

# Metadata

*Introduction to Data Management Practices course*

NBIS DM Team

data@nbis.se

<https://nbisweden.github.io/module-metadata-dm-practices/index.html>



---

*“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.”*

*“Your primary collaborator is yourself six months from now, and your past self don’t answer e-mails.”*

The data about the data (or anything really)

*“One persons metadata, is another persons data”*

- Describe data at different levels
  - e.g. a whole study vs the samples

## *Examples*

- Creators
- File types and formats of the data
- Licence for re-use of the data
- Methodology for data collection
- Analytical and procedural information
- Sources of samples
- Sample treatment
- Geolocation(s) of samples

# What problems do you see with the descriptions of these samples?

	animal ID	date	mouse line	strain	age
1	800793	2018-01-06	Alk3	BALB/cJ	10
2	804396	2019-01-07	Vegfr	C57BL/6	6
3	805431	2018-01-12	Vegfr	C57BL/6	P9
4	805992	2019-01-13	Vegfr	C57BL/6	P10
5	808935	2020-01-14	Alk3	BALB/cJ	12
6	810875	2019-01-16	alk6	C57BL/6	E16
7	812308	2018-01-19	Alk3	BALB/cJ	12
8	814334	2019-01-19	Vegfr	C57BL/6	P9
9	816649	2018-01-20	Vegfr	C57BL/6	P10

[samples\\_metadata\\_lesson.csv](#)

- 
- Date formats
  - Different terms for the same information
  - Misspelled terms
  - Not clear what a data point means
  - Not clear what unit

- 
- Descriptions must be understandable over time - *not only for you*
  - FAIR principles → also for computers
  - Consistency
    - Date formats
    - Units
    - Terms

- 
- What is necessary for you to do your particular analysis
  - What is necessary for someone to understand the data
  - All the metadata you have
  - *“How can I make this dataset as useful as possible for others?”*



---

*“A biologist would rather share a toothbrush with another biologist than share a gene name”*

- Consistency and stringency
- **Controlled vocabularies**
- **Ontologies**
- Thesauruses (Thesauri)
- Taxonomies

---

**How many terms for *Breast Cancer* can you think of?**

Breast Neoplasm  
Neoplasm, Breast  
Breast Tumors  
Breast Tumor  
Tumor, Breast  
Tumors, Breast  
Neoplasms, Breast  
Breast Cancer  
Cancer, Breast  
Mammary Cancer  
Cancer, Mammary  
Cancers, Mammary  
Mammary Cancers  
Malignant Neoplasm of Breast  
Breast Malignant Neoplasm  
Breast Malignant Neoplasms  
Malignant Tumor of Breast  
Breast Malignant Tumor  
Breast Malignant Tumors

Cancer of Breast  
Cancer of the Breast  
Mammary Carcinoma, Human  
Carcinoma, Human Mammary  
Carcinomas, Human Mammary  
Human Mammary Carcinomas  
Mammary Carcinomas, Human  
Human Mammary Carcinoma  
Mammary Neoplasms, Human  
Human Mammary Neoplasm  
Human Mammary Neoplasms  
Neoplasm, Human Mammary  
Neoplasms, Human Mammary  
Mammary Neoplasm, Human  
Breast Carcinoma  
Breast Carcinomas  
Carcinoma, Breast  
Carcinomas, Breast

- List of terms to describe some domain of knowledge
- Only one term per phenomenon
- Term definition
- List synonyms
- Each term has a unique identifier

## **Medical Subject Headings - MeSH**

### **Breast Neoplasms**

*Definition:* Tumors or cancer of the human BREAST

*Synonyms:* Breast Tumors, Breast Tumor, Breast Cancer, ...

