

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 person’s metadata, is another person’s 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



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

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

	A	B	C	D	E
1	sample id	patient id	sex	date	geographic location
2	PE300_COVseq_OAS-1	OAS-1	female	31 March	Italy, Turin, Nizza Mille
3	PE150_COVseq_OAS-1	OAS-1	Female	32 March	Italy, Turin, Nizza Mille
4	NEBNext_OAS-1	OAS-1	female	33 March	Italy, Turin, Nizza Mille
5	PE300_COVseq_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin
6	PE150_COVseq_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin
7	NEBNext_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin
8	PE300_COVseq_OAS-11	OAS-11	male	2020-03-31	Italy, Turin, Piemonte
9	PE150_COVseq_OAS-11	OAS-11	Male	2020-03-31	Italy, Turin, Piemonte
10	NEBNext_OAS-11	OAS-11	Male	2020-03-31	Italy, Turin, Piemonte

[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 different medical conditions do you think this list of terms describes?

Bloodstream Infection, Circulatory Failure, Toxic Shock Syndrome, Pyemia, Circulatory Collapse, Blood Poisoning, Endotoxin Shock, Pyohemia, Hypovolemic Shock, Septicemia, Sepsis-associated hypotension, Pyaemia

Sepsis	Shock	Septic shock
Blood Poisoning	Circulatory Collapse	Endotoxin Shock
Bloodstream Infection	Circulatory Failure	Sepsis-associated hypotension
Pyæmia	Hypovolemic Shock	Toxic Shock Syndrome
Pyemia		
Pyohemia		
Septicemia		

- 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

Sepsis

Definition: Systemic inflammatory response syndrome with a proven or suspected infectious etiology.

Synonyms: Blood Poisoning, Bloodstream Infection, Pyaemia, Pyemia, ...

MeSH Unique ID: D018805

- 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  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  Copy

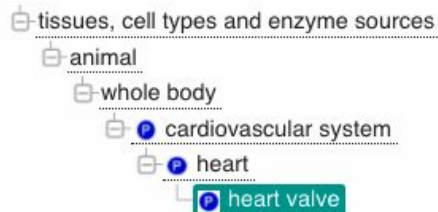
Search


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.

<https://xkcd.com/927/>

CC BY-NC 2.5

- 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

Term Name: creator	
URI:	http://purl.org/dc/elements/1.1/creator
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.
Term Name: date	
URI:	http://purl.org/dc/elements/1.1/date
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] http://www.w3.org/TR/NOTE-datetime
Term Name: description	
URI:	http://purl.org/dc/elements/1.1/description
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.
Term Name: format	
URI:	http://purl.org/dc/elements/1.1/format
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] http://www.iana.org/assignments/media-types/

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

[CC BY 3.0](#)

- *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

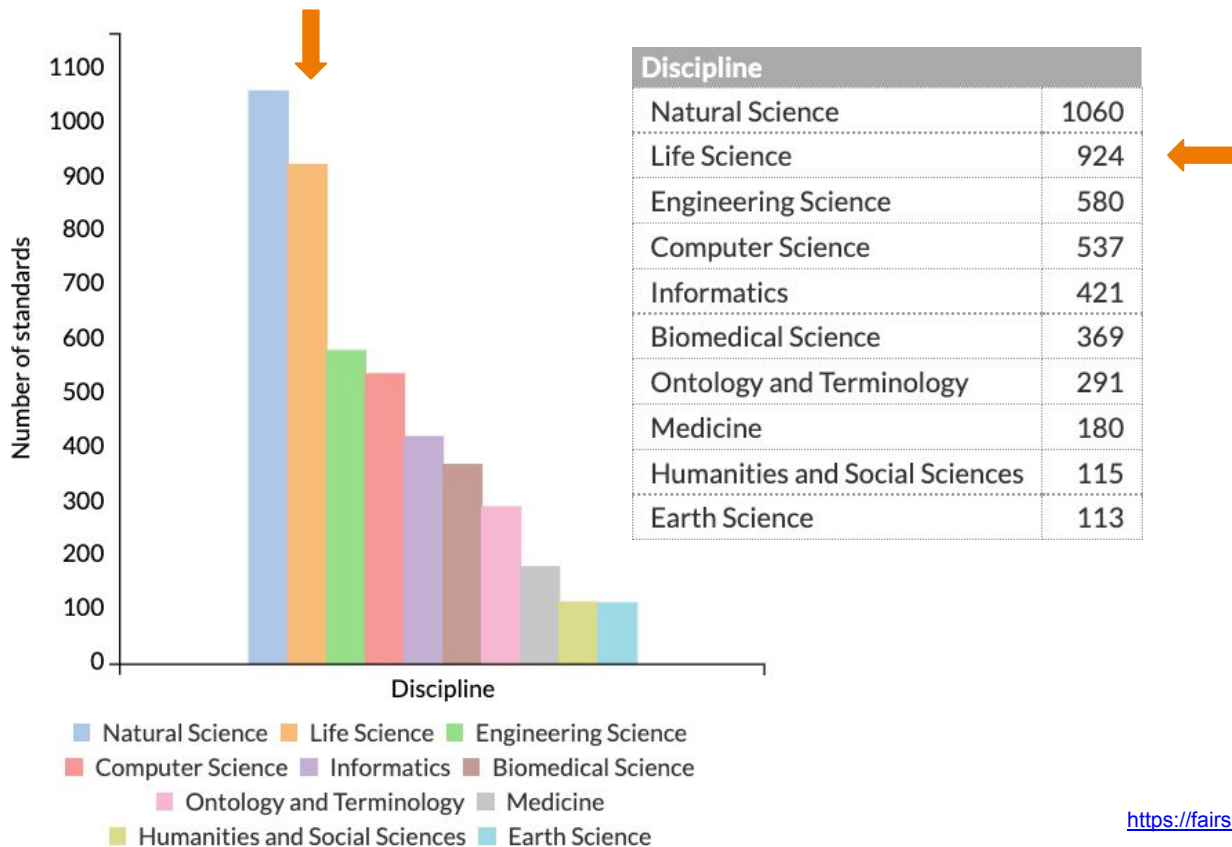
Filter fields... 

