Uploading data to repositories

NCCR Network Ventures
Open Research Data Meetings
Lausanne, 8 May 2023





What is a data repository?



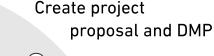
Data Repositories

Define the research question Create project

Verify and re-use

Preserve







Evaluate and select

Publish and manage access

> Process, analyse and interpret

A centralized place to

- Hold data over the long term
- Make data available for others to use

Requirements for the results of SNSF-funded projects

Researchers must upload the data underlying a publication to a repository that conforms with the FAIR Data Principles.

- Findable
- Accessible
- Interoperable
- Re-useable

There must be a long-term preservation plan for the archived data.



FAIR Repositories

- Findable: dataset is given a persistent identifier
- Accessible: researcher can choose a license
- Interoperable: repository supports metadata
- Re-useable: citation information and metadata are publicly accessible

What is NOT a repository?

Supplementary Materials: don't meet SNSF requirements

nature microbiology

Explore content > About the journal > Publish with us >

nature > nature microbiology > articles > article

Article Published: 30 May 2022

Mapping phyllosphere microbiota interactions in planta to establish genotype-phenotype relationships

Martin Schäfer, Christine M. Vogel, Miriam Bortfeld-Miller, Maximilian Mittelviefhaus & Julia A. Vorholt ⊡

Nature Microbiology 7, 856-867 (2022) | Cite this article

4754 Accesses | 5 Citations | 52 Altmetric | Metrics

Abstract

Host-associated microbiomes harbour hundreds of bacterial species that co-occur, creating the opportunity for manifold bacteria–bacteria interactions, which in turn contribute to the overall community structure. The mechanisms that underlie this self-organization among bacteria remain largely elusive. Here, we studied bacterial interactions in the phyllosphere microbiota. We screened for microbial interactions in planta by adding 200 endogenous

Mapping phyllosphere microbiota interactions in planta to establish genotype-phenotype relationships

Supplementary information

Supplementary Information

Supplementary Note and Figs. 1-13.

Reporting Summary

Peer Review File

Supplementary Video

Time-lapse video of Leaf374 cell lysis upon addition of Leaf245 supernatant. Imaging was started 1 min after supernatant addition and a picture was recorded every 20 s over the course of 20 min.

Supplementary Table 1

Supplementary Tables 1-7.

Supplementary Data 1

Demultiplexed 16S RNA gene amplicon sequencing read counts and corresponding metadata.

Supplementary Data 2

Source data for Supplementary Figs. 3, 4, 6 and 9-13.

Supplementary Materials: don't meet SNSF requirements

May not be Findable: I can only find the data if I find the publication first

May not be Accessible: Journals do not prompt authors to assign a license

No long-term preservation plan

Github: doesn't meet SNSF requirements

From SNSF:

Github is a commercial platform that is well-suited to code development, but it is not a data archiving tool.

Therefore, a copy of the code has to be archived in a data repository.

For instance, you can set up a connection between Zenodo and Github to ensure a permanent record of your code and make it citable.



What we'll cover

Finding the right repository for your data set.

Annotating data in a standard way, so that other researchers can understand them.

Streamlining the process of uploading data.

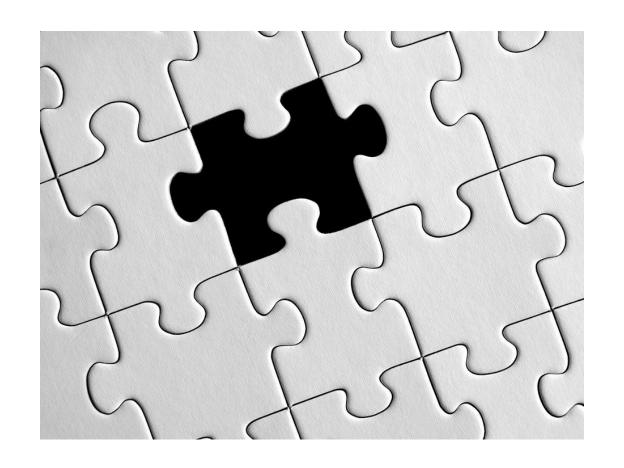
Publishing protocols to ensure reproducibility and reuseability.