*MeSH Unique ID:* D001943

- A controlled vocabulary
- Captures term relationships, e.g.
  - *is a*
  - *part of*
  - *contained in*
  - *produced by*
- Hierarchy / Tree
  - A term can be present at several places in the hierarchy

OLS / Human Phenotype Ontology **HP** / **HP:0001658**  Copy



## Myocardial infarction

 [http://purl.obolibrary.org/obo/HP\\_0001658](http://purl.obolibrary.org/obo/HP_0001658)  Copy








Necrosis of the myocardium caused by an obstruction of the blood supply to the heart and often associated with chest pain, shortness of breath, palpitations, and anxiety as well as characteristic EKG findings and elevation of serum markers including creatine kinase-MB fraction and troponin. [ HPO:probinson ]

Synonyms: **MI** **Heart attack**

 Tree view

 Term mappings

 Term history

 All  
 Phenotypic abnormality  
 Abnormality of the cardiovascular system  
 Abnormal cardiovascular system physiology  
 **Myocardial infarction**

 Graph view

Reset tree

Show all siblings

☒ Preferred root terms

☐ All terms

### Term information

#### database cross reference

- MSH:D009203
- UMLS:C0027051
- SNOMEDCT\_US:22298006

#### layperson term

Heart attack [ ORCID:0000-0001-5208-3432 ]

#### abbreviation

MI

#### has obo namespace

human\_phenotype

#### id

HP:0001658

### Term relations

#### Subclass of:

- Abnormal cardiovascular system physiology

OLS / The BRENDA Tissue Ontology (BTO) **BTO** / **BTO:0000564**  Copy




## heart valve

 [http://purl.obolibrary.org/obo/BTO\\_0000564](http://purl.obolibrary.org/obo/BTO_0000564)  Copy

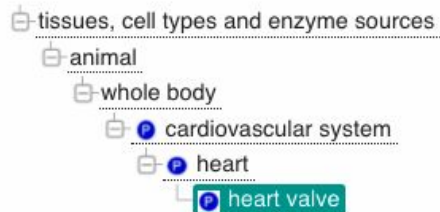
Search


A structure especially in a vein or lymphatic that closes temporarily a passage or orifice or permits movement of fluid in one direction only. [ From\_Merriam-Webster's\_Online\_Dictionary\_at\_www.Merriam-Webster.com:http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=valve ]

 Tree view

 Term mappings

 Term history



 Graph view

Reset tree

Show all siblings

### Term information

**has obo namespace**

BrendaTissueOBO

**id**

BTO:0000564

### Term relations

**Subclass of:**

- *part of* some heart

HOW STANDARDS PROLIFERATE:  
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION:  
THERE ARE  
14 COMPETING  
STANDARDS.

14?! RIDICULOUS!  
WE NEED TO DEVELOP  
ONE UNIVERSAL STANDARD  
THAT COVERS EVERYONE'S  
USE CASES.



**SOON:**

SITUATION:  
THERE ARE  
15 COMPETING  
STANDARDS.



- At what point does it make sense to use something that exists?
  - Number of terms
  - Nature of terms
  - Relationships of terms
  - Terms management
    - Definitions
- FAIRness
  - Unique identifiers
  - Home brew vocabularies makes it harder to achieve machine readability

- Collections of metadata **elements** of relevance for a particular purpose
- Elements
  - Mandatory, Recommended, or Optional
  - Defined input value type
    - Free text, data, geographical position, numerical values, ontology terms
  - Can itself be an ontology term
- Stricter → potentially increased FAIRness
- Generic to Specific

- Describing digital and physical resources
- 15 elements

<b>Term Name: creator</b>	
URI:	<a href="http://purl.org/dc/elements/1.1/creator">http://purl.org/dc/elements/1.1/creator</a>
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
<b>Term Name: date</b>	
URI:	<a href="http://purl.org/dc/elements/1.1/date">http://purl.org/dc/elements/1.1/date</a>
Label:	Date
Definition:	A point or period of time associated with an event in the lifecycle of the resource.
Comment:	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
References:	[W3CDTF] <a href="http://www.w3.org/TR/NOTE-datetime">http://www.w3.org/TR/NOTE-datetime</a>
<b>Term Name: description</b>	
URI:	<a href="http://purl.org/dc/elements/1.1/description">http://purl.org/dc/elements/1.1/description</a>
Label:	Description
Definition:	An account of the resource.
Comment:	Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
<b>Term Name: format</b>	
URI:	<a href="http://purl.org/dc/elements/1.1/format">http://purl.org/dc/elements/1.1/format</a>
Label:	Format
Definition:	The file format, physical medium, or dimensions of the resource.
Comment:	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
References:	[MIME] <a href="http://www.iana.org/assignments/media-types/">http://www.iana.org/assignments/media-types/</a>

- *ENA virus pathogen reporting standard checklist*
- Reporting metadata of virus pathogen samples associated with genomic data
- 35 elements - 9 mandatory and 15 recommended

## Checklist: ERC000033

### ENA virus pathogen reporting standard checklist













Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

#### Checklist Fields

Filter fields... 

