Research Data Management for Bioimage Data at the HHU

Metadata Curation:
What are ontologies?
Why and how to use them?

Tom Boissonnet

I3D:bio
Information Infrastructure for BioImage Data

Adapted from: Schmidt C., Bortolomeazzi M., Boissonnet T., Fortmann-Grote C. *et al.* (2023). I3D:bio's OMERO training material: Re-usable, adjustable, multi-purpose slides for local user training. Zenodo. DOI: 10.5281/zenodo.8323588

Heinrich Heine Universität Düsseldorf

Metadata details in form of Key-Value Pair annotation

Key-Value Pairs allow (standardized) annotation of detailed metadata

Consists of

- **Key:** Denotes a real-world object or an abstract concept that can be assigned a specific value (of different possible values)
- **Value:** Number or text string that specifies the object denoted under „Key“

Examples:

Key: „cell type“ **Value:** „CD4+ T cell“

Key: „disease model“ **Value:** „Experimental Autoimmune Encephalomyelitis“

Standardize Key-Value pairs?

Key: „cell type“ **Value:** „CD4+ T cell“

Key: „disease model“ **Value:** „Experimental Autoimmune Encephalomyelitis“

„cell type“ „type of cell“ „cell-type“ „cellular entity“ „cellular identity“

„CD4+ T cell“ „CD4-positive T-lymphocyte“ „naive, CD4-positive T cell“ ???

„CD4-positive, alpha-beta T cell“ „Th0 cell“ „CD4+ T helper cell“

„Experimental Autoimmune Encephalomyelitis“ „EAE“ „Allergic Encephalomyelits“

How to avoid ambiguity?
How to describe the data objectively?
How to make the metadata machine-interpretable?

Controlled vocabularies

A **controlled vocabulary** provides a list of terms.

- a definition of each term
- a unique identifier of each term
- different types exist, e.g.,
  - Alphabetical list
  - Thesaurus (a collection of synonyms)
  - Taxonomy (hierarchical or network-like list of terms)
  - (ontology)

→ Allows standardized usage of terms

https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/controlled-vocabularies/

Controlled vocabularies – *example: MeSH*

Medical Subject Headings (MeSH)

Controlled vocabulary in the form of a thesaurus

curated by the National Library of Medicine (US)

National Library of Medicine
National Center for Biotechnology Information

Log in

MeSH

MeSH | Search

Limits Advanced

Help

Full

Send to:

CD4-Positive T-Lymphocytes

A critical subpopulation of T-lymphocytes involved in the induction of most immunological functions. The HIV virus has selective tropism for the T4 cell which expresses the CD4 phenotypic marker, a receptor for HIV. In fact, the key element in the profound immunosuppression seen in HIV infection is the depletion of this subset of T-lymphocytes.

Tree Number(s): A11.118.637.555.567.569.200, A15.145.229.637.555.567.569.200, A15.382.490.555.567.569.200
MeSH Unique ID: D015496
Entry Terms:

- CD4 Positive T Lymphocytes
- CD4-Positive T-Lymphocyte
- T-Lymphocyte, CD4-Positive
- T-Lymphocytes, CD4-Positive
- CD4-Positive Lymphocytes
- CD4-Positive Lymphocyte
- Lymphocyte, CD4-Positive
- Lymphocytes, CD4-Positive
- T4 Cells
- T4 Cell
- T4 Lymphocytes
- T4 Lymphocyte

Previous Indexing:

- T Lymphocytes (1986-1988)

PubMed Search Builder

MeSH

Lymphocytes MeSH

"Lymphocytes"[MeSH Terms] (1) MeSH

house mouse (1) Taxonomy

rat (2) Taxonomy

See more...

https://www.ncbi.nlm.nih.gov/mesh/ (last access: 2023-08-17)

hhu.

