University of Basel

BIOZENTRUM
The Center for
Molecular Life Sciences

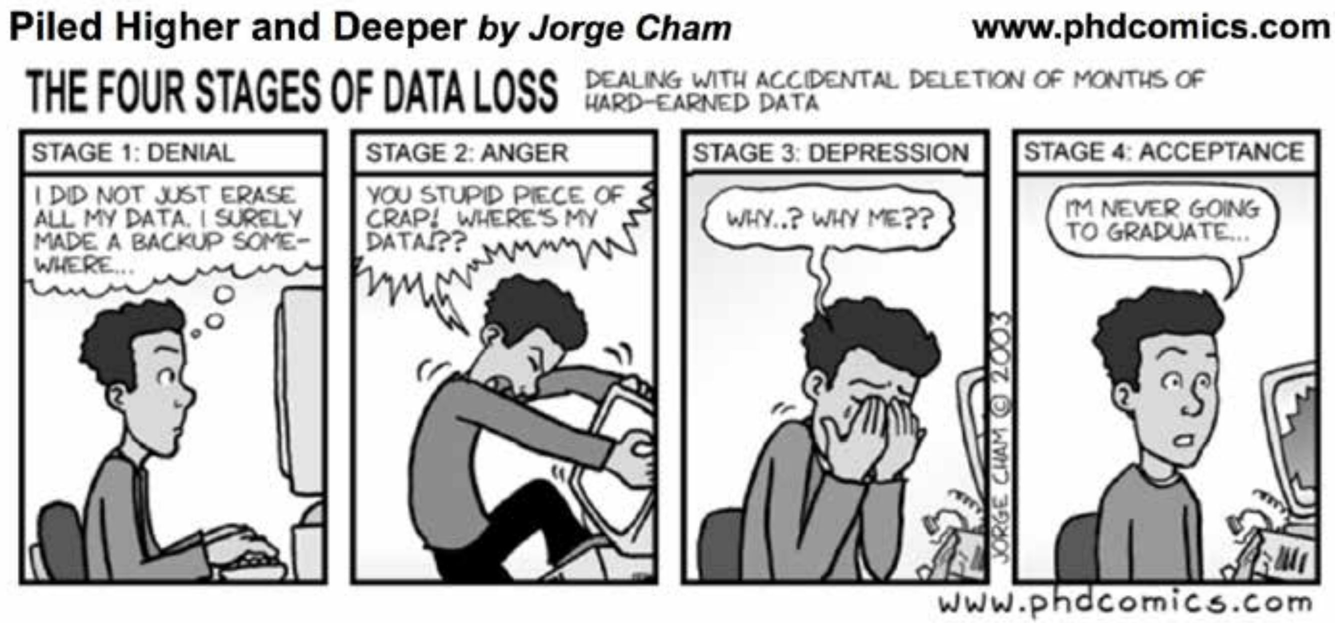
Metadata and controlled vocabularies – a treasure map to your data

Oliver Biehlmaier

Imaging Core Facility, Biozentrum, University of Basel

Why to manage and share your data

- You act **according to funding agencies requests**


Piled Higher and Deeper by Jorge Cham
www.phdcomics.com
THE FOUR STAGES OF DATA LOSS
DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA
STAGE 1: DENIAL
I DID NOT JUST ERASE ALL MY DATA. I SURELY MADE A BACKUP SOMEWHERE...
STAGE 2: ANGER
YOU STUPID PIECE OF CRAP! WHERE'S MY DATA!??
STAGE 3: DEPRESSION
WHY..? WHY ME??
JORGE CHAM © 2003
STAGE 4: ACCEPTANCE
I'M NEVER GOING TO GRADUATE...
www.phdcomics.com

title: "Stages of Data Loss" - originally published 10/30/2003

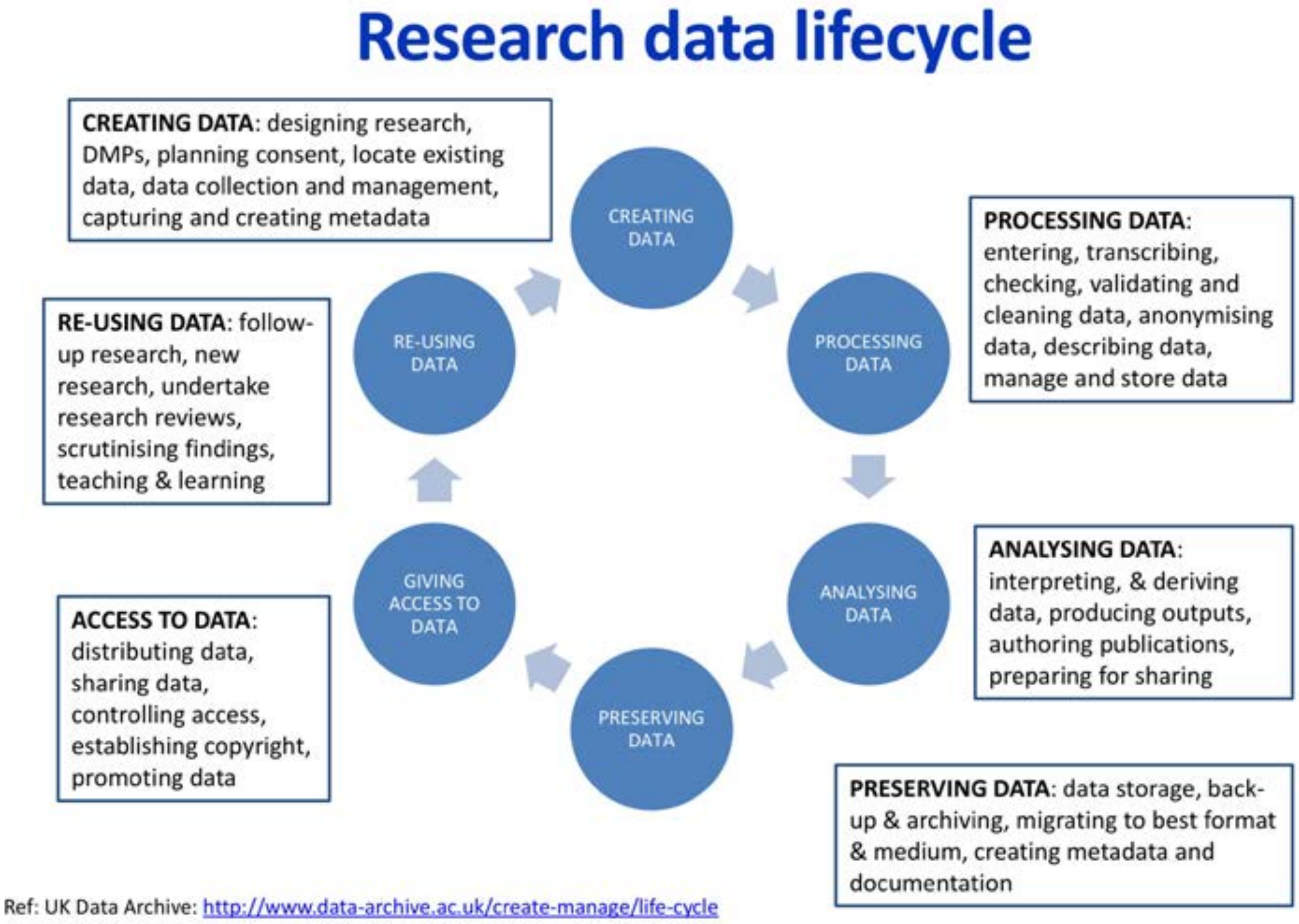
Enables more collaboration and advances research

Why to manage and share your data

- You act **according to funding agencies requests (SNF, ERC)**
  =>that's where the actual momentum comes from!
- **Find and understand your data when you need to use it**
- **Continuity if project staff leave or new researchers join**
- **Avoid unnecessary duplication** *(e.g. re-collecting or re-working data on different storage shares)*
- **Data underlying publications are maintained**, allowing for validation of results
- Enables **more collaboration and advances research**

How/where to start? Where to go?

- Think about your data creation process and beyond!