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?



-
- 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

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	sample id					
3	patient id					
4	sex					
5	date					
6	geographic location					
7	age					
8	health state					
9	symptoms					
10	disease outcome					
11	tissue					

2. [data_dictionary](#)

1. [samples_metadata_lesson.csv](#)

4. Copy headings from [samples_metadata_lesson.csv](#) to rows in [data_dictionary](#)

- Add some definitions
- Add some units
- Add some allowed value definitions

	A	B	C	D	E	F	G	H	I	J
1	sample id	patient id	sex	date	geographic location	age	health state	symptoms	disease outcome	tissue
2	PE300_COVseq_OAS-1	OAS-1	female	31 March	Italy, Turin, Nizza Millefonti	48	ill	fever, sore throat	dead	laryngopharynx
3	PE150_COVseq_OAS-1	OAS-1	Female	32 March	Italy, Turin, Nizza Millefonti	48	ill	fever, sore throat	dead	laryngopharynx
4	NEBNext_OAS-1	OAS-1	female	33 March	Italy, Turin, Nizza Millefonti	48	ill	fever, sore throat	dead	laryngopharynx
5	PE300_COVseq_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin	35		N/A		lung
6	PE150_COVseq_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin	35		N/A		lung
7	NEBNext_OAS-10	OAS-10	male	2020-03-31	Italy, Turin, Turin	35		N/A		lung
8	PE300_COVseq_OAS-11	OAS-11	male	2020-03-31	Italy, Turin, Piemonte	59	healthy	N/A	healthy	nasopharynx
9	PE150_COVseq_OAS-11	OAS-11	Male	2020-03-31	Italy, Turin, Piemonte	59	healthy	N/A	healthy	nasopharynx
10	NEBNext_OAS-11	OAS-11	Male	2020-03-31	Italy, Turin, Piemonte	59	healthy	N/A	healthy	nasopharynx
11	PE300_COVseq_OAS-12	OAS-12	female	2020-03-31	Italy, Turin, Turin	60	healthy	N/A	healthy	nasopharynx
12	PE150_COVseq_OAS-12	OAS-12	female	2020-03-31	Italy, Turin, Turin	60	healthy	N/A	healthy	nasopharynx
13	NEBNext_OAS-12	OAS-12	female	2020-03-31	Italy, Turin, Turin	60	healthy	N/A	healthy	nasopharynx
14	PE300_COVseq_OAS-13	OAS-13	female	31/3/2020	Italy, Turin, Torino	83	ill	fatigue, loss of taste	dead	laryngopharynx
15	PE150_COVseq_OAS-13	OAS-13	female	31/3/2020	Italy, Turin, Torino	83	ill	fatigue, loss of taste	dead	laryngopharynx
16	NEBNext_OAS-13	OAS-13	female	31/3/2020	Italy, Turin, Torino	83	ill	fatigue, loss of taste	dead	laryngopharynx
17	PE300_COVseq_OAS-14	OAS-14	Male	4/1/2020	Italy, Turin, Campidoglio	21	ill	fever	dead	laryngopharynx
18	PE150_COVseq_OAS-14	OAS-14	M	4/1/2021	Italy, Turin, Campidoglio	21	ill	fever	dead	laryngopharynx
19	NEBNext_OAS-14	OAS-14	M	4/1/2022	Italy, Turin, Campidoglio	21	ill	fever	dead	laryngopharynx

Data dictionary - start

	A	B	C	D	E	F
1	Current variable name	ENA variable name	Measurement unit	Allowed values	Definition	Description
2	sample id					
3	patient id					
4	sex			male, female, unknown	Sex of individual	
5	date			format: YYYY-MM-DD, >=proj_start_date & <=today	Date of sampling	
6	location					
7	age		years		Age of the individual at	
8	health state				Health state of individual at	
9	symptoms			fever, sore throat, fatigue, loss of taste, not applicable	Symptoms experienced in connection with illness	
10	disease outcome			healthy, dead	Final outcome of disease	
11	tissue				Tissue sampled	
12						

-
- 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 **ERC000033 ENA virus pathogen reporting standard 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.
 - a. Are all mandatory fields present, or will you need to add fields?
 - b. Are there fields that need to be split into more fields?
 - c. Are there controlled vocabularies you should adhere to?