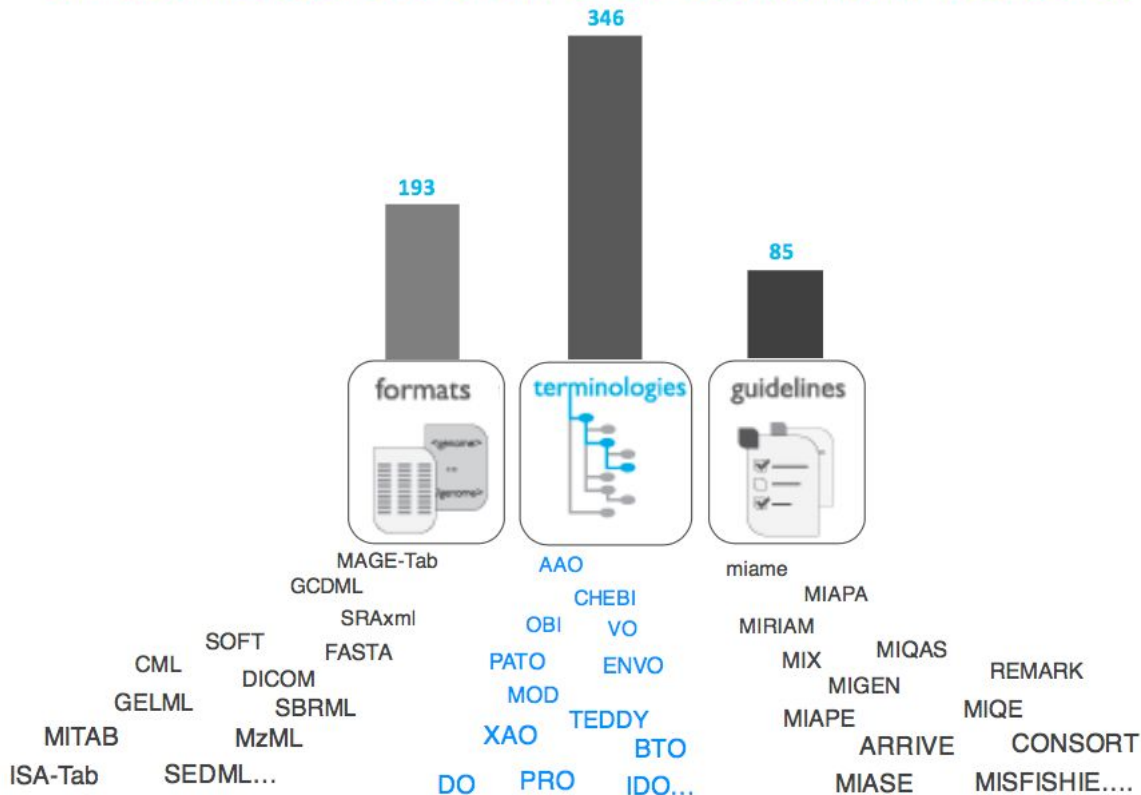
Filter by type:

Human surveillance data
Collection event information
sample collection
host disorder
host description
Virus isolate information
General collection event information
Serology detection
Infraspecies

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
subject exposure	 free text		optional	
subject exposure duration	 free text		optional	
type exposure	 free text		optional	
personal protective equipment	 free text		optional	
hospitalisation	 text choice	options 	optional	
illness duration	 free text		optional	
illness symptoms	 free text		optional	
collection date	 restricted text	regular expression 	recommended	
geographic location (country and/or sea)	 text choice	options 	mandatory	

# How do I know what to use?

In the life sciences there are **>600 content standards**



- 
- Your own metadata standard
  - Document what type of information is supposed to be entered for the metadata fields
  - Name, units, allowed values, definitions, ...

---

# **Exercise: Start a data dictionary**

# Start a Data dictionary

1. Open [samples\\_metadata\\_lesson.csv](#)
2. Create a new [data\\_dictionary](#) file
3. Add headings to [data\\_dictionary](#)

- Current variable name
- ENA variable name
- Measurement unit
- Allowed values
- Definition
- Description

4.