Research data lifecycle

CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata

PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage and store data

ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing

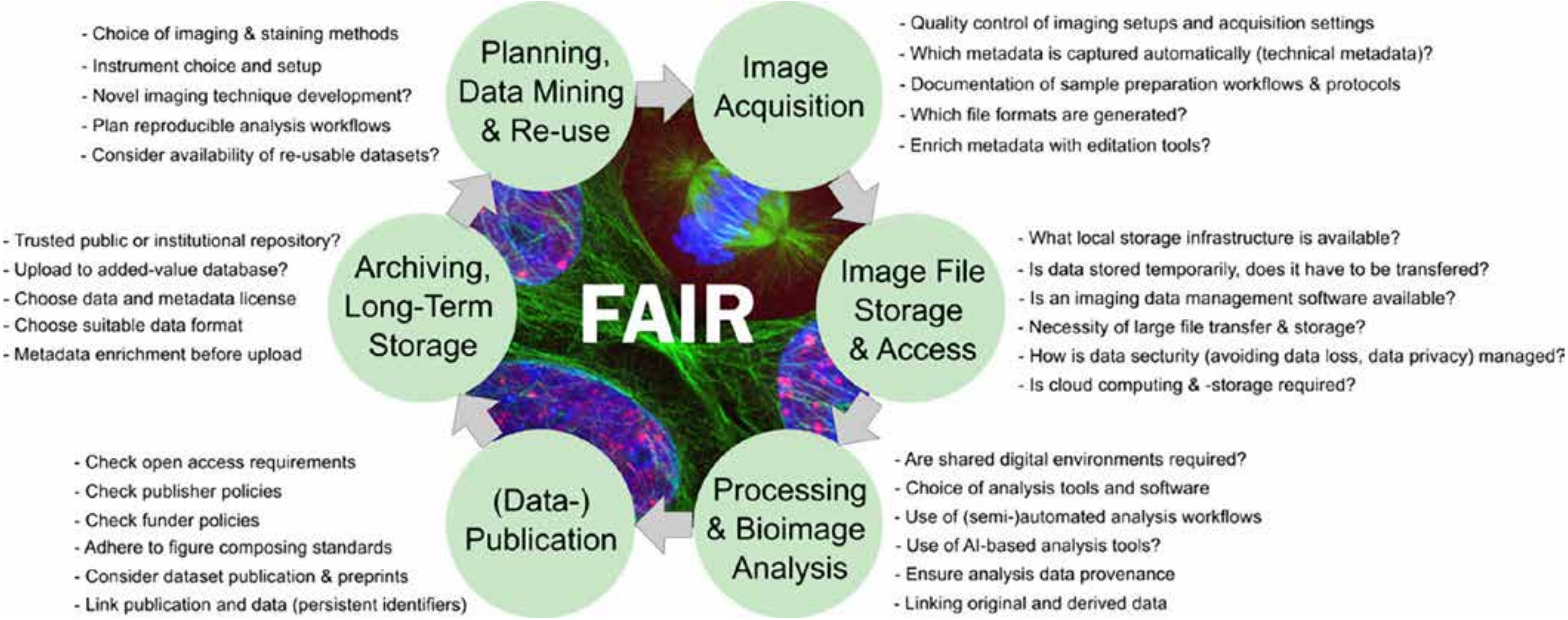
PRESERVING DATA: data storage, back-up & archiving, migrating to best format & medium, creating metadata and documentation

ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning

Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

Lifecycle in bioimaging



Planning, Data Mining & Re-use

- Choice of imaging & staining methods
- Instrument choice and setup
- Novel imaging technique development?
- Plan reproducible analysis workflows
- Consider availability of re-usable datasets?

Image Acquisition

- Quality control of imaging setups and acquisition settings
- Which metadata is captured automatically (technical metadata)?
- Documentation of sample preparation workflows & protocols
- Which file formats are generated?
- Enrich metadata with editation tools?

Image File Storage & Access

- What local storage infrastructure is available?
- Is data stored temporarily, does it have to be transfered?
- Is an imaging data management software available?
- Necessity of large file transfer & storage?
- How is data sectuity (avoiding data loss, data privacy) managed?
- Is cloud computing & -storage required?

Processing & Bioimage Analysis

- Are shared digital environments required?
- Choice of analysis tools and software
- Use of (semi-)automated analysis workflows
- Use of AI-based analysis tools?
- Ensure analysis data provenance
- Linking original and derived data

(Data-) Publication

- Check open access requirements
- Check publisher policies
- Check funder policies
- Adhere to figure composing standards
- Consider dataset publication & preprints
- Link publication and data (persistent identifiers)

Archiving, Long-Term Storage

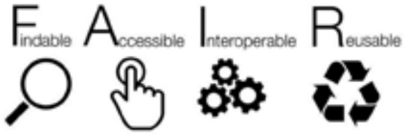
- Trusted public or institutional repository?
- Upload to added-value database?
- Choose data and metadata license
- Choose suitable data format
- Metadata enrichment before upload

from I3D:bio – Information Infrastructure for BioImage Data

Two most important points to care about when planning data management

1. Make sure you provide all possible **Metadata**

2. Be **F.A.I.R.**



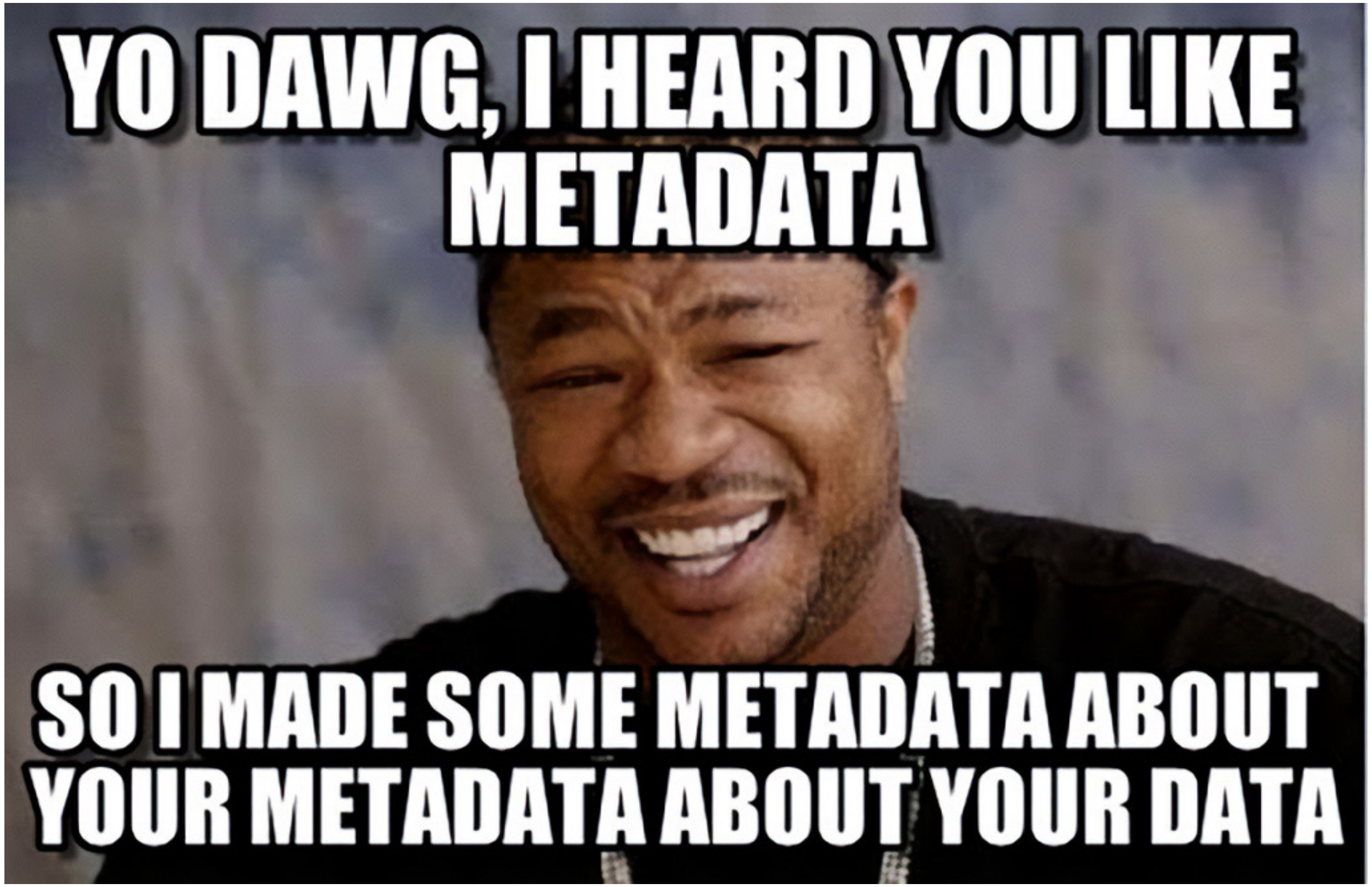
What is METADATA?

“We kill people based on metadata”

Gen. Michael Hayden
Former head of the National Security Agency

What is METADATA?

Data about your data…


YO DAWG, I HEARD YOU LIKE METADATA
SO I MADE SOME METADATA ABOUT YOUR METADATA ABOUT YOUR DATA

What is METADATA?

METADATA = a set of data that describes and gives information about other data.

- **scientific images** – image size, objective (mag., NA), filters, laser wavelength, exposure time, opt. slice thickness, stack size, xy resolution, etc.
- **photography** – image size, exposure, time & date, objective, zoom, flash/no flash, GPS data, etc. (EXIF information)
- **movies** – file format, movie length, size, actors, director, producer, etc.
- **ebooks** – format, author, editor, year of publication, # pages, ISBN, etc.
- **Music streaming** – artist, album, title, length of song, genre, etc.

What is METADATA? – a definition

“Metadata is constructed, constructive, and actionable.”