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 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

Intraspecies
information

Associated host
information

host details

Environmental
information


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	1	Current variable nan	ENA variable name	Measurement unit
hospitalisation	2	sample id	host subject id	Allowed values
illness duration	3	patient id	host sex	male, female, not collected
illness symptoms	4	sex	collection date	format: YYYY-MM-DD, >=proj_start_date & <=today
collection date	5	date	geographic location (country	<country>
geographic location (country and/or sea)	6	location	geographic location (region	<region>, <city>, ...
geographic location (latitude)	7	age	host age	years
geographic location (longitude)	8	health state	host health state	diseased, healthy, not applicable, not collected, not provided, restricted access
geographic location (region and locality)	9	symptoms	illness symptoms	fever, sore throat, fatigue, loss of taste, not applicable
	10	disease outcome	host disease outcome	recovered, dead
	11	tissue	isolation source host-associated	
	12	isolate	isolate	
	13			
	14			
		restricted text	regular expression	recommended DD
		free text	recommended	

<https://www.ebi.ac.uk/ena/browser/view/ERC000033>

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 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... 	Field Name	Field Format (Field Restriction)	Requirement	(Units)
Filter by type:				
Human surveillance data	subject exposure	free text	optional	
Collection event information	subject exposure duration	free text	optional	
sample collection	type exposure	free text	optional	
host disorder	personal protective equipment	free text	optional	
host description	hospitalisation	text choice options	optional	
Virus isolate information	illness duration	free text	optional	
General collection event information	illness symptoms	free text	optional	
Serology detection	collection date	restricted text regular expression	recommended	
Intraspecies information	geographic location (country and/or sea)	text choice options	mandatory	
Associated host information	geographic location (latitude)	restricted text regular expression	recommended	DD
host details	geographic location (longitude)	restricted text regular expression	recommended	DD
Environmental information	geographic location (region and locality)	free text	recommended	

- This standard is 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://www.ebi.ac.uk/ontology-tooling/)
 - [Zooma](https://zooma.ebi.ac.uk/) - map free text to ontology terms
 - [Ontology Lookup Service - OLS](https://ontology.ebi.ac.uk/ols/)



- 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

<https://fairsharing.org/>

[CC BY-SA 4.0](#)

ZOOMA

ONTOLOGY ANNOTATION

Home | Explore | Help | About ZOOMA

What's this? ⓘ

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
 Senawa syndrome autosomal recessive phenotype

Annotate Clear

 [Configure Datasources](#)

Zooma is a tool for mapping free text annotations to ontology term based on a curated repository of annotation knowledge.

Where mappings are not found in the curated repository one or more ontologies can be selected from the [Ontology Lookup Service](#) to increase coverage. For example if you want to map GWAS annotations select the GWAS datasource and a common disease ontology such as EFO or DOID to maximise coverage when terms have no curated mappings.

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](#).

<https://www.ebi.ac.uk/training/online/courses/cellular-microscopy-phenotype-ontology-quick-tour/annotating-data-with-cmpo/>

CC BY 4.0

<https://www.ebi.ac.uk/spot/zooma/>



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...



EBISPOT/ontoto...
Configuration to ...
github.com

<https://www.ebi.ac.uk/ols/>

Exercise:

Find suitable ontologies for your data

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

- **illness symptoms**, using OLS
- **isolation source host-associated**, using FAIRsharing.org

	A	B	C	D	E	F
1	Current variable name	ENA variable name	Measurement unit	Allowed values	Definition	Description
2	sample id					
3	patient id	host subject id				
4	sex	host sex		male, female, not collected	Sex of individual	
5	date	collection date		format: YYYY-MM-DD, >=proj_start_date & <=today	Date of sampling	
6	location	geographic location (country and/or sea)		<country>		
7		geographic location (region and locality)		<region>, <city>, ...		
8	age	host age	years		Age of the individual at	
9	health state	host health state		diseased, healthy, not applicable, not collected, not provided, restricted access	Health state of individual at time of sampling	
10	symptoms	illness symptoms		NCIT ontology: Fever (NCIT:C3038), Sore Throat (NCIT:C50747), Fatigue (NCIT:C3036), Ageusia (NCIT:C116374), not applicable		
11	disease outcome	host disease outcome		recovered, dead	Final outcome of disease	
12	tissue	isolation source host-associated		FMA ontology: Laryngopharynx (FMA:54880), Nasopharynx (FMA:54878), Lung (FMA:7195)	Tissue sampled	
13	experiment type					
14	isolate	isolate			individual isolate from which the sample was obtained	
15						

-
- 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