230.124.049

MeSH Unique ID: D044822

Entry Terms:

- Biological Diversity
- Diversity, Biological

[All MeSH Categories](#)

[Phenomena and Processes Category](#)

[Biological Phenomena](#)

[Ecological and Environmental Phenomena](#)

[Environment](#)

[Ecosystem](#)

Biodiversity

[Biota](#)

[Microbiota](#) +

[Ecotype](#)

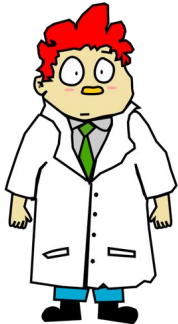
[Endangered Species](#)

[Introduced Species](#)

Metadata transformation

A simple example:

	Place	
	Strängnäs	



One location
=
One metadata value

Does anyone need to
know more?

Metadata transformation

Yes!

**More detail let others
(and your future self)
know what you have
done**

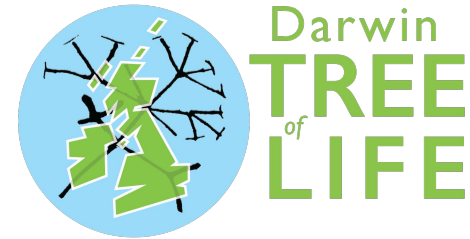
	Geographic location (country and/or sea)	Geographic location (region and locality)	Geographic location (latitude)	Geographic location (longitude)	
	Sweden	Strängnäs	59.29	17.12	

Rich metadata is the key to scientific reliability!

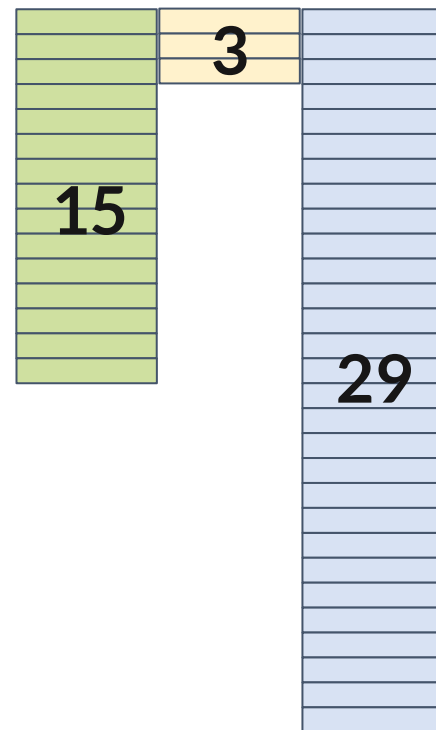
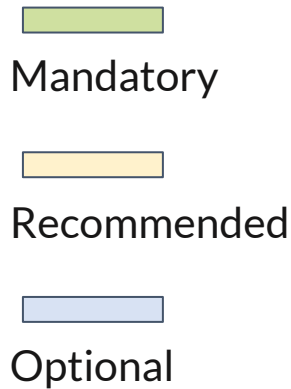
What defines high quality metadata?

Biodiversity community standards

- Rich in details
- Subject specific
- Precise
- Reliable
- Persistent over time



Tree of Life (ERC000053)



Default

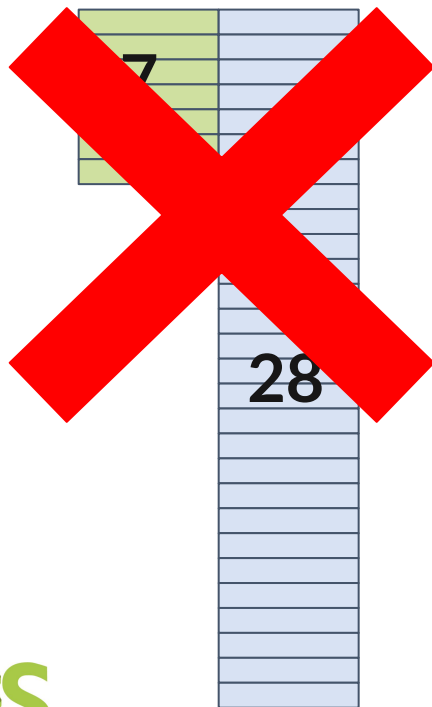
taxonomy identifier
scientific name
sample alias
sample title
sample description
collection date
geographic location (country and/or sea)