What we'll cover

Concepts, not specifics.

I'm going to point out gaps to you.

You'll have to fill them in yourself.



Finding a Repository



General	Discipline-specific
Subject independentData from many fields	 Specific to a subject or data type Well-known in a particular field



Types of Repositories

General	Discipline-specific
Upside	
 Indexed by major search engines Upload is usually simple Provide a digital object identifier (DOI) Integrate with Github 	 Your data will be seen by the people who are most likely to appreciate and re-use it Data are usually curated or verified Provide a persistent identifier
Downside	
May be treated as a "data dump"	 Upload may be tricky: "What does this repository want from me?!?"



Exercise: Choose a repository

- Search for: Nature Data Repository Guidance.
 - If a subject-specific data repository exists for your research, it is a good idea to use it.

nature > scientific data > policies > data repository guidance

Policies

Editorial & Publishing Policies

For Referees

Data Policies

Data Repository Guidance

Data Repository Guidance

Scientific Data mandates the release of datasets accompanying our Data Descriptors, but we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible. Where a suitable discipline-specific resource does not exist, data should be submitted to a generalist repository.

Authors must deposit their data to a data repository as part of the manuscript submission process; manuscripts will not otherwise be sent for review. If data have not been deposited to a repository prior to manuscript submission, authors can upload their data to figshare or the Dryad Digital Repository during the submission process. Data may also be deposited to these resources temporarily, if the main host repository does not support confidential peer review.

Repositories need to meet our requirements for anonymous peer-review, data access, preservation, resource stability, and suitability for use by all researchers with the



Exercise: Choose a repository

- Search for: Nature Data Repository Guidance.
 - If a subject-specific data repository exists for your research, it is a good idea to use it.
- Answer the questions on your worksheet.
- Create an account.
 - If you're not sure about which repository to choose, you can create a test account at sandbox.zenodo.org.

Annotating Data

An example from my personal life
Transferring dental records from the USA to Switzerland



Data annotation in my personal life

Hey Kendra,

Here are your X-rays you requested. The FMX was taken in 2019 and the BWX Was taken in may of 2021. If you have any further questions feel free to call!

	BROWN^KENDRA_MN01_MP01_006.JPG		BROWN^KENDRA_MN01_MP08_002.JPG	Ser.	BROWN^KENDRA_MN01_MP15_001.JPG
1730	BROWN^KENDRA_MN01_MP02_013.JPG	The	BROWN^KENDRA_MN01_MP09_000.JPG		BROWN^KENDRA_MN01_MP16_014.JPG
	BROWN^KENDRA_MN01_MP03_019.JPG	118	BROWN^KENDRA_MN01_MP10_007.JPG	122	BROWN^KENDRA_MN01_MP17_011.JPG
	BROWN^KENDRA_MN01_MP04_012.JPG	(T)	BROWN^KENDRA_MN01_MP11_017.JPG	120	BROWN^KENDRA_MN01_MP18_004.JPG
000	BROWN^KENDRA_MN01_MP05_003.JPG		BROWN^KENDRA_MN01_MP12_015.JPG	200	BROWN^KENDRA_MN04_MP01_028.JPG
	BROWN^KENDRA_MN01_MP06_010.JPG	411/	BROWN^KENDRA_MN01_MP13_008.JPG	-	BROWN^KENDRA_MN04_MP02_029.JPG
	BROWN^KENDRA_MN01_MP07_009.JPG		BROWN^KENDRA_MN01_MP14_005.JPG		BROWN^KENDRA_MN04_MP03_030.JPG



My dental x-rays and FAIRness



Implementation Networks News Events Resources About GO FAIR Q

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

Home > FAIR Principles > F3: Metadata clearly and explicitly include the identifier of the data they describe

> FAIR Principles

- F1: (Meta) data are assigned globally unique and persistent identifiers
- F2: Data are described

What does this mean?