Use of controlled vocabularies in practice

Key: „cell type“ **Value:** „CD4+ T cell“

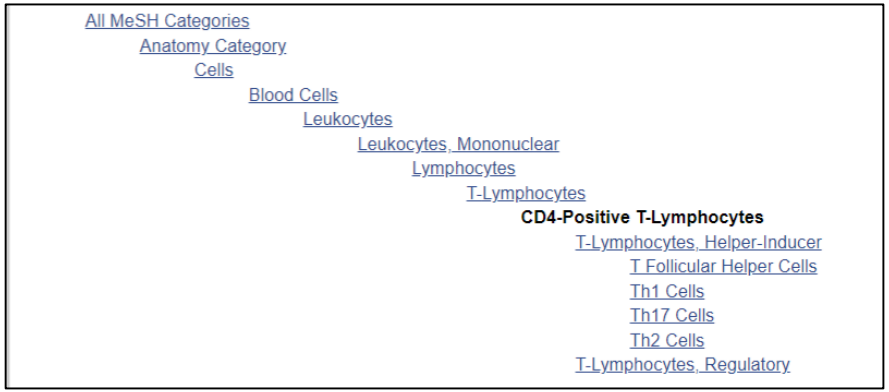
„CD4+ T cell“ „CD4-positive T-lymphocyte“ „Th0 cell“ „naive, CD4-positive T cell“

Example of controlled vocabulary usage in Key-Value Pairs:

Key: cell type **Value:** CD4-positive T-lymphocyte
Key: cell type term accession number **Value:** http://id.nlm.nih.gov/mesh/D015496

→ **The term is defined**
→ **Some hierarchical information is contained in MeSH** *see* →

→ Attributes / properties / relationships are missing



https://www.ncbi.nlm.nih.gov/mesh/68015496 (last access: 2023-08-17)

Ontologies

An **ontology** is a conceptual framework of how specific terms are used to represent *domain knowledge* in a (research) domain.

- Defines term attributes/properties, and relationships between the terms
- Terms with shared attributes are grouped into classes
- Terms in different ontologies are mapped to each other or adopted
- Can be extended over time with the evolving domain knowledge (i.e., an ontology is versioned)
- Formalized, i.e., ontologies can be expressed in ontology formats (machine-interpretable)

Examples of Ontologies:

- Experimental Factor Ontology (EFO) – curated by the EMBL EBI
- Biological Imaging Methods Ontology (FBbi) – curated by the Cell Image Library
- Cell Line Ontology (CLO) – community-based, curated at the University of Michigan

Browse different ontologies: https://bioportal.bioontology.org/

Ontologies consist of classes with attributes

Class	Represents a real-world object (e.g., „microscope objective lense“) or an abstract concept (e.g., „disease model“) A class comprises subclasses or individual terms (instances) sharing attributes. Classes have specific relationships with each other.
Attribute	Specific property of a class (can be in form of Key-Value Pairs), e.g.: **Key: Definition** **Value: „This is the term definition (and a reference to a paper that first described it).“**
Relationship	Relationship between classes

Note: The Key-Value Pairs in OMERO are not the same as the Key-Value Pairs for ontology class attributes. Both use the same concept independently.

Use of ontologies in practice

Key: „cell type“ **Value:** „CD4+ T cell“

„CD4+ T cell“ „CD4-positive T-lymphocyte“ „naive, CD4-positive T cell“

„CD4-positive, alpha-beta T cell“ „Th0 cell“ „CD4+ T helper cell“

Example of ontology usage in Key-Value Pairs:

Key: cell type **Value:** CD4-positive, alpha-beta T cell

Key: cell type term accession number **Value:** http://purl.obolibrary.org/obo/CL_0000624

Several ontologies can use the *same term,* e.g.:

- Experimental Factor Ontology (EFO)
- Cell Ontology (CL)
- Uber Anatomy Ontology (UBERON)
- others