Tree of Life additional

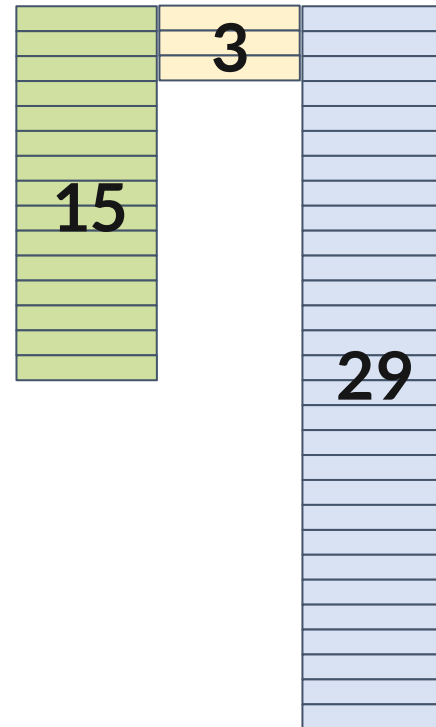
geographic location (region and locality)
organism part
lifestage
habitat
sex
collecting institution
collected by
project name
geographic location (latitude)
geographic location (longitude)
specimen_voucher

Default (ERC000011)



 Mandatory
 Recommended
 Optional

Tree of Life (ERC000053)



Capture the information as early as possible

- “It is the doom of men that they forget...”
- What metadata should be captured? - Plan ahead!



Unlucky example - How not to...

- ✗ Biodiversity study using a default checklist
- ✗ Unknown information on samples
- ✗ Low precision and accuracy metadata values

For example:

- ✗ Sample location only provided as “Strängnäs”
- ✗ 2 decimal digits for gps location (~3 km error margin)
- ✗ Bad enough for an aquatic species to end up on land

A Good Example

Biosample: SAMEA112878232

ce0be2db-efbc-4e1d-a5b6-c004dccc9e6d-ERGA-specimen

Organism:

Cladocora caespitosa

Scientific Name:

Cladocora caespitosa

Sample Accession:

SAMEA112878232

Location:

45.51562 N 13.57029 E

Center Name:

Earlham Institute

Sample Alias:

642d8ed0fe6059b46659b673

Checklist:

ERC000053

Broker Name:

COPO

Sample Title:

ce0be2db-efbc-4e1d-a5b6-c004dccc9e6d-ERGA-specimen

Original Collection Date:

2022-12-14

Habitat:

sea

Collector ORCID ID:

0000-0002-3312-382X|0000-0001-5488-0793

Collection Date:

2022-12-14

Geographic Location (Longitude):

13.57029

Collected By:

DAVID STANKOVIC|BORUT MAVRIC

Original Geographic Location (Longitude):

13.57029

Proxy Biomaterial:

PMS TIS 31

Sample Coordinator ORCID ID:

0000-0003-0714-5301

Specimen Id:

ERGA DS 382X 06 01

Identified By:

DAVID STANKOVIC|BORUT MAVRIC

Proxy Voucher:

PMSL-Invertebrata-24

Lifestage:

adult

Geographic Location (Country And/or Sea):

Slovenia

Original Geographic Location:

Slovenia|North Adriatic|Bernardin

ENA-CHECKLIST:

ERC000053

Sex:

HERMAPHRODITE MONOEICIOUS

Voucher Institution Url:

<https://www.pms-lj.si/en/>

Geographic Location (Latitude):

45.51562

Organism Part:

WHOLE ORGANISM

Sample Collection Method:

hand collected during scuba diving

Geographic Location (Region And Locality):

North Adriatic|Bernardin

GAL Sample Id:

NOT PROVIDED

Identifier Affiliation:

National Institute of Biology|Department of Organisms and Ecosystems Research|Marine Biology Station Piran

Original Geographic Location (Latitude):

45.51562

Sample Coordinator:

ELENA BUZAN

Specimen Voucher:

NOT_APPLICABLE

Sample Coordinator Affiliation:

University of Primorska

G A L:

SciLifeLab

Project Name:

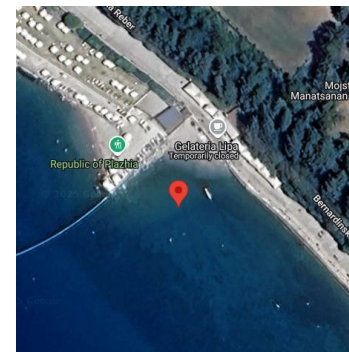
ERGA

Collecting Institution:

National Institute of Biology

Tollid:

jaClaCaes3



Takeaway Message

- A chain is only as strong as its weakest link
- Science is a game of trust
- For reference data, always aim as high as possible with metadata, not just for the bare necessities (or even less)
- Provide all information that you have, even if it isn't mandatory
- Record rich and precise metadata for each step of the process
- Define responsibilities and maintain contact throughout project