Karen Coyle, Digital Librarian, Author of Coyle’s InFormation

- Constructed - a man-made artifice, not naturally occurring
- Constructive - serving a useful purpose, to solve some problem
- Actionable - can be acted upon, processed by humans and machines

METADATA types

There are 3 main types of metadata:

- **Descriptive metadata** enables **discovery**, **identification**, and **selection** of resources. It can include elements such as title, author, and subjects.

- **Administrative metadata** facilities the **management of resources**.
  It can include elements such as technical, preservation, rights, and use.

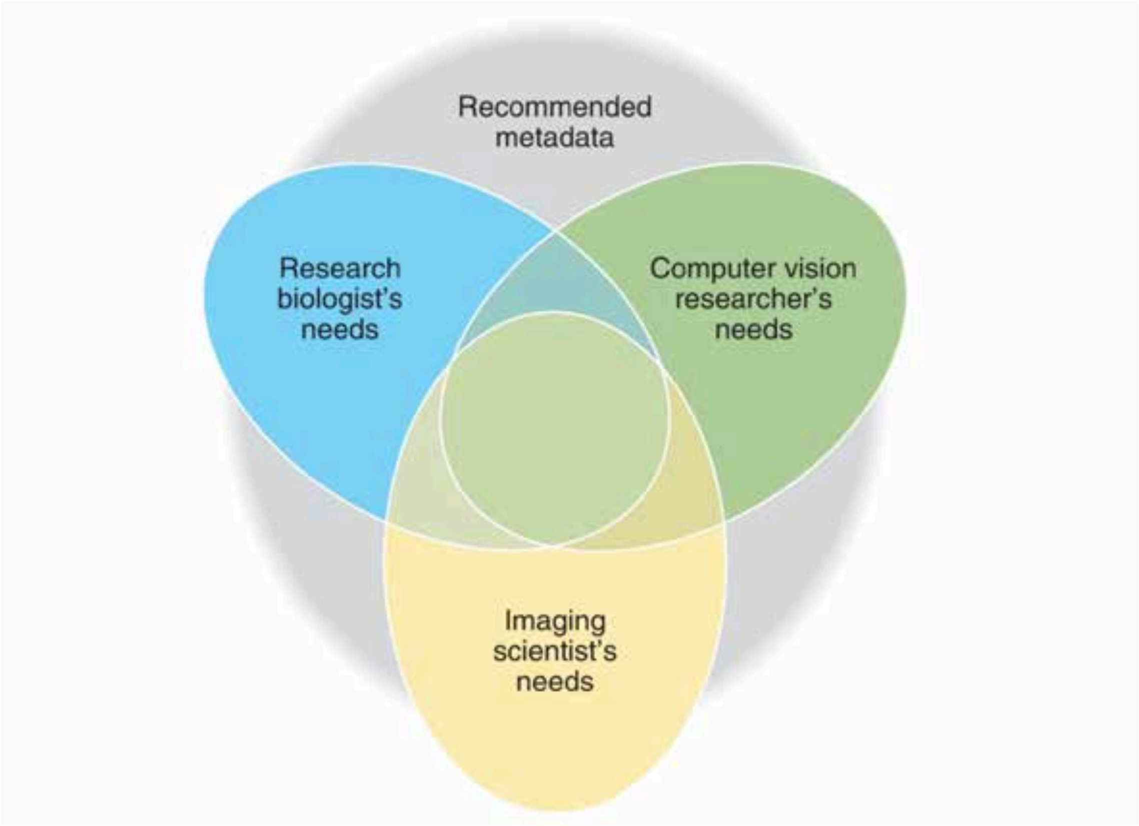
- **Structural metadata,** generally **used in machine processing**, describes **relationships** among various parts of a resource, such as chapters in a book.

METADATA in bioimaging (example)

- **Technical metadata** contains information on hardware and settings used to acquire an image.
  - includes **device specifications**, **objective** lens specifications, **light-source**, **laser** and **filter** settings, number of **channels**, **camera**, bit depth, etc.
  - is **automatically recorded** by most microscopes and **stored in the metadata header** of the image file.
  - can be accessed and edited using metadata editing tools.

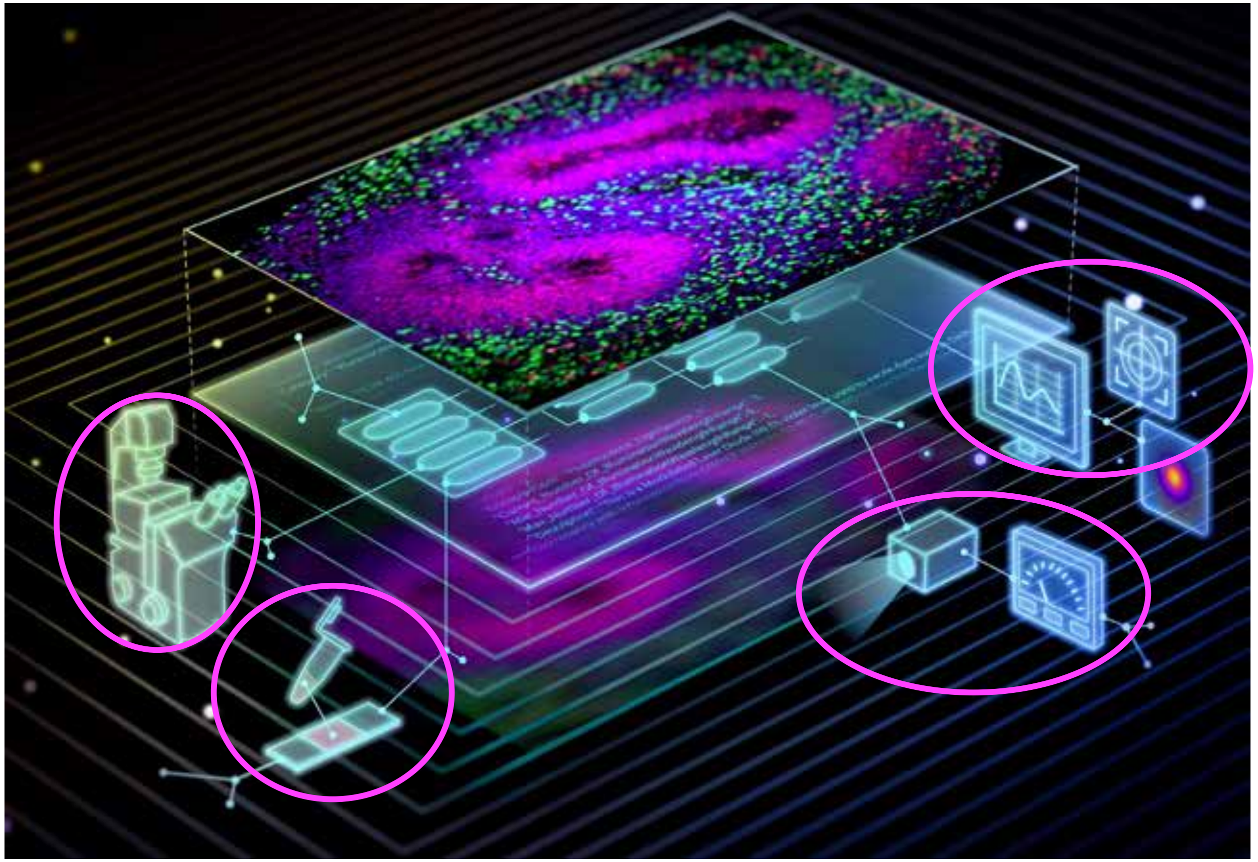
- **Experimental and sample preparation metadata** contains information about the specimen that is imaged (organism, cell line, organ, expression contstructs, etc.)
  - The **experimenter must add** all this information to the data.
  - This experimental metadata **should also cover**
    - **how a sample has been prepared** for imaging
      - fixed and stained with antibodies
      - conjugated to fluorescent dyes?
    - whether **live** cells and under what conditions were they imaged?

METADATA you need in bioimaging - visualization


Recommended metadata
Research biologist's needs
Computer vision researcher's needs
Imaging scientist's needs

from *REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology*

METADATA in bioimaging



© Thao Do (Allen Institute, Seattle, WA, USA)

Nature Methods FOCUS issue on Reporting and Reproducibility in Microscopy
https://www.nature.com/collections/djciihhjh

All information that is needed to interpret, evaluate the quality, reproduce and share microscopy images

- Sample preparation
- Image Acquisition
  - Hardware configuration
  - Acquisition setting
  - Quality Control
- Image data processing and analysis

How to create useful experimental METADATA

You want to describe your experimental data in a F.A.I.R. way?

What would you do?

You should use

- specific and defined **keywords or tags**
- **Controlled vocabularies** or **taxonomies**

Treasure map = METADATA and controlled vocabularies

How do you **find your treasure**?

- Randomly checking every island you find?
- Searching every corner of the island and do some digging at suspicious locations?