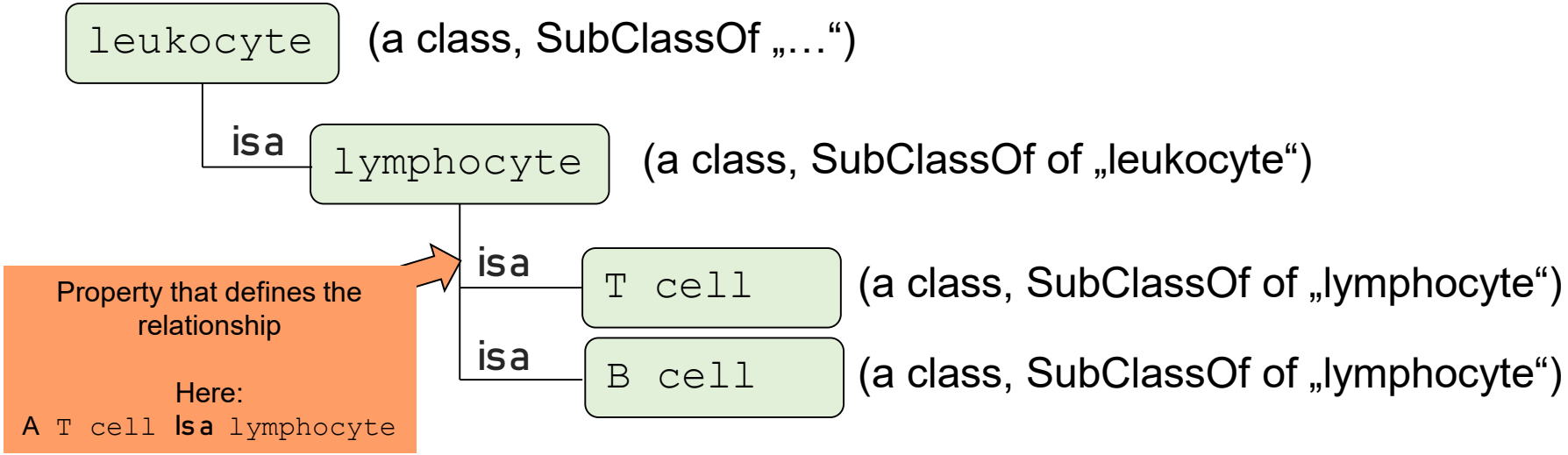
Why are there many different ontologies?

Look up ontologies: https://bioportal.bioontology.org/

Why are there so many ontologies?

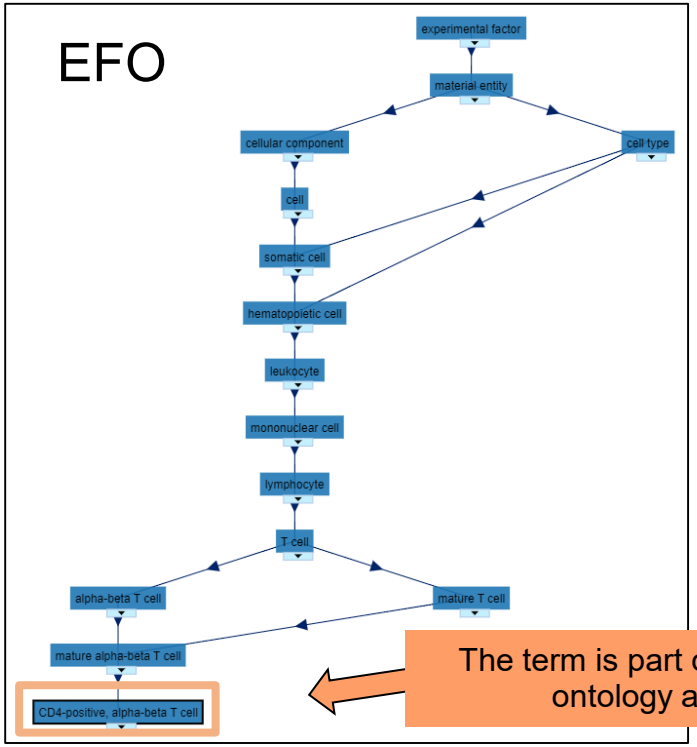
Different ontologies are designed to optimally ***represent their respective domain knowledge*** (for example, the relationship between terms)