3.	A	B	C	D	E	F
1	Current variable name	ENA Variable name	Measurement unit	Allowed values	Definition	Description
2	animal ID					
3	date					
4	mouse line					
5	strain					
6	age					
7	developmental stage					
8	sex					
9	organism part					
10	genotype					
11	experiment type					
12	experiment reference					
13	researcher					

2. [data\\_dictionary](#)

	A	B	C	D	E	F	G	H	I	J	K	L
1	animal ID	date	mouse line	strain	age	developmental stage	sex	organism part	genotype	experiment type	experiment reference	researcher
2	800793	2018-01-06	Alk3	BALB/cJ	10	adult	F		wt	behaviour	Jan0618_B	Kim
3	804396	2019-01-07	Vegfr	C57BL/6	6	adult	mael	lung	wild type genotype	sequencing assay	up_201_2	Sam
4	805431	2018-01-12	Vegfr	C57BL/6	P9	pup	female	lung	wild type genotype	IHC	Jan1218_IHC	Sam
5	805992	2019-01-13	Vegfr	C57BL/6	P10	pup	male	lung	wild type genotype	IHC	Jan1319_IHC	Sam
6	808935	2020-01-14	Alk3	BALB/cJ	12	adult	M		wt	behaviour	Jan1420_B	Kim
7	810875	2019-01-16	alk6	C57BL/6	E16	embryo	N/A	brain	KO	culture	Jan1619_C	Jo
8	812308	2018-01-19	Alk3	BALB/cJ	12	adult	F		wt	behaviour	Jan1918_B	Kim
9	814334	2019-01-19	Vegfr	C57BL/6	P9	pup	mael	lung	wild type genotype	IHC	Jan1919_IHC	Sam
10	816649	2018-01-20	Vegfr	C57BL/6	P10	pup	female	lung	wild type genotype	IHC	Jan2018_IHC	Sam
11	819947	2019-01-24	vegfr	C57BL/6						sequencing assay	up_432_1	Alex
12	820421	2019-01-31	Vegfr	C57BL/6	P9	pup	female	lung	wild type genotype	IHC	Jan3119_IHC	Sam
13	821756	2018-02-04	vegfr	C57BL/6	P9	pup	male	lung	wild type genotype	IHC	Feb0418_IHC	Alex
14	877817	2019-04-13	vegfr	C57BL/6	6	adult	male	lung	wild type genotype	sequencing assay	up_432_2	Alex
15	821844	2002-07-18	Vegfr	C57BL/6	P10	pup	female	lung	wild type genotype	IHC	Feb0718_IHC	Sam
16	826176	2019-02-14	Vegfr	C57BL/6	P10	pup	male	lung	wild type genotype	IHC	Feb1419_IHC	Sam
17	832626	2020-02-16	Vegfr	C57BL/6	P9	pup	Male	lung	wild type genotype	IHC	Feb1620_IHC	Sam
18	834217	2020-02-18	Alk3	BALB/cJ	4	adult	male	lung	Vegfr2 Y949F/Y949F	sequencing assay	up_235_1	Kim

1. [samples\\_metadata\\_lesson.csv](#)

4. Copy headings from [samples\\_metadata\\_lesson.csv](#) to rows in [data\\_dictionary](#)

- Add some definitions
- Add some units



# Data dictionary - start

	A	B	C	D	E	F
1	Current variable name	ENA Variable n	Measurement	Allowed values	Definition	Description
2	animal ID					
3	date			format: YYYY-MM-DD, >=proj_start_date & <=today	Date of experiment ???	
4	mouse line					
5	strain				The mouse strain of the animal	
6	age		days (?)		Age of animal	
7	developmental stage					
8	sex			male, female, unknown	Sex of the animal	
9	organism part					
10	genotype					
11	experiment type					
12	experiment reference					
13	researcher					
14						

- 
- Use standards of deposition databases were you plan to publish your data
  - Helps with selecting elements
  - Makes data submission much easier

## **Exercise:**

**Look up an ENA checklist to improve the data dictionary**

1. Go to <https://www.ebi.ac.uk/ena/browser/checklists> to see the available checklists
2. Scroll down the listing until you find the **ERC000011 ENA default sample checklist**
3. Go through the data dictionary and find suitable field names in the ENA default sample checklist for those fields. Add them to the ENA Variable name column of your data dictionary file.