This is a simple and obvious principle, but of critical importance to FAIR. The metadata and the dataset they describe are usually separate files. The association between a metadata file and the dataset should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata. As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose.



FMX? BWX?

Full Mouth ("FMX"): an FMX captures detailed images of each tooth and the surrounding structures. Uses a combination of Bitewing and Periapical x-rays.

Bitewing ("BWX"): shows the upper and lower back teeth in a single view.



Radiographie complète de la bouche?

Les types de radiographie dentaire

Les médecins-dentistes et hygiénistes dentaires à Lausanne ou Fribourg peuvent réaliser plusieurs types de radiographies. Le choix dépend des besoins particuliers du patient, car chacune offre un point de vue différent de la bouche.

- Radiographie panoramique : pour une vue de l'ensemble des structures de l'anatomie bucco-dentaire.
- Radiographie interproximale : pour visualiser les surfaces interdentaires, non-visibles lors de l'examen clinique.
- Radiographie périapicale : pour évaluer l'état précis de deux dents entières de la racine jusqu'à la couronne.
- Radiographie céphalométrique : avant d'établir un traitement orthodontique.
- Radiographie 3D (scans) : dans certaines situations de traumatisme ou d'implantologie par exemple.



My dental x-rays and FAIRness



FAIR Principles Implementation Networks News Events Resources About GO FAIR Q

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

Home > FAIR Principles > 11: (Meta) data use a formal, accessible, shared, and broadly applicable language for knowledge representation

- > FAIR Principles
 - > F1: (Meta) data are assigned globally unique and persistent identifiers
 - F2: Data are described

What does this mean?

Humans should be able to exchange and interpret each other's data (so preferably do not use dead languages). But this also applies to computers, meaning that data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. The main goal of this principle is to provide a "common understanding" of digital objects by means of a language for knowledge representation to be used to represent these objects.



Parallels to research data

Good intentions

The curse of knowledge

When you know something, you tend to assume others know it too

Language barriers

Different vocabularies used in adjacent fields

Annotating Data

Providing metadata with the data you upload



Metadata and FAIRness



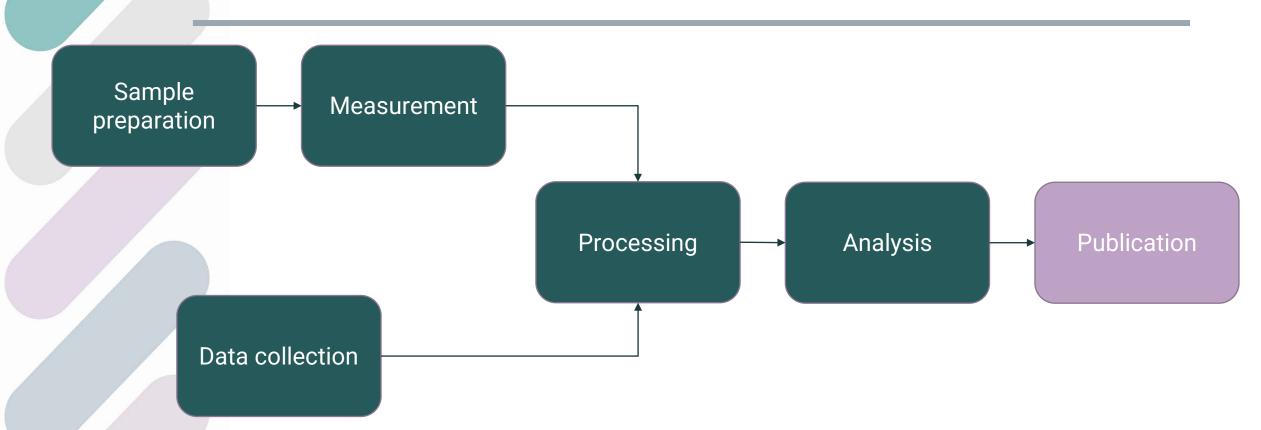
AIR Principles Implementation Networks News Events Resources About GO FAIR Q