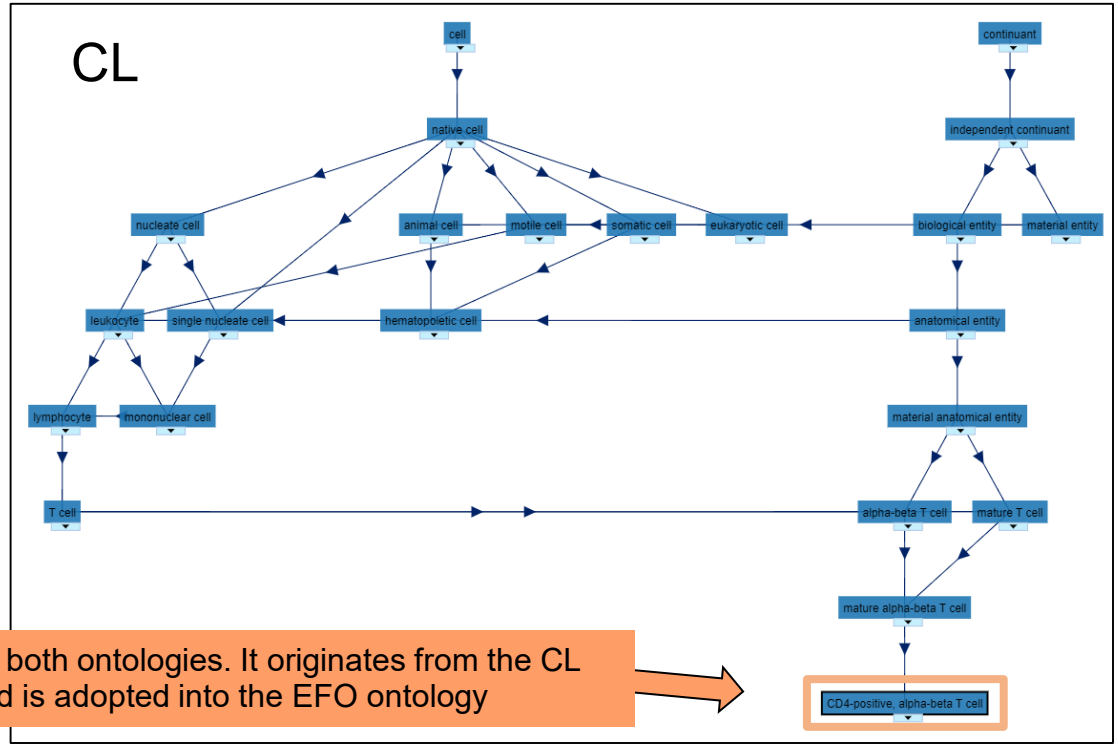
This knowledge can be represented as a tree structure or „knowledge graph“.
Example:



Graph visualizations of different ontologies

Term: CD4-positive, alpha-beta T cell; http://purl.obolibrary.org/obo/CL_0000624



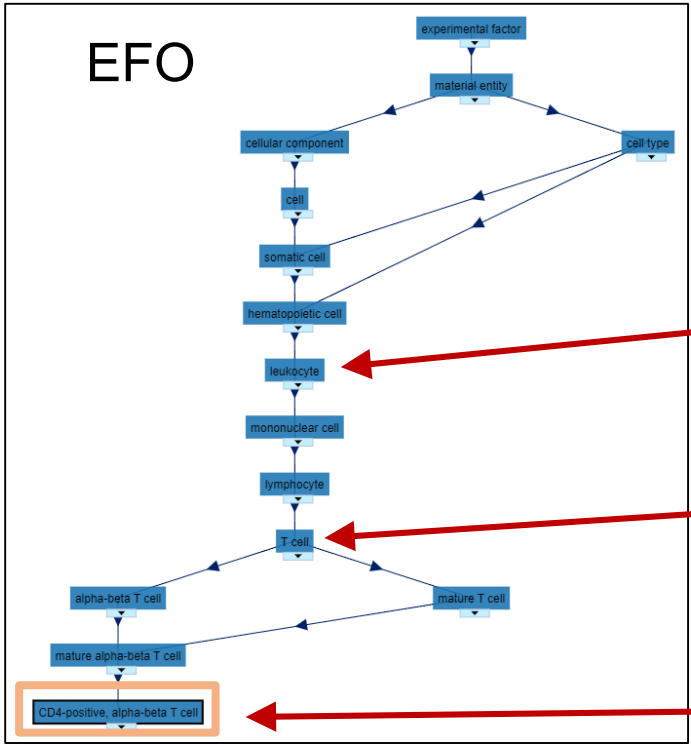


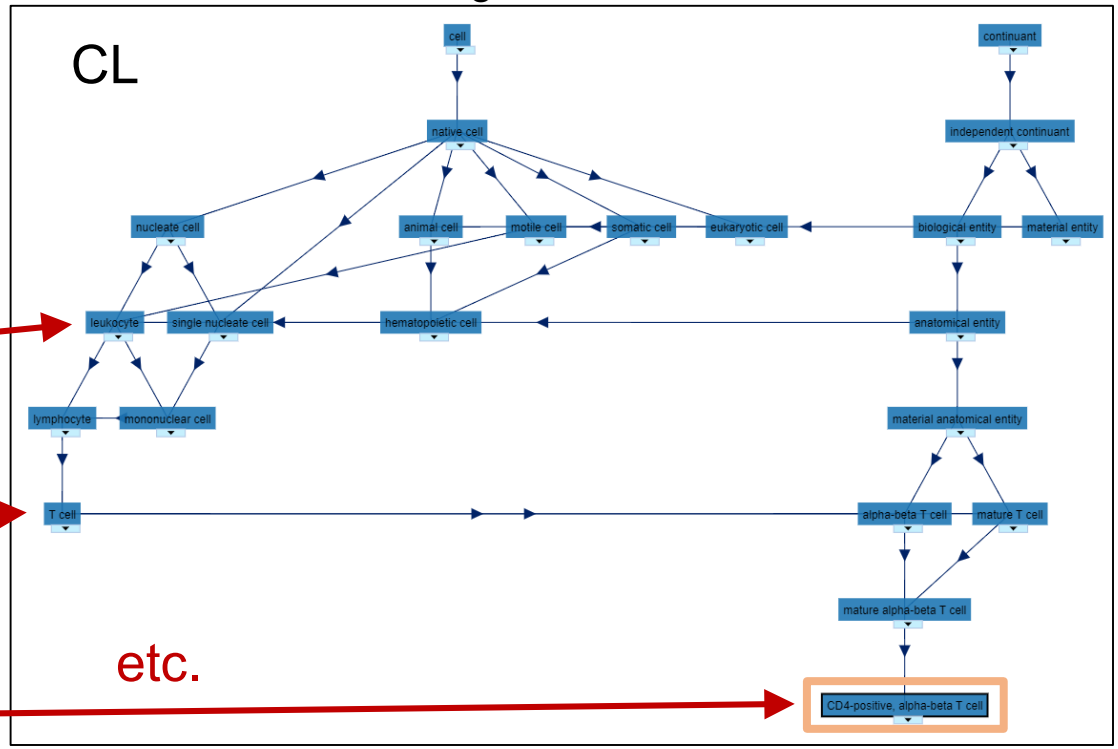
The term is part of both ontologies. It originates from the CL ontology and is adopted into the EFO ontology