## Checklist: ERC000011

### ENA default sample checklist

Minimum information required for the sample

### Checklist Fields

Filter fields... 

Filter by type:

- Part and developmental stage of organism
- Collection event information
- sample collection
- Organism characteristics
- host description
- Pointer to physical material
- Intraspecies information

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
bio_material	① free text		optional	
culture_collection	① free text		optional	
specimen_voucher	① free text		optional	
cultivar	① free text		optional	
ecotype	① free text		optional	
isolate	① free text		optional	
sub_species	① free text		optional	
variety	① name or identifier of a genetically or otherwise modified strain from which sample was obtained, derived from a parental strain (which should be annotated in the strain field; sub_strain from which sample was obtained)		optional	
sub_strain	①		optional	
cell_line	①		optional	
serotype	① free text		optional	

	A	B	C	D	E	F
1	Current variable name	ENA Variable n	Measurement	Allowed values	Definition	Description
2	animal ID					
3	date			format: YYYY-MM-DD, >=proj_start_date & <=today	Date of experiment ???	
4	mouse line	sub_strain				
5	strain	strain			The mouse strain of the animal	
6	age		days (?)		Age of animal	
7	developmental stage	dev_stage				
8	sex	sex		male, female, unknown	Sex of the animal	
9	organism part	tissue_type				
10	genotype					
11	experiment type					
12	experiment reference					
13	researcher					
14						

## Checklist: ERC00011

### ENA default sample checklist

Minimum information required for the sample

Checklist Fields			
Filter fields...			
Filter by type:			
Part and developmental stage of organism			
Collection event information			
sample collection			
Organism characteristics			
host description			
Pointer to physical material			
Intraspecies information			
Field Name	Field Format (Field Restriction)	Requirement	(Units)
cell_type	free text	optional	
dev_stage	free text	optional	
germline	free text	optional	
tissue_lib	free text	optional	
tissue_type	free text	optional	
collection_date	restricted text <a href="#">regular expression</a>	optional	
isolation_source	free text	optional	
lat_lon	free text	optional	
collected_by	free text	optional	
geographic location (country and/or sea)	text choice <a href="#">options</a>	optional	
geographic location (region and locality)	free text	optional	
identified by	free text	optional	

- This standard is very liberal when it comes the allowed values for the different fields
- *We can do better!*
- Use ontology terms
  - Improves FAIRness
  - But which ontologies...?

- Tools
  - [FAIRsharing.org](https://fairsharing.org)
  - [EBI Ontology Tooling page](https://eutils.ebi.ac.uk/ontology/tooling/)
    - [Zooma](https://zooma.ebi.ac.uk/) - map free text to ontology terms
    - [Ontology Lookup Service - OLS](https://eutils.ebi.ac.uk/ontology/lookup/)
- Not an exact science... There is no perfect way...
- Sometimes hard
- Trial and error



**A curated, informative and educational resource on data and metadata  
*standards*, inter-related to *databases* and data *policies*.**

## HOW CAN WE HELP?

We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.



### Societies, unions and community alliances

Raise awareness around standards, databases, repositories and data policies, as well as mobilise your community to take action to promote the registration, use and citation of key resources...

[\[read more\]](#)

Researchers

Developers & Curators

Journal Publishers

Librarians & Trainers

**Societies & Alliances**

Funders

Find

Discover

Learn



[EMBL-EBI](#)
[Services](#)
[Research](#)
[Training](#)
[About us](#)
[Search](#)

EMBL-EBI 



ONTOLOGY ANNOTATION

[Home](#)
[Help](#)
[About](#)

## Query

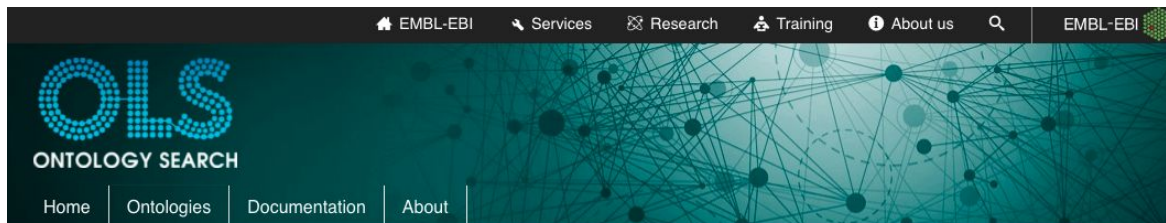
Use the text box to find possible ontology mappings for free text terms in the ZOOMA repository of curated annotation knowledge. You can add one term (e.g. 'Homo sapiens') per line. If you also have a type for your term (e.g. 'organism'), put this after the term, separated by a tab. If you are new to ZOOMA, take a look at our getting started guide.