F2: Data are described with rich metadata

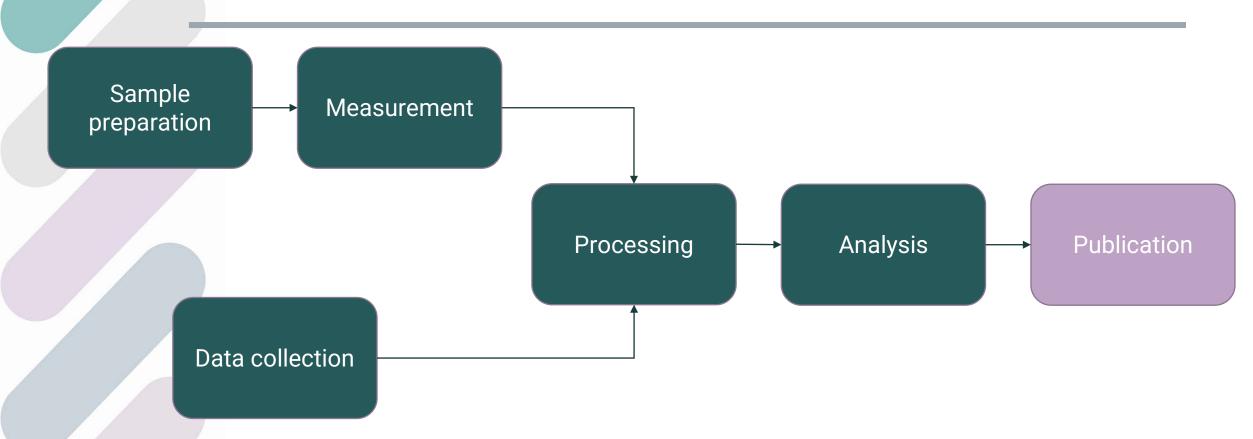
Researchers in your field (including your future self) need the metadata in order to understand, reproduce and re-use your work!

Rich metadata implies that you should not presume that you know who will want to use your data, or for what purpose.

Metadata associated with different research steps



Exercise: Identify important metadata for your research



In three minutes, write down as many metadata types as you can think of that are relevant to your research.



Metadata standards provide a "checklist" of fields that are relevant to the data set you're submitting.

Checklists often have some elements that are mandatory, some recommended, and some that are optional.

They should define what type of input values are expected for each element, e.g. free text, date in a specific format, geographical positions, numerical values, and controlled vocabularies or ontologies.



Finding metadata standards

If you are using a discipline-specific repository, it may impose metadata standards on you.

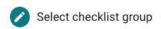
Examples:

- European Nucleotide Archive
 - Minimum Information about any (x) Sequence: MIxS
- Biolmage Archive
 - Recommended Metadata for Biological Images (REMBI)

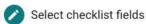
Register Samples using spreadsheet template

Download spreadsheet to register samples

Please select the most appropriate checklist group, checklist and checklist fields. Download an empty spreadsheet template, fill in the spreadsheet and submit the spreadsheet using Webin.







Download spreadsheet template

You have selected **Environmental Checklists**. Please select the most appropriate checklist from the list below.