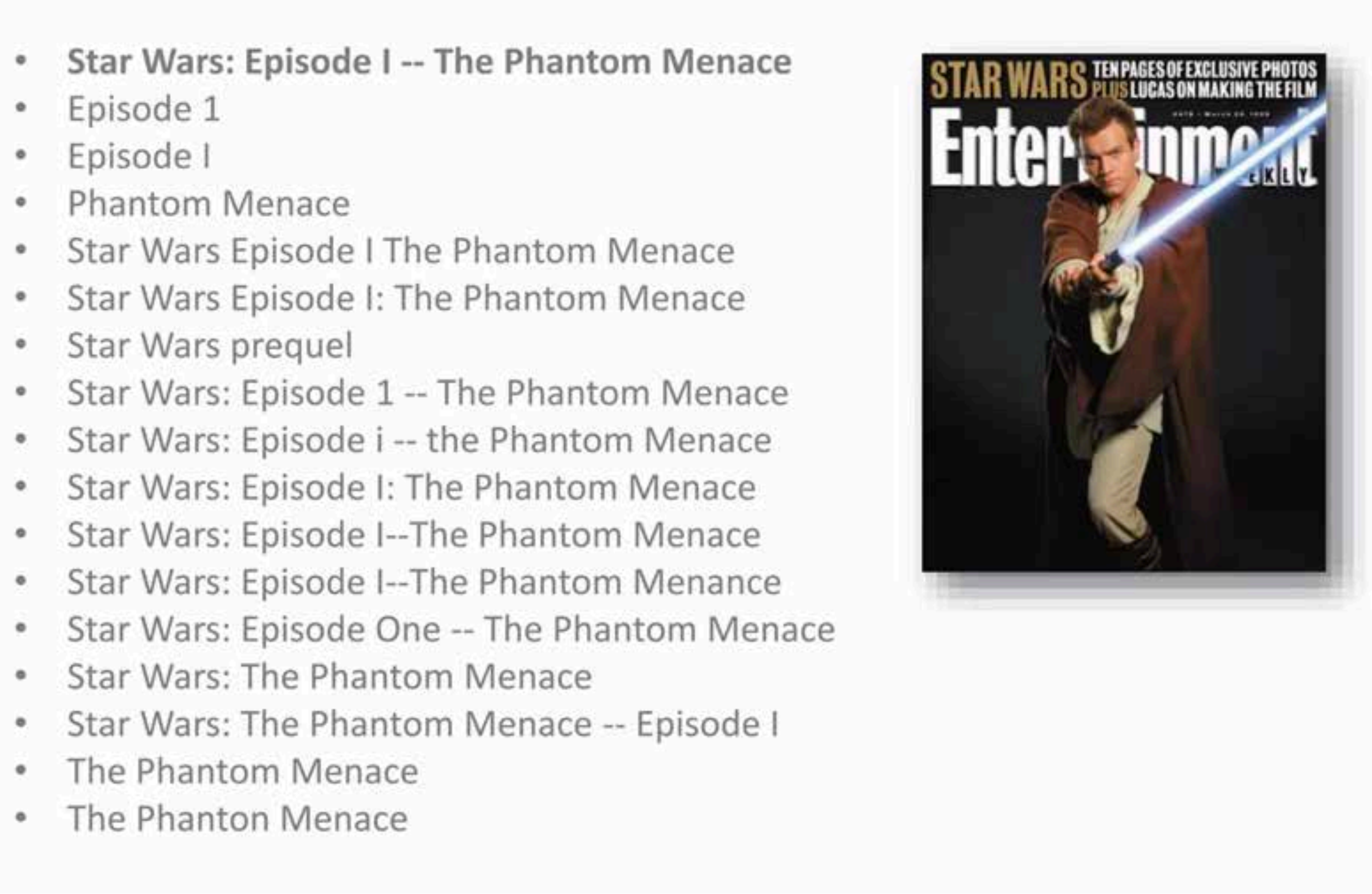
No!!!!

You would only start searching **if you had the treasure map**!

- Directly navigate to the island that fits the map
- Follow the path indicated on the map which is referring to **specific known or easily identifiable landmarks**
- **Search at the location** where the data treasure is hidden

Metadata is a map. Metadata is a means by which the complexity of an object is represented in a simpler form.

Standardizing keywords – an example



- **Star Wars: Episode I -- The Phantom Menace**
- Episode 1
- Episode I
- Phantom Menace
- Star Wars Episode I The Phantom Menace
- Star Wars Episode I: The Phantom Menace
- Star Wars prequel
- Star Wars: Episode 1 -- The Phantom Menace
- Star Wars: Episode i -- the Phantom Menace
- Star Wars: Episode I: The Phantom Menace
- Star Wars: Episode I--The Phantom Menace
- Star Wars: Episode I--The Phantom Menance
- Star Wars: Episode One -- The Phantom Menace
- Star Wars: The Phantom Menace
- Star Wars: The Phantom Menace -- Episode I
- The Phantom Menace
- The Phanton Menace

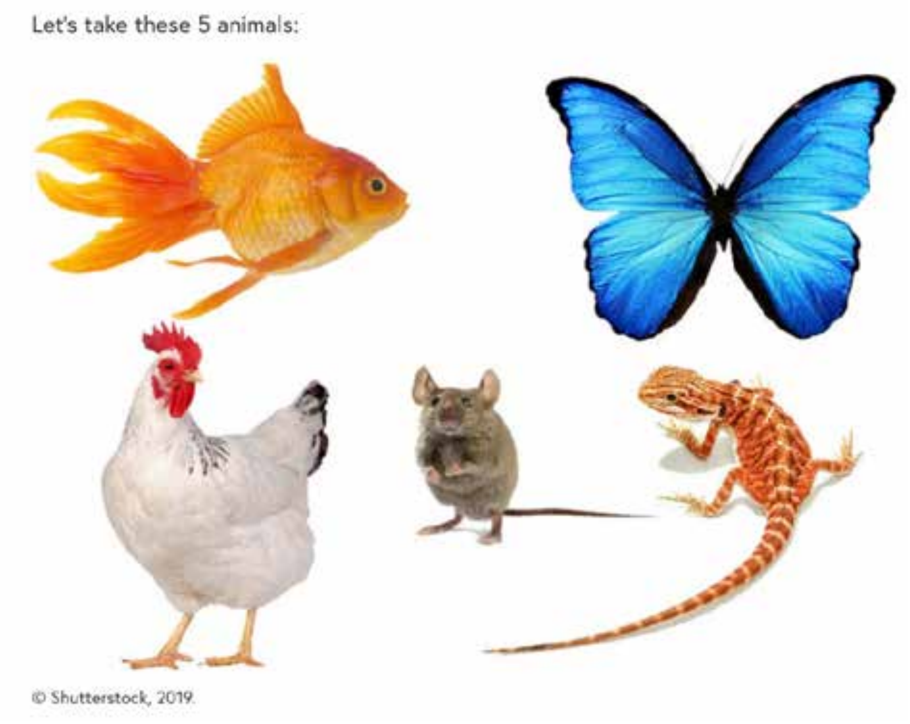
from https://www.slideshare.net/rlovinger/metadata-is-a-love-note-to-the-future

Controlled vocabularies example: Biological Taxonomy

Carolus Linnaeus, first person to combine binomial definition (nomenclature) with a hierarchical structure of classification.

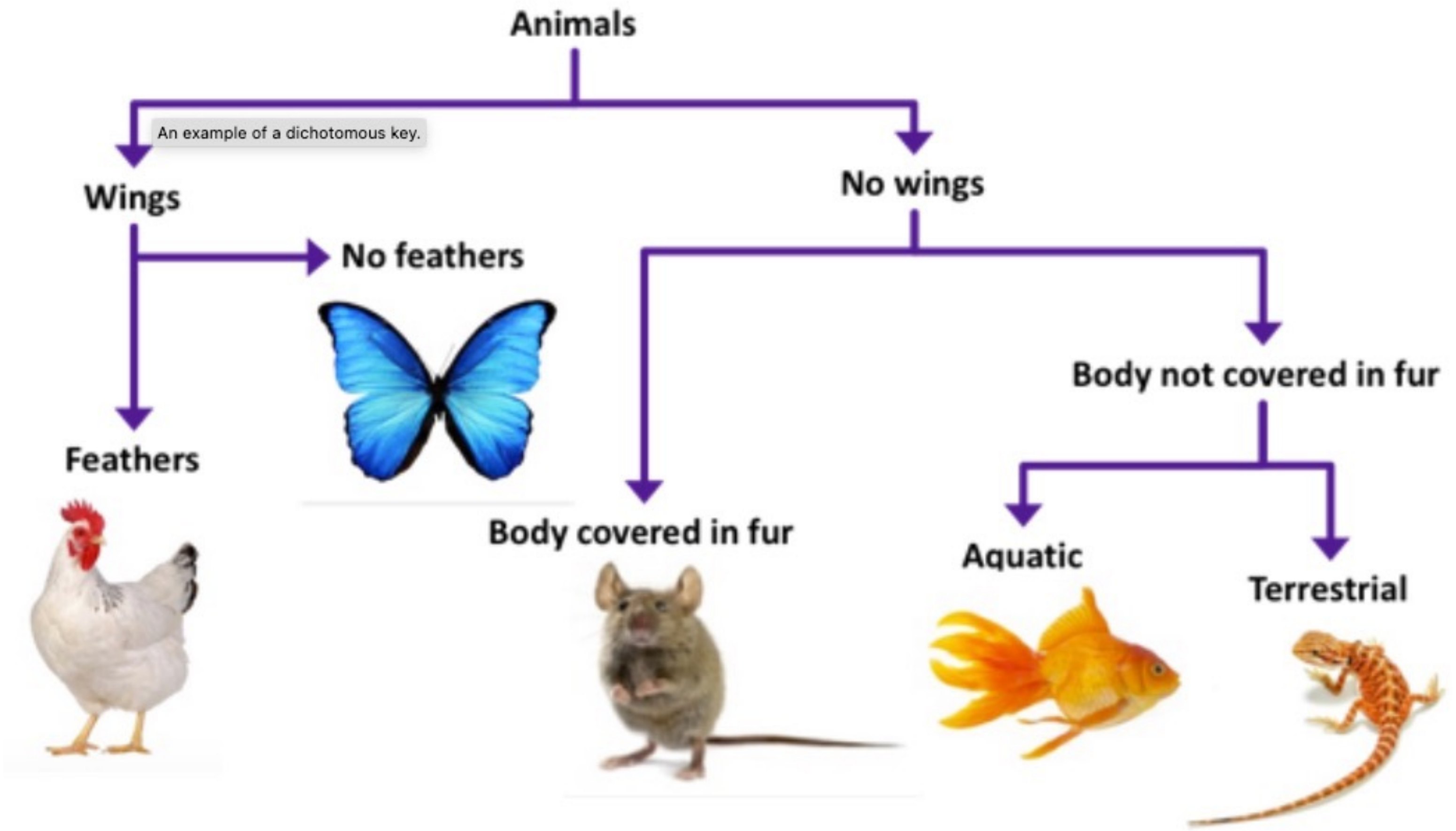
His **system organized both plants and animals from the level of the kingdom, right down to species**.