Both visualizations taken from: https://bioportal.bioontology.org/

Mapping between ontologies

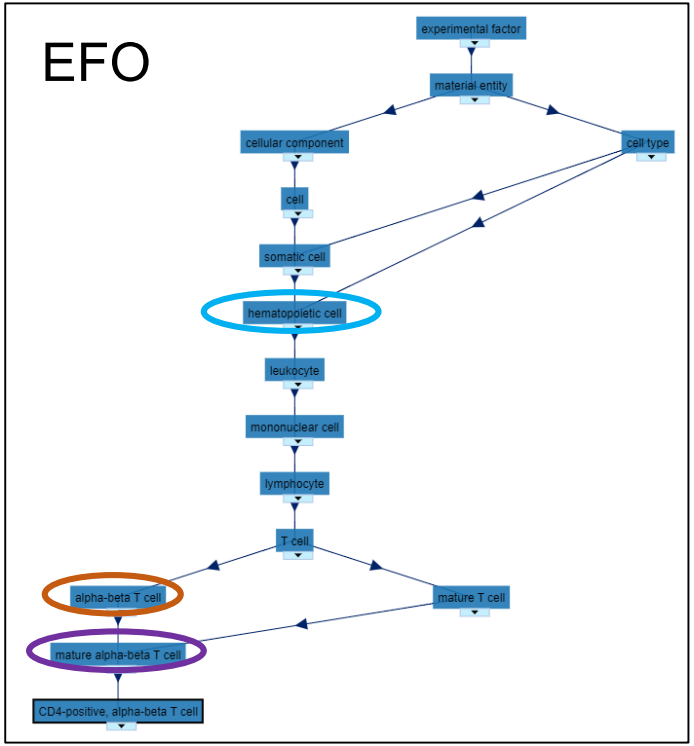
Terms are adopted from other ontologies, or synonyms in different ontologies are mapped to each other.
→ Semantic knowledge *across* domains!





Both visualizations taken from: https://bioportal.bioontology.org/

The advantage of using ontologies



A single Key-Value Pair can carry extended domain knowledge!

„CD4-positive, alpha-beta T cell“ following an ontology (here: EFO) includes more information from the domain knowledge formalized in the ontology (and cross-domain knowledge formalized by mapping):

- Is carrying a T cell receptor with αβ-chains
- Has completed thymic selection (i.e., is mature)
- Is a cell of the hematopoietic system
- etc…

Due to the ontology format, a computer can read the knowledge!

Visualization taken from: https://bioportal.bioontology.org/

Using ontologies in OMERO 1/2

There is no unified standard for the use of ontologies in OMERO.
But we can start working with some recommendations.

Suggestion *(based on REMBI[1] items, and ISA-TAB[2]):*
To create machine-actionable metadata, make use of **ontology terms** and **ontology term source references**:

- Use the ontology-derived term as the Value for a specific Key
- Add the ontology term URL as the Value for a second Key using the <Key> + „Term Accession Number“

Example

Key: Biological entity
Key: Biological entity Term Accession Number

Value: CD4-positive, alpha-beta T cell
Value: http://purl.obolibrary.org/obo/CL_0000624

1) Sarkans et al., 2021: https://doi.org/10.1038/s41592-021-01166-8
2) https://isa-specs.readthedocs.io/en/latest/isatab.html (last access: 2023-08-16)

Using ontologies in OMERO 2/2

When and why to include the ontology source reference?

Ontologies allow for *cross-domain* referencing. I.e., a specific term in one ontology may be adopted from another ontology.

How do you know? Example:
A term was chosen from EFO ontology but the term ID implies CL ontology:
http://purl.obolibrary.org/obo/CL_0000624

Term ID points to CL (not EFO)

Solution? Include the ontology source URL:

Example

Key: Biological entity
Key: Biological entity Term Accession Number
Key: Biological entity Term Source REF

Value: CD4-positive, alpha-beta T cell
Value: http://purl.obolibrary.org/obo/CL_0000624
Value: http://www.ebi.ac.uk/efo/efo.owl

https://isa-specs.readthedocs.io/en/latest/isatab.html (last access: 2023-08-16)

Benefits of using ontologies for data annotation

Pre-publication benefits:

- Avoid term ambiguity in collaborative research settings
- Use ontology identifiers for automated image analysis workflows
- Enable semantic search in your own dataset(s)

Post-publication benefits:

- Your publication & data are more likely to be found by others, which can increase citations and facilitate collaboration
- Your data is retrievable for semantic search across domains and can thus generate a higher scientific impact

Getting started with ontologies – Open Biological and Biomedical Ontologies (OBO) Acadamy

OBO Semantic Engineering Training

Search

GitHub 59 32

Introduction to ontologies

Based on CL editors training by David Osumi-Sutherland

Why do we need ontologies?