GSC MIxS air

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

GSC MIxS host associated

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

GSC MIxS human associated

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

GSC MIxS human gut

Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

Selection	Field Name	Validation	Units	Description
~	tax_id	Text field	None	Taxonomy ID of the organism as in the NCBI Taxonomy database. Entries in the NCBI Taxonomy database have integer taxon IDs. See our tips for sample taxonomy here
~	scientific_name	Text field	None	Scientific name of the organism as in the NCBI Taxonomy database. Scientific names typically follow the binomial nomenclature. For example, the scientific name for humans is Homo sapiens.
~	sample_alias	Text field	None	Unique name of the sample. If not selected system will auto generate an unique alias
✓	sample_title	Text field	None	Title of the sample
~	sample_description	Text field	None	Description of the sample
~	project name	Text field	None	Name of the project within which the sequencing was organized
~	collection date	Regular expression ▼	None	The date the sample was collected with the intention of sequencing, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.
~	geographic location (country and/or sea)	Permitted values •	None	The location the sample was collected from with the intention of sequencing, as defined by the country or sea. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html).
~	geographic location (latitude)	Regular expression 🔻	DD	The geographical origin of the sample as defined by latitude. The values should be reported in decimal degrees and in WGS84 system
✓	geographic location (longitude)	Regular expression 🔻	DD	The geographical origin of the sample as defined by longitude. The values should be reported in decimal degrees and in WGS84 system
~	depth	Regular expression 🔻	m	The vertical distance below local surface, e.g. for sediment or soil samples depth is measured from sediment or soil surface, respectively. Depth can be reported as an interval for subsurface samples.
V	broad-scale environmental conte	_{Xt} Text field	None	Report the major environmental system the sample or specimen came from. The system(s) identified should have a coarse spatial grain, to provide the general environmental context of where the sampling was done (e.g. in the desert or a rainforest). We recommend using subclasses of EnvO's biome class: http://purl.obolibrary.org/obo/ENVO_0000428. EnvO documentation about how to use the field: https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.
>	local environmental context	Text field	None	Report the entity or entities which are in the sample or specimen's local vicinity and which you believe have significant causal influences on your sample or specimen. We recommend using EnvO terms which are of smaller spatial grain than your entry for "broad-scale environmental context". Terms, such as anatomical sites, from other OBO Library ontologies which interoperate with EnvO (e.g. UBERON) are accepted in this field. EnvO documentation about how to use the field: https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.



If you are using a general repository, you can search for the metadata standards that are widely used in your field.

Metadata Standards Catalog:

https://rdamsc.bath.ac.uk/

Fairsharing.org:

https://fairsharing.org/search?fairsharingRegistry=Standard



Finding metadata standards

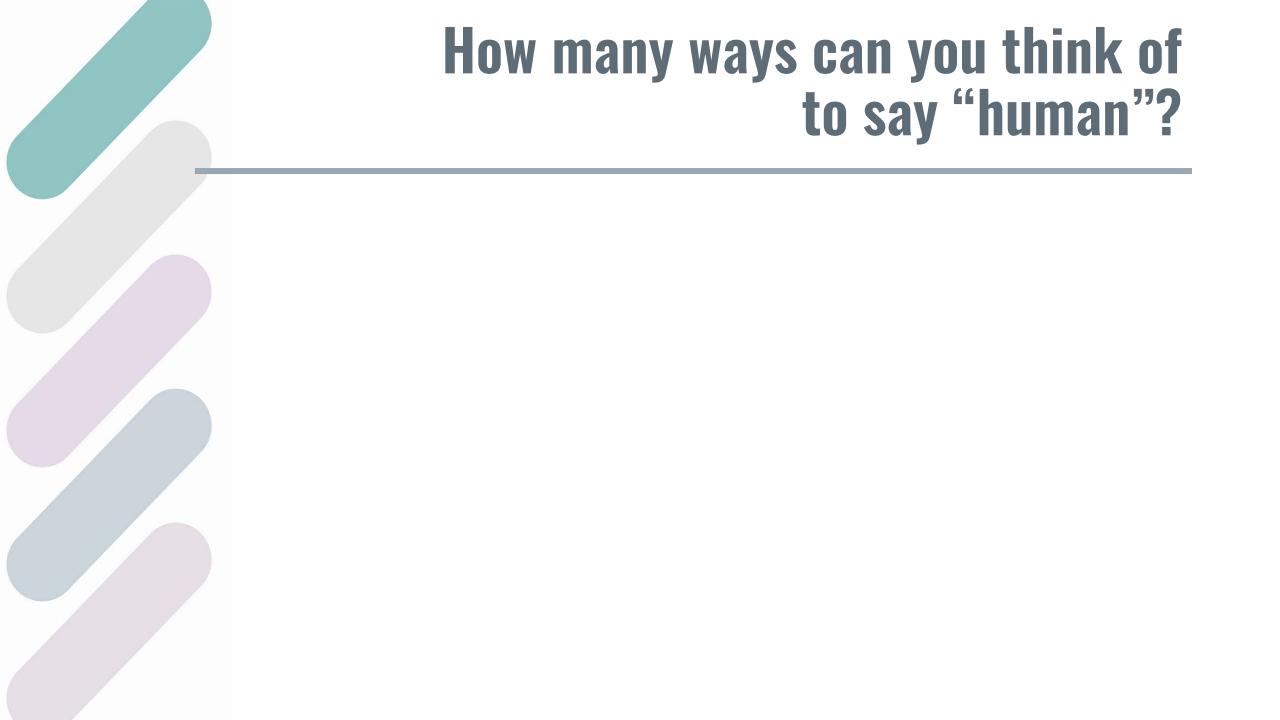
Warning: This is messy!