He used this system consistently to identify every species of plant and animal he came across and this is the basis of the system we use today.



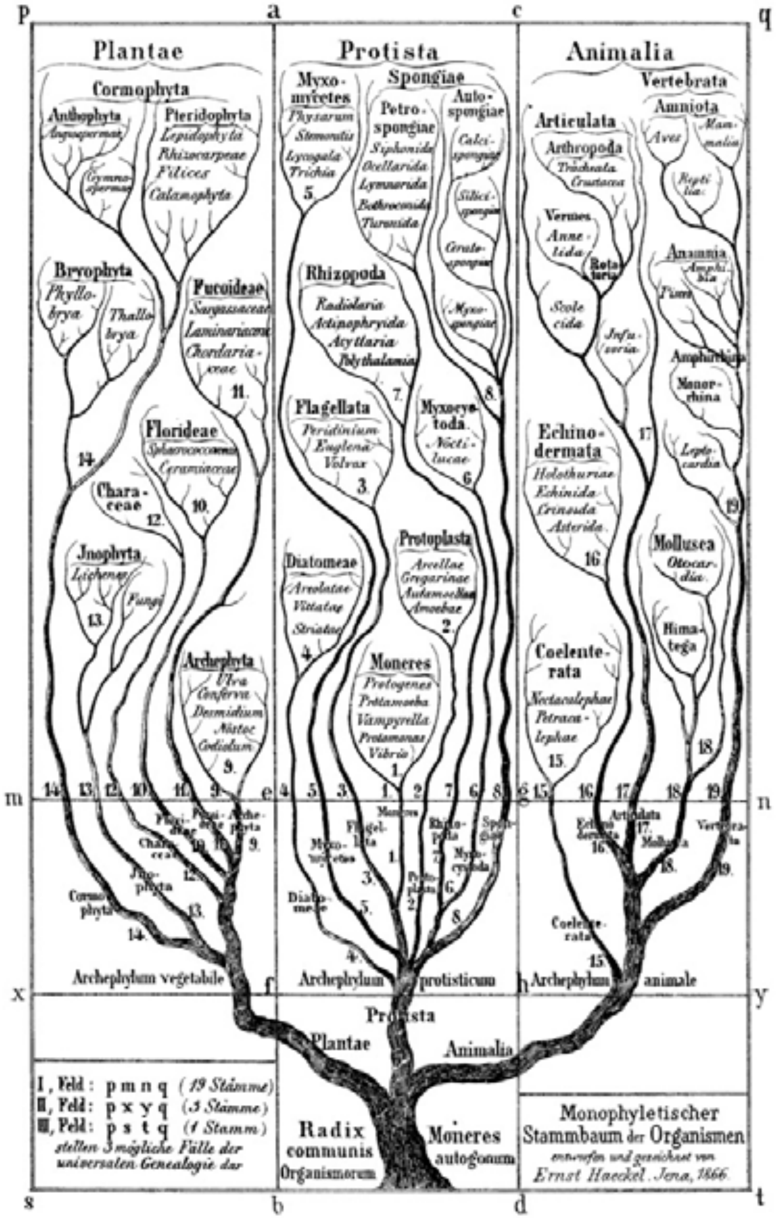
© Shutterstock, 2019.

Controlled vocabularies example: Biological Taxonomy

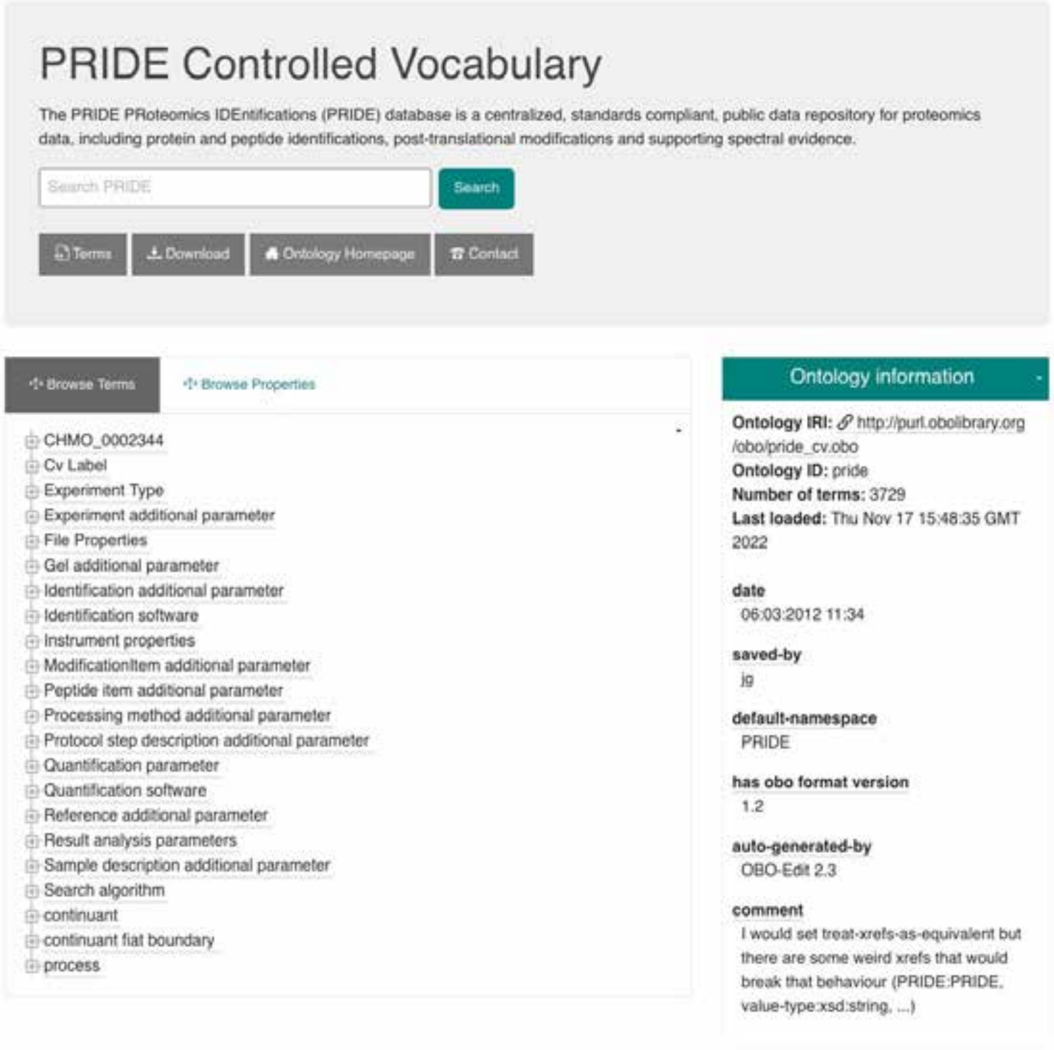


©Shutterstock, 2019

Controlled vocabularies example: Biological Taxonomy



Controlled vocabularies example in life sciences today - PRIDE



PRIDE Controlled Vocabulary

The PRIDE PRoteomics IDEntifications (PRIDE) database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.