[Show me some examples...](#)

Bright nuclei  
 Agammaglobulinemia 2   phenotype  
 Reduction in IR-induced 53BP1 foci in HeLa   cell  
 Impaired cell migration with increased protrusive activity   phenotype  
 C57Black/6   strain  
 nuclei stay close together  
 Retinal cone dystrophy 3B   disease  
 segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses  
 Segawa syndrome autosomal recessive   phenotype  
 BRCA1   gene  
 Deafness, autosomal dominant 17 phenotype  
 cooked broccoli   compound

## Datasources

ZOOMA maps text to ontology terms based on curated mappings from selected datasources (more preferred), and by searching ontologies directly (less preferred). Here, you can select which curated datasources to use, optionally ranked in order of preference. You can also select



Welcome to the EMBL-EBI Ontology Lookup Service



Examples: [diabetes](#), [GO:0098743](#)

[Looking for a particular ontology?](#)

## Data Content

Updated 18 Feb 2021

07:58

- 260 ontologies
- 6,466,998 terms
- 31,530 properties
- 497,537 individuals

## About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the [Samples, Phenotypes and Ontologies Team \(SPOT\)](#) at EMBL-EBI.

## Related Tools

In addition to OLS the SPOT team also provides the [OxO](#), [Zooma](#) and [Webulous](#) services. [OxO](#) provides cross-ontology mappings between terms from different ontologies. [Zooma](#) is a service to assist in mapping data to ontologies in OLS and [Webulous](#) is a tool for building ontologies from spreadsheets.

## Report an Issue

For feedback, enquiries or suggestion about OLS or to request a new ontology please use our [GitHub issue tracker](#). For announcements relating to OLS, such as new releases and new features sign up to the [OLS announce mailing list](#)

## Tweets by [@EBIOLS](#)



**EBISpot OLS**  
[@EBIOLS](#)

A number of our users have custom installations of OLS, [OxO](#) and [Zooma](#). [@NicoMatentzogl](#) has created a page where you can tell us about your custom EBI Ontology Tools installation and your use case:  
[github.com/EBISpot/ontoto...](https://github.com/EBISpot/ontoto...)



**EBISpot/ontoto...**  
Configuration to ...  
[github.com](https://github.com)

---

# **Exercise:**

## **Find suitable ontologies for your data**

Try finding and deciding on suitable ontologies and terms to use for the data file

- **strain**, using OLS
- **dev\_stage**, using Zooma
- **tissue\_type**, using FAIRsharing.org

# Update data dictionary

	A	B	C	D	E	F
1	<b>Current variable name</b>	<b>ENA Variable name</b>	<b>Measurement</b>	<b>Allowed values</b>	<b>Definition</b>	<b>Description</b>
2	animal ID					
3	date			format: YYYY-MM-DD, >=proj_start_date & <=today	Date of experiment ???	
4	mouse line	sub_strain				
5	strain	strain		NCIT ontology: C56BL/6 Mouse (NCIT:C14424), BALB/cJ Mouse (NCIT:C14657)	The mouse strain of the animal	
6	age		days (?)		Age of animal	
7	developmental stage	dev_stage		BTO ontology: pup (BTO:0004377), adult (BTO:0001043), embryo (BTO:0000379)		
8	sex	sex		male, female, unknown	Sex of the animal	
9	organism part	tissue_type		MA ontology: lung (MA:0000415), brain (MA:0000168)		
10	genotype					
11	experiment type					
12	experiment reference					
13	researcher					

- 
- Information about data is called **metadata**
  - Good metadata is a necessity for understanding the data - FAIRness
  - Try to be **consistent** when describing data
  - Use **controlled vocabularies** and **ontologies** when specifying metadata
  - **Metadata standards** - generic and domain specific
  - Use **data dictionaries** to document standards for your data
  - There are tools to help you decide on ontologies and terms to use