- Metadata standards are not well-defined in every field.
- Metadata standards can become outdated as research evolves over time
- Consider becoming part of the conversation

Exercise: Find metadata standards for your data

General Repository	Discipline-specific repository
 Search the Metadata Standards Catalog 	 Search the repository website for metadata standards
 Search Fairsharing.org Search Google Scholar for "metadata standards" + keywords from your field 	 Try a test submission to find the metadata standards

Controlled vocabularies



How many ways can you think of to say "human"?

person

human being

individual

man/woman/child

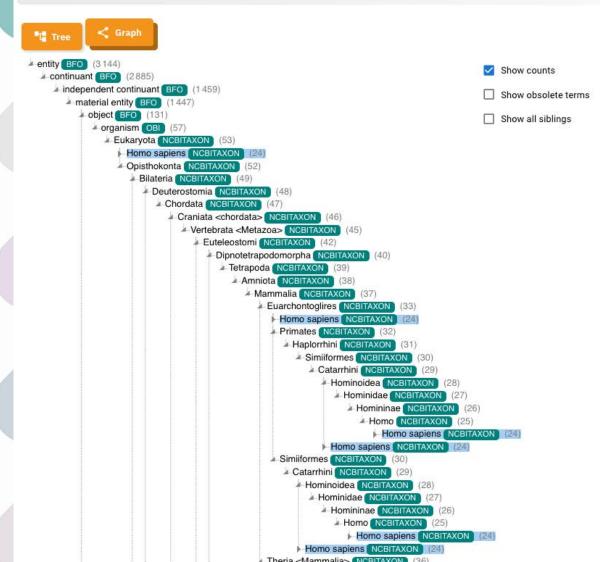
Homo sapiens

hominin

NCBITaxon:9606

homonid





▼ Class Information

alternative label

- · human being
- man
- · person
- human

editor preferred term

Homo sapiens

has_dbxref

GC_ID:1

has_obo_namespace

ncbi_taxonomy

has rank

species NCBITAXON

has_related_synonym

humans

imported from

- · http://purl.obolibrary.org/obo/omrse.owl
- · http://purl.obolibrary.org/obo/envo.owl
- · Ontology for Biomedical Investigations
- · http://purl.obolibrary.org/obo/ncbitaxon.owl

▼ Class Relations

Subclass of

- Eukaryota NCBITAXON
- Euarchontoglires NCBITAXON
- Catarrhini NCBITAXON
- Homo NCBITAXON
- has quality RO some biological sex PATO

Ontology Lookup Service



What is an ontology?