We face an ever-increasing deluge of biological data analysis. Ensuring that this data and analysis are Findable, Accessible, Interoperable, and Re-usable (FAIR) is a major challenge. Findability, Interoperabiltiy, and Resuability can all be enhanced by standardising metadata. Well-standardised metadata can make it easy to *find* data and analyses despite variations in terminology ('Clara cell' vs 'nonciliated bronchiolar secretory cell' vs 'club cell') and precision ('bronchial epithelial cell' vs 'club cell'). Understanding which entities are referred to in metadata and how they relate to the annotated material can help users work out if the data or analysis they have found is of interest to them and can aid in its re-use and interoperability with other data and analyses. For example, does an annotation of sample data with a term for breast cancer refer to the health status of the patient from which the sample was derived or that the sample itself comes from a breast cancer tumor?

We can't find what we're looking for

Given variation in terminology and precision, annotation with free text alone is not sufficient for findability. One very lightweight solution to this problem is to rely on user-generated keyword systems, combined with some method of allowing users to choose from previously used keywords. This can produce some degree of annotation alignment but also results in fragmented annotation and varying levels of precision with no clear way to relate annotations.

For example, trying to refer to feces, in NCBI BioSample:

Query	Records
Feces	22,592

https://oboacademy.github.io/obook/explanation/intro-to-ontologies/ (last access: 2023-08-16)

Getting started with ontologies – FAIR Cookbook

FAIRCOOKBOOK

GITHUB

4. Introduction to terminologies and ontologies

Recipe Overview

Reading Time
15 minutes

Executable Code
No

Difficulty

Introducing terminologies and ontologies

Recipe Type
Survey / Review

Audience
Data Curator, Data Manager, Data Scientist

Maturity Level & Indicator
not applicable

Cite me with FCB019

4.1. Main objectives

The aim of this recipe is to provide a compact introduction about controlled terminologies and ontologies, why these resources are central to the preservation of knowledge and data mining and how such resources are developed.

4.2. Controlled terminology or ontology: what's the difference?

The need for controlled vocabulary often arises in situations where validation of textual information is necessary for operational requirements. The main initial driver for data entry harmonization is to increase query recall. In its most basic form, keywords may be used to perform indexation. However, if relying on user input alone, the chances of typographic errors increases with the number of users. These unavoidable events accumulate over time and end up hurting the accuracy of search results and this is the reason for offering sets of predefined values. It reduces the noise. However, this can come at the cost of precision, as the predefined terms may not cover the exact thing users may need to describe. Furthermore, term mis-selection by the user is not eliminated and introduces another type of error.

A controlled terminology is a *normative* collection of terms, the spelling of which is fixed and for which additional information may be provided such as a definition, a set of synonyms, an editor, a version, as well as a license determining the condition of use. The set of information about a specific controlled terminology term is designated as term metadata. In a controlled terminology, terms appear as a flat list, meaning that no relationship between any of the entities the controlled terminology represents is captured in any formal way. This is the main drawback and limitation of controlled terminologies, which are often developed to support a data model or an application.

https://faircookbook.elixir-europe.org/content/recipes/interoperability/introduction-terminologies-ontologies.html

hhu.

Getting started with ontologies – BioPortal BioOntology

Check out:
- The Annotator
- The Recommender

BioPortal Ontologies Search Annotator Recommender Mappings

Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma

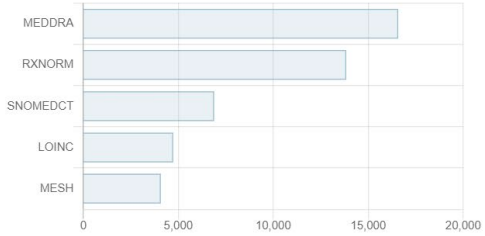
Advanced Search

Find an ontology

Start typing ontology name, then choose from list

Browse Ontologies

Ontology Visits (June 2023)



More

BioPortal Statistics

Ontologies	1,062
Classes	15,915,705
Properties	36,286
Mappings	79,636,946

https://bioportal.bioontology.org/ (last access: 2023-08-16)

Getting started with ontologies – Ontology Lookup Service (by EMBL-EBI)

OLS ONTOLOGY SEARCH

Home | Ontologies | Help | About | Downloads

Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... **Search**

☐ Exact match ☐ Include obsolete terms ☑ Include imported terms

Examples: diabetes, GO:0098743

Looking for a particular ontology?