- CHMO_0002344
- Cv Label
- Experiment Type
- Experiment additional parameter
- File Properties
- Gel additional parameter
- Identification additional parameter
- Identification software
- Instrument properties
- ModificationItem additional parameter
- Peptide item additional parameter
- Processing method additional parameter
- Protocol step description additional parameter
- Quantification parameter
- Quantification software
- Reference additional parameter
- Result analysis parameters
- Sample description additional parameter
- Search algorithm
- continuant
- continuant fiat boundary
- process

Ontology information

Ontology IRI: http://purl.obolibrary.org/obo/pride_cv.obo
Ontology ID: pride
Number of terms: 3729
Last loaded: Thu Nov 17 15:48:35 GMT 2022

date
06:03:2012 11:34

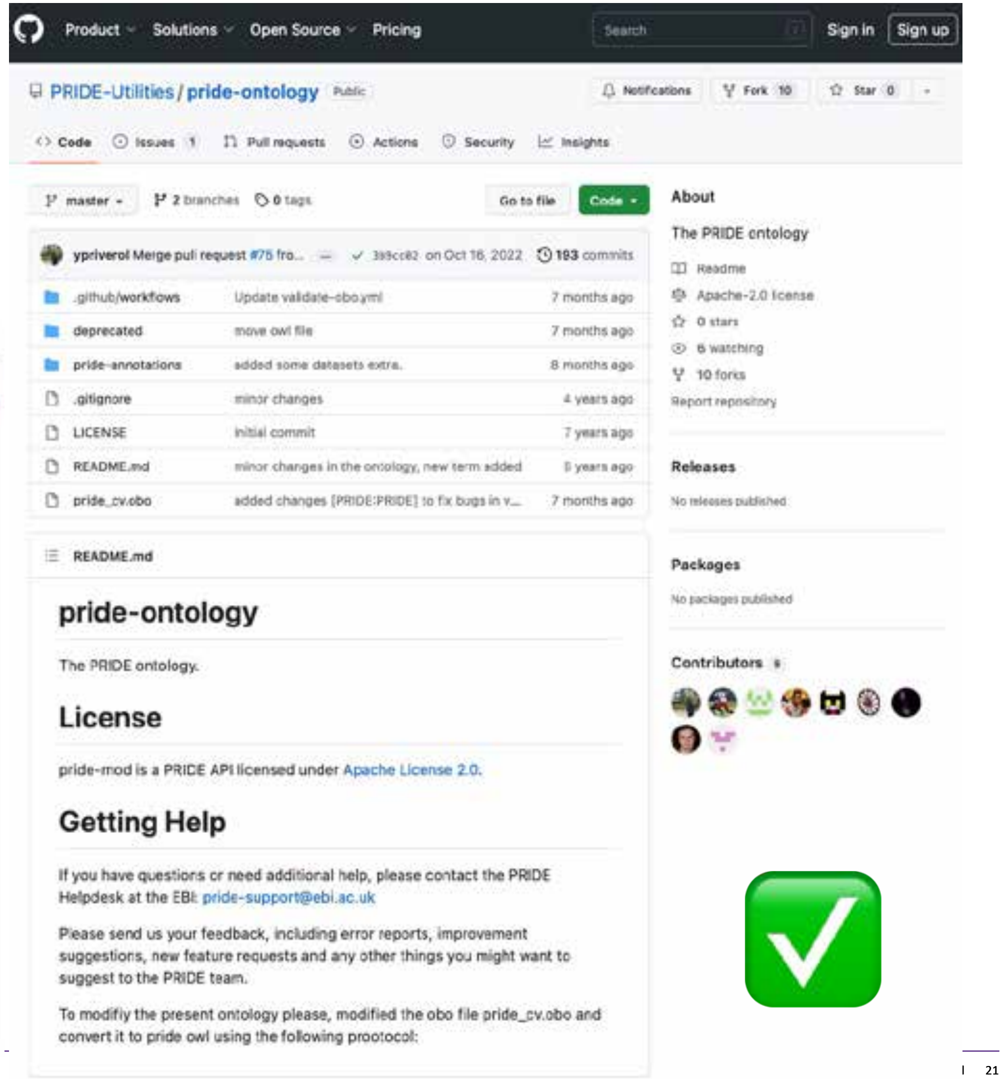
saved-by
jg

default-namespace
PRIDE

has obo format version
1.2

auto-generated-by
OBO-Edit 2.3

comment
I would set treat-xrefs-as-equivalent but there are some weird xrefs that would break that behaviour (PRIDE:PRIDE, value-type:xsd:string, ...)



PRIDE-Utilities / pride-ontology Public

ypriverol Merge pull request #75 fro... ✓ 399cc82 on Oct 16, 2022 193 commits

.github/workflows	Update validate-obo.yml	7 months ago
deprecated	move owl file	7 months ago
pride-annotations	added some datasets extra.	8 months ago
.gitignore	minor changes	4 years ago
LICENSE	initial commit	7 years ago
README.md	minor changes in the ontology, new term added	5 years ago
pride_cv.obo	added changes [PRIDE:PRIDE] to fix bugs in v...	7 months ago

README.md

pride-ontology

The PRIDE ontology.

License

pride-mod is a PRIDE API licensed under Apache License 2.0.

Getting Help

If you have questions or need additional help, please contact the PRIDE Helpdesk at the EBI: pride-support@ebi.ac.uk

Please send us your feedback, including error reports, improvement suggestions, new feature requests and any other things you might want to suggest to the PRIDE team.

To modifiy the present ontology please, modified the obo file pride_cv.obo and convert it to pride owl using the following prootocol:

About

The PRIDE ontology

Readme
Apache-2.0 license
0 stars
6 watching
10 forks
Report repository

Releases

No releases published

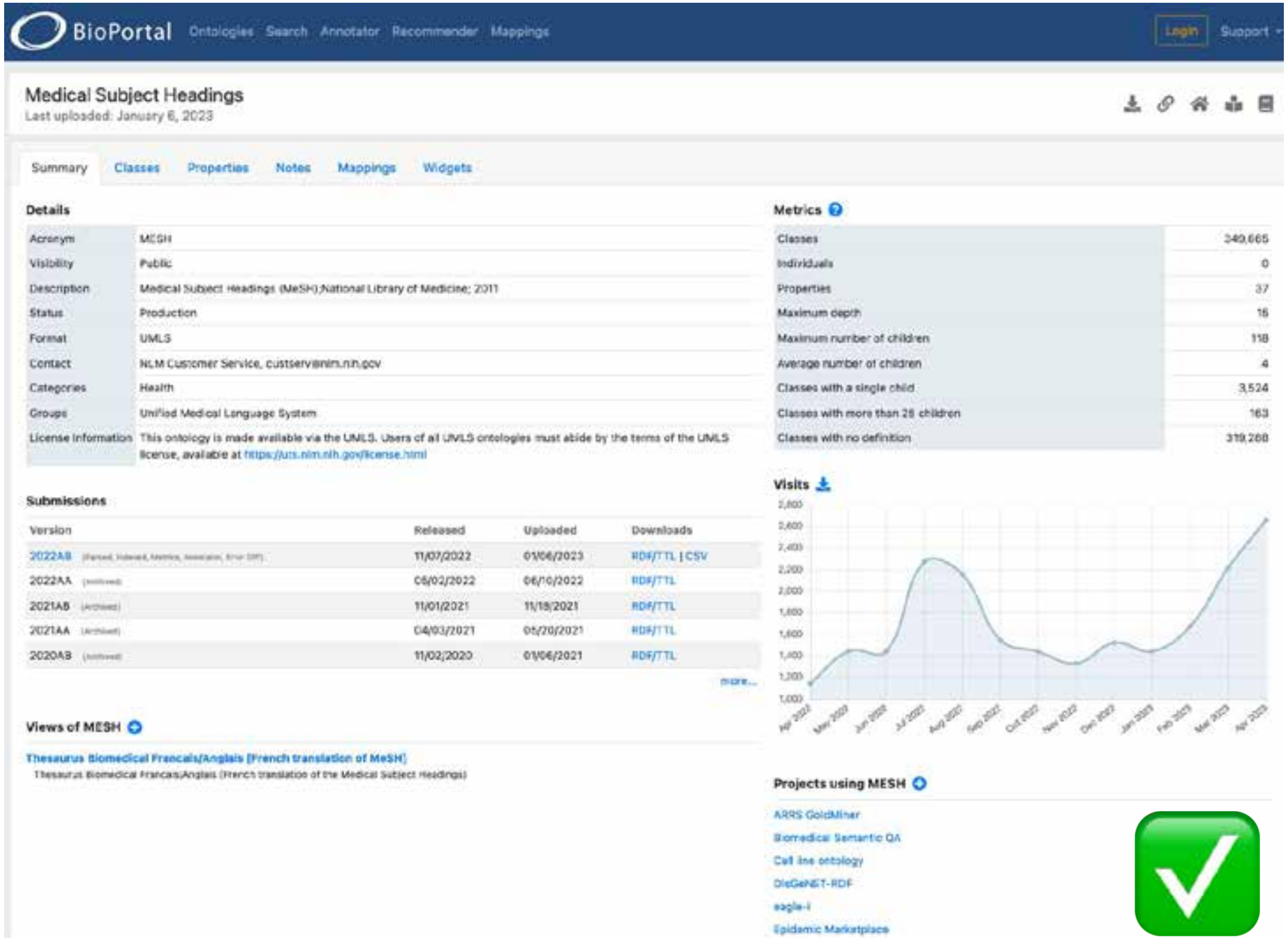
Packages

No packages published

Contributors 9

✅

Medical Subject Headings - MESH



Medical Subject Headings

Last uploaded: January 6, 2023

Details

Acronym	MESH
Visibility	Public
Description	Medical Subject Headings (MeSH);National Library of Medicine; 2011
Status	Production
Format	UMLS
Contact	NLM Customer Service, custserv@nlm.nih.gov
Categories	Health
Groups	Unified Medical Language System
License information	This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html