A set of concepts and categories in a subject area that shows:

- their definitions
- the relations between them



Ontologies and FAIRness



Implementation Networks News Events Resources About GO FAIR Q

12: (Meta)data use vocabularies that follow the FAIR principles

Use controlled vocabularies (from ontologies) so that others (humans or machines) can find, access, interoperate and reuse the data.



Why use an ontology?

Knowledge: organized information, meaning

Information: linked elements, context

Data: raw elements

Limit complexity
Organize data into information and knowledge

Selection	Field Name	Validation	Units	Description
~	tax_id	Text field	None	Taxonomy ID of the organism as in the NCBI Taxonomy database. Entries in the NCBI Taxonomy database have integer taxon IDs. See our tips for sample taxonomy here
~	scientific_name	Text field	None	Scientific name of the organism as in the NCBI Taxonomy database. Scientific names typically follow the binomial nomenclature. For example, the scientific name for humans is Homo sapiens.
~	sample_alias	Text field	None	Unique name of the sample. If not selected system will auto generate an unique alias
✓	sample_title	Text field	None	Title of the sample
~	sample_description	Text field	None	Description of the sample
~	project name	Text field	None	Name of the project within which the sequencing was organized
~	collection date	Regular expression ▼	None	The date the sample was collected with the intention of sequencing, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.
~	geographic location (country and/or sea)	Permitted values •	None	The location the sample was collected from with the intention of sequencing, as defined by the country or sea. Country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html).
~	geographic location (latitude)	Regular expression 🔻	DD	The geographical origin of the sample as defined by latitude. The values should be reported in decimal degrees and in WGS84 system
✓	geographic location (longitude)	Regular expression 🔻	DD	The geographical origin of the sample as defined by longitude. The values should be reported in decimal degrees and in WGS84 system
~	depth	Regular expression 🔻	m	The vertical distance below local surface, e.g. for sediment or soil samples depth is measured from sediment or soil surface, respectively. Depth can be reported as an interval for subsurface samples.
V	broad-scale environmental conte	_{Xt} Text field	None	Report the major environmental system the sample or specimen came from. The system(s) identified should have a coarse spatial grain, to provide the general environmental context of where the sampling was done (e.g. in the desert or a rainforest). We recommend using subclasses of EnvO's biome class: http://purl.obolibrary.org/obo/ENVO_0000428. EnvO documentation about how to use the field: https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.
>	local environmental context	Text field	None	Report the entity or entities which are in the sample or specimen's local vicinity and which you believe have significant causal influences on your sample or specimen. We recommend using EnvO terms which are of smaller spatial grain than your entry for "broad-scale environmental context". Terms, such as anatomical sites, from other OBO Library ontologies which interoperate with EnvO (e.g. UBERON) are accepted in this field. EnvO documentation about how to use the field: https://github.com/EnvironmentOntology/envo/wiki/Using-ENVO-with-MIxS.



Exercise: Explore some ontologies

Use the Ontology Lookup Service to search for the metadata items you identified in the last session.

In which ontologies do they appear?

Does a certain ontology seem to work well for your research?

Streamlining



Streamlining the upload

"Begin with the end in mind"

Which metadata do you need?

Treat the upload as an important step in your research work

- Develop a protocol to share with your community
- Write a program if possible



Look at the metadata checklist and ontology you identified for your research.

How can you incorporate these into your own data workflow?

Work individually for 5 minutes.



Look at the metadata checklist and ontology you identified for your research.

How can you incorporate these into your own data workflow?

Work individually for 5 minutes.

Work in groups of 2 or 3 and discuss your ideas.



Look at the features of your repository.

Could you program

- The upload?
- The preparation of your files?

Discuss in groups of 2-3.

Protocols



You're doing great!

You can upload your data to a repository.

This brings you into compliance with SNSF requirements.

You know how to annotate your dataset with structured and rich metadata.

Your data are aligned with the FAIR principles.



But what if...?