Data Content

Updated 29 Jul 2023 Sat 08:07(+02:00)

- 245 ontologies
- 7.839.758 classes
- 42.825 properties
- 22.956 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 and ZOOMA services. OxO provides cross-ontology mappings between terms from different ontologies. ZOOMA is a service to assist in mapping data to ontologies in OLS.

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.

Check out:

- The Related Tools

https://www.ebi.ac.uk/ols4

https://www.ebi.ac.uk/ols4 (last access: 2023-08-16)

hhu.

Getting started with ontologies – ISA tools software suite

isatools | Software Suite | Model and Formats | Blog | Support | User Community and Publications

The open source **ISA framework** and tools help to manage an increasingly diverse set of life science, environmental and biomedical experiments that employing one or a combination of technologies.

Built around the '**I**nvestigation' (the project context), '**S**tudy' (a unit of research) and '**A**ssay' (analytical measurement) data model and serializations (tabular, JSON and RDF), the ISA framework helps you to provide rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are reproducible and reusable.

Collect and curate, following standards
Describe the experimental steps using community-defined minimum reporting requirements and ontologies, where possible.

Store and browse, locally or publicly
Create your own repository to search and browse the experimental description and associated data, hosted openly or privately.

Submit to public repositories
When required, reformat experiments for submission to supported public repositories or directly export to those already using ISA formats.

Analyse with existing tools
Upload experimental descriptions and associated data to a growing number of well-known analysis systems that ISA formats connect with.

Release, reason and nanopublish
Explore and reason over your experiments, open them to the linked data universe, or publish nano-statements of your discoveries.

Publish data alongside your article
Directly export your experiments to a new generation of data journals that are accepting submissions in ISA formats.

Software tools (outside of OMERO) for metadata annotation

https://isa-tools.org/ (last access: 2023-08-17)

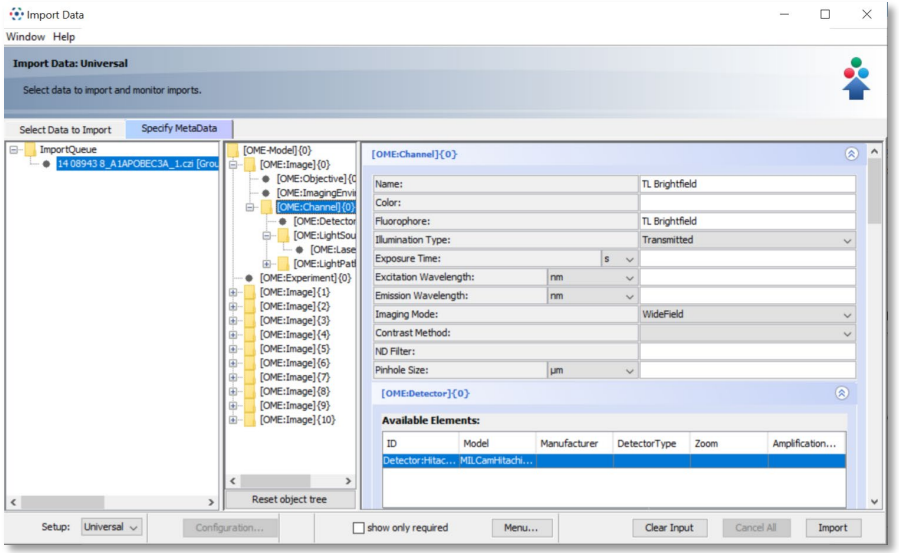
MDE.mic (OMERO.mde) for ontology-compliant annotation

Intermediate step during the data import to OMERO:

Review and Annotate metadata using **OMERO.mde**, a metadata editor.

It allows to edit:

- metadata of individual files,
- metadata the import queue in batch,
- and is supported by standardized, but configurable metadata fields and ontology term look-up



Kunis et al., 2021, Nat Meth: https://doi.org/10.1038/s41592-021-01288-z