Metrics

Classes	349,665
Individuals	0
Properties	37
Maximum depth	16
Maximum number of children	118
Average number of children	4
Classes with a single child	3,524
Classes with more than 25 children	163
Classes with no definition	319,268

Submissions

Version	Released	Uploaded	Downloads
2022AB (Parsed, Indexed, Metrics, Annotator, Error Diff)	11/07/2022	01/06/2023	RDF/TTL | CSV
2022AA (Archived)	05/02/2022	06/10/2022	RDF/TTL
2021AB (Archived)	11/01/2021	11/18/2021	RDF/TTL
2021AA (Archived)	04/03/2021	05/20/2021	RDF/TTL
2020AB (Archived)	11/02/2020	01/06/2021	RDF/TTL

more...

Visits

Views of MESH

Thesaurus Biomedical Francais/Anglais [French translation of MeSH]

Thesaurus Biomedical Francais/Anglais (French translation of the Medical Subject Headings)

Projects using MESH

ARRS GoldMiner

Biomedical Semantic QA

Cell line ontology

DisGeNET-RDF

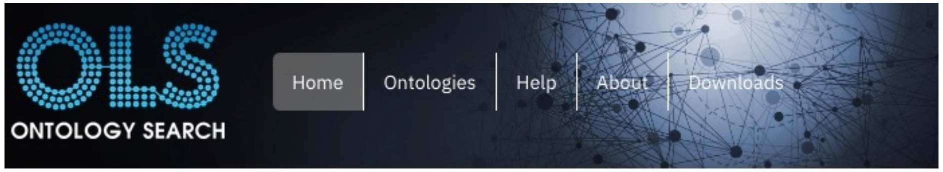
eagle-i

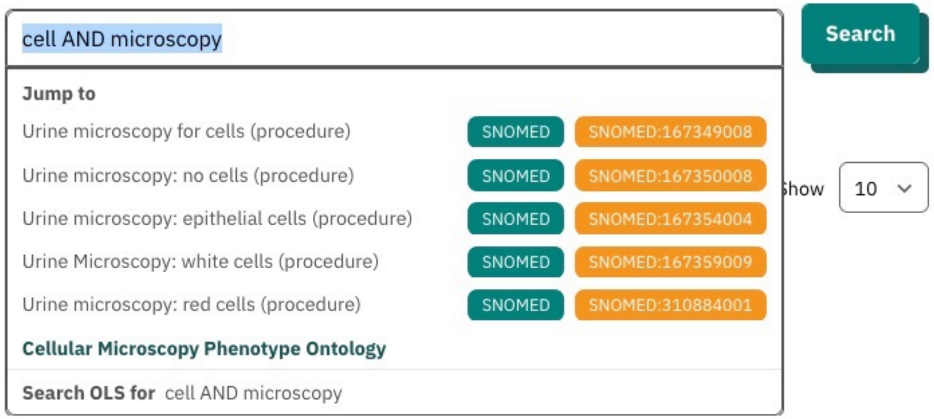
Epidemic Marketplace

✓

What if you do not know “your” specific ontology yet?

Search OLS (Ontology Lookup Service)





cell AND microscopy

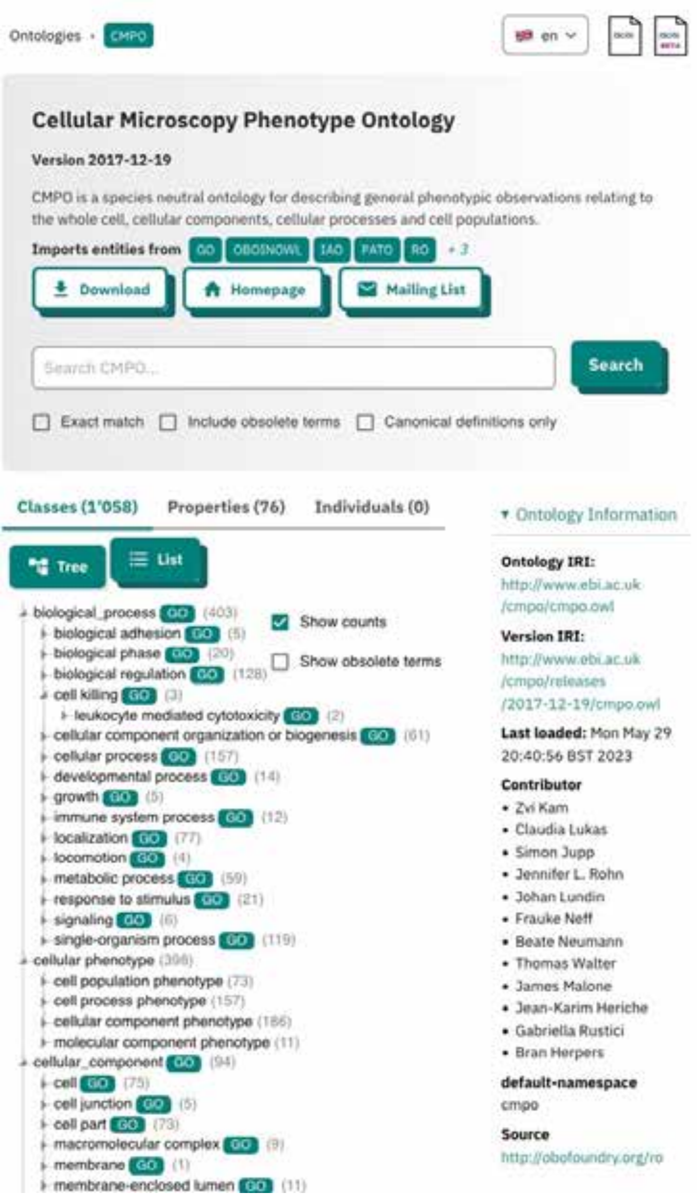
Search

Jump to

Urine microscopy for cells (procedure)	SNOMED	SNOMED:167349008
Urine microscopy: no cells (procedure)	SNOMED	SNOMED:167350008
Urine microscopy: epithelial cells (procedure)	SNOMED	SNOMED:167354004
Urine Microscopy: white cells (procedure)	SNOMED	SNOMED:167359009
Urine microscopy: red cells (procedure)	SNOMED	SNOMED:310884001

Cellular Microscopy Phenotype Ontology

Search OLS for cell AND microscopy



Ontologies • CMPO

en

Cellular Microscopy Phenotype Ontology

Version 2017-12-19

CMPO is a species neutral ontology for describing general phenotypic observations relating to the whole cell, cellular components, cellular processes and cell populations.

Imports entities from GO OBOINOWL IAO PATO RO + 3

Download Homepage Mailing List

Search CMPO... Search

☐ Exact match ☐ Include obsolete terms ☐ Canonical definitions only

Classes (1'058) **Properties (76)** **Individuals (0)**

Tree List

☑ Show counts

☐ Show obsolete terms

- biological_process GO (403)
  - biological adhesion GO (5)
  - biological phase GO (20)
  - biological regulation GO (128)
  - cell killing GO (3)
    - leukocyte mediated cytotoxicity GO (2)
  - cellular component organization or biogenesis GO (61)
  - cellular process GO (157)
  - developmental process GO (14)
  - growth GO (5)
  - immune system process GO (12)
  - localization GO (77)
  - locomotion GO (4)
  - metabolic process GO (59)
  - response to stimulus GO (21)
  - signaling GO (6)
  - single-organism process GO (119)
- cellular phenotype (306)
  - cell population phenotype (73)
  - cell process phenotype (157)
  - cellular component phenotype (186)
  - molecular component phenotype (11)
- cellular_component GO (94)
  - cell GO (75)
  - cell junction GO (5)
  - cell part GO (73)
  - macromolecular complex GO (9)
  - membrane GO (1)
  - membrane-enclosed lumen GO (11)