But it still might be difficult for someone (including your future self) to start with your raw data and reproduce the results in your paper.

Publishing protocols can provide the missing link here.



Publishing protocols

Where can I publish a protocol?

- Protocols.io
- WorkflowHub
- Protocol journals



Take a few minutes to write an outline of one of your research protocols.

 Depending on your stage, this may be rough or detailed.

How could you share your protocol with others who could help you refine it?

Discussion: Have I missed anything?

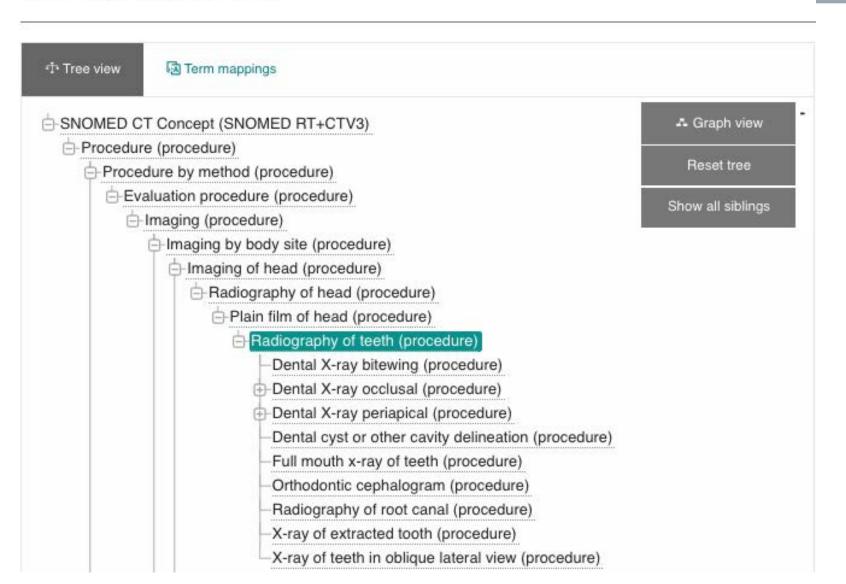
Thank you!







Radiography of teeth (procedure)









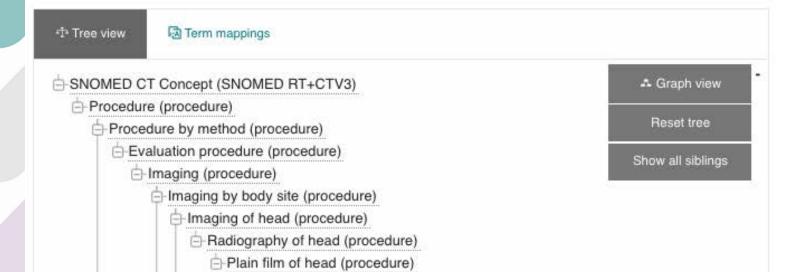


Dental X-ray bitewing (procedure)

If http://snomed.info/id/241046008 Copy

Search SNOMED

Search



Radiography of teeth (procedure)

Radiography of teeth (procedure)

Plain film of body region (procedure)

Plain film of head (procedure)

Radiography of head (procedure)

Radiographic imaging procedure (procedure)

Plain film of head (procedure)

Radiographic imaging procedure by site (procedure)

Dental X-ray bitewing (procedure)

Dental X-ray bitewing (procedure) Radiography of face, head AND/OR neck (procedure)

-Radiography of teeth (procedure)

Dental X-ray bitewing (procedure)

Term information

database cross reference

CTV3:X70di

altLabel

. BW - Dental bitewing X-ray

hasDbXref

CTV3:X70di

prefLabel

. Dental X-ray bitewing

Term relations

Subclass of:

- Radiography of teeth (procedure)
- · Radiography of teeth (procedure) and

Role group (attribute) some (

Method (attribute) some

Radiographic imaging - action (qualifier value)

and

Procedure site - Direct (attribute) some

Tooth structure (body structure))