▾ Ontology Information

Ontology IRI:
http://www.ebi.ac.uk/cmpo/cmpo.owl

Version IRI:
http://www.ebi.ac.uk/cmpo/releases/2017-12-19/cmpo.owl

Last loaded: Mon May 29 20:40:56 BST 2023

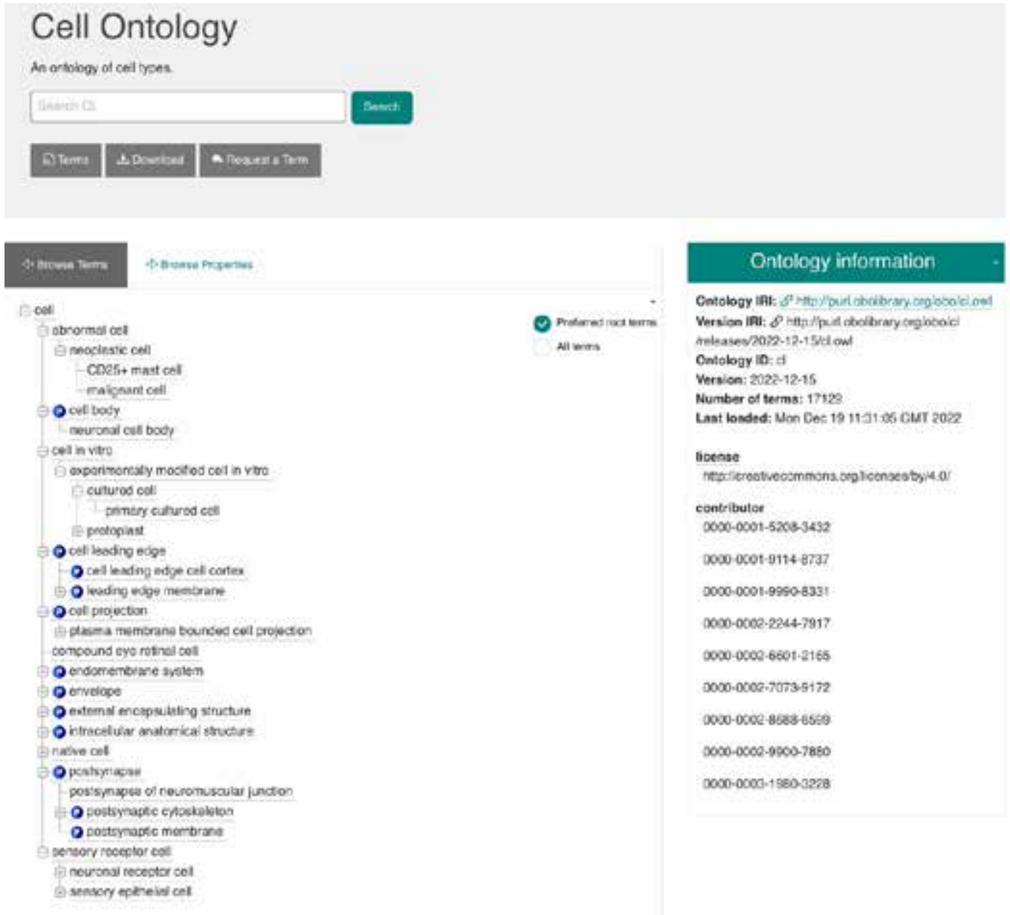
Contributor

- Zvi Kam
- Claudia Lukas
- Simon Jupp
- Jennifer L. Rohn
- Johan Lundin
- Frauke Neff
- Beate Neumann
- Thomas Walter
- James Malone
- Jean-Karim Heriche
- Gabriella Rustici
- Bran Herpers

default-namespace
cmpo

Source
http://obofoundry.org/ro

Ontologies that might be useful for NCCR AntiResist



Cell Ontology

An ontology of cell types.

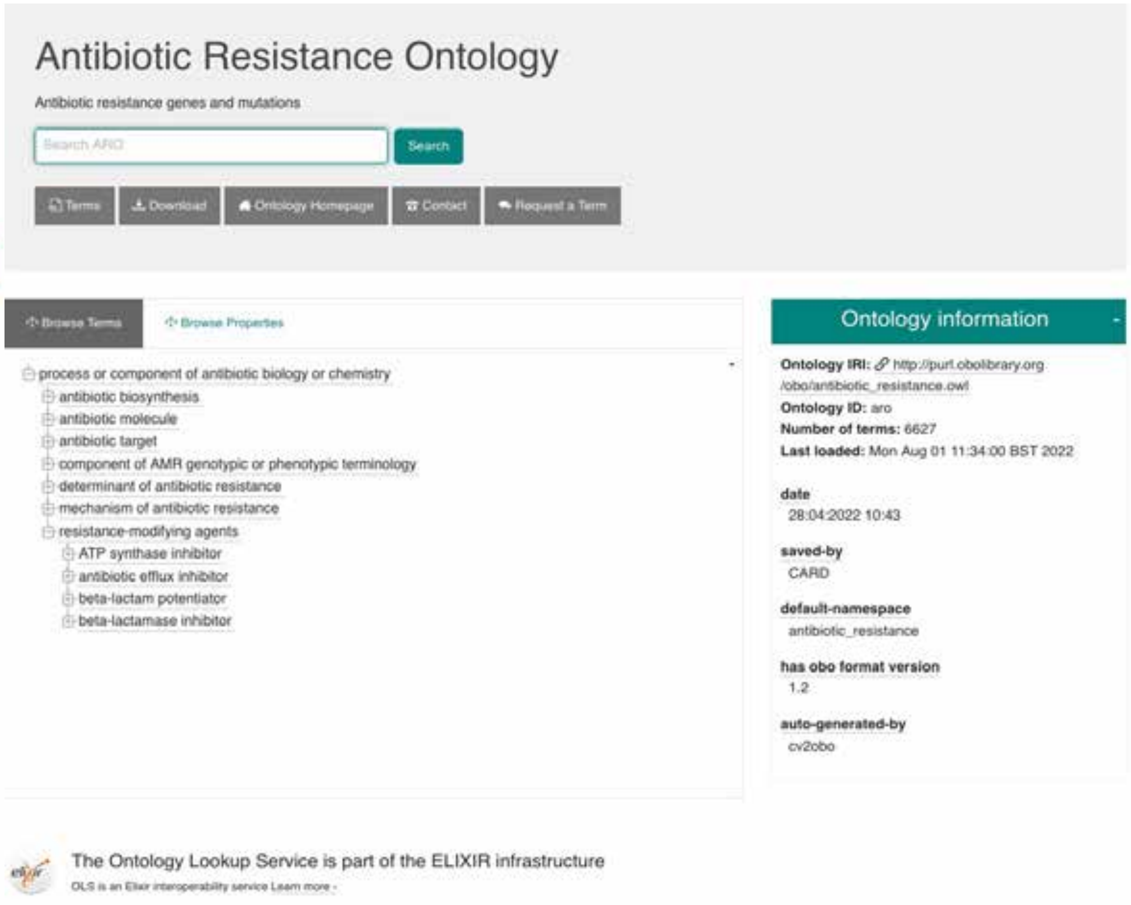
- cell
  - abnormal cell
    - neoplastic cell
      - CD25+ mast cell
      - malignant cell
  - cell body
    - neuronal cell body
  - cell in vitro
    - experimentally modified cell in vitro
      - cultured cell
        - primary cultured cell
      - protoplast
  - cell leading edge
    - cell leading edge cell cortex
    - leading edge membrane
  - cell projection
    - plasma membrane bounded cell projection
  - compound eye retinal cell
  - endomembrane system
  - envelope
  - external encapsulating structure
  - intracellular anatomical structure
  - native cell
  - postsynapse
    - postsynapse of neuromuscular junction
    - postsynaptic cytoskeleton
    - postsynaptic membrane
  - sensory receptor cell
    - neuronal receptor cell
    - sensory epithelial cell

Ontology information

Ontology IRI: http://purl.obolibrary.org/obo/cl.owl
Version IRI: http://purl.obolibrary.org/obo/cl/releases/2022-12-15/cl.owl
Ontology ID: cl
Version: 2022-12-15
Number of terms: 17129
Last loaded: Mon Dec 19 11:31:05 GMT 2022

license
http://creativecommons.org/licenses/by/4.0/

contributor
0000-0001-5208-3432
0000-0001-9114-8737
0000-0001-9990-8331
0000-0002-2244-7917
0000-0002-6601-2165
0000-0002-7073-9172
0000-0002-8688-6599
0000-0002-9900-7880
0000-0003-1980-3228



Antibiotic Resistance Ontology

Antibiotic resistance genes and mutations

- process or component of antibiotic biology or chemistry
  - antibiotic biosynthesis
  - antibiotic molecule
  - antibiotic target
  - component of AMR genotypic or phenotypic terminology
  - determinant of antibiotic resistance
  - mechanism of antibiotic resistance
  - resistance-modifying agents
    - ATP synthase inhibitor
    - antibiotic efflux inhibitor
    - beta-lactam potentiator
    - beta-lactamase inhibitor

Ontology information

Ontology IRI: http://purl.obolibrary.org/obo/antibiotic_resistance.owl
Ontology ID: aro
Number of terms: 6627
Last loaded: Mon Aug 01 11:34:00 BST 2022

date
28:04:2022 10:43

saved-by
CARD

default-namespace
antibiotic_resistance

has obo format version
1.2

auto-generated-by
cv2obo

The Ontology Lookup Service is part of the ELIXIR infrastructure

OLS is an Elixir interoperability service Learn more ›

What if you do not find a suitable ontology?

Create your own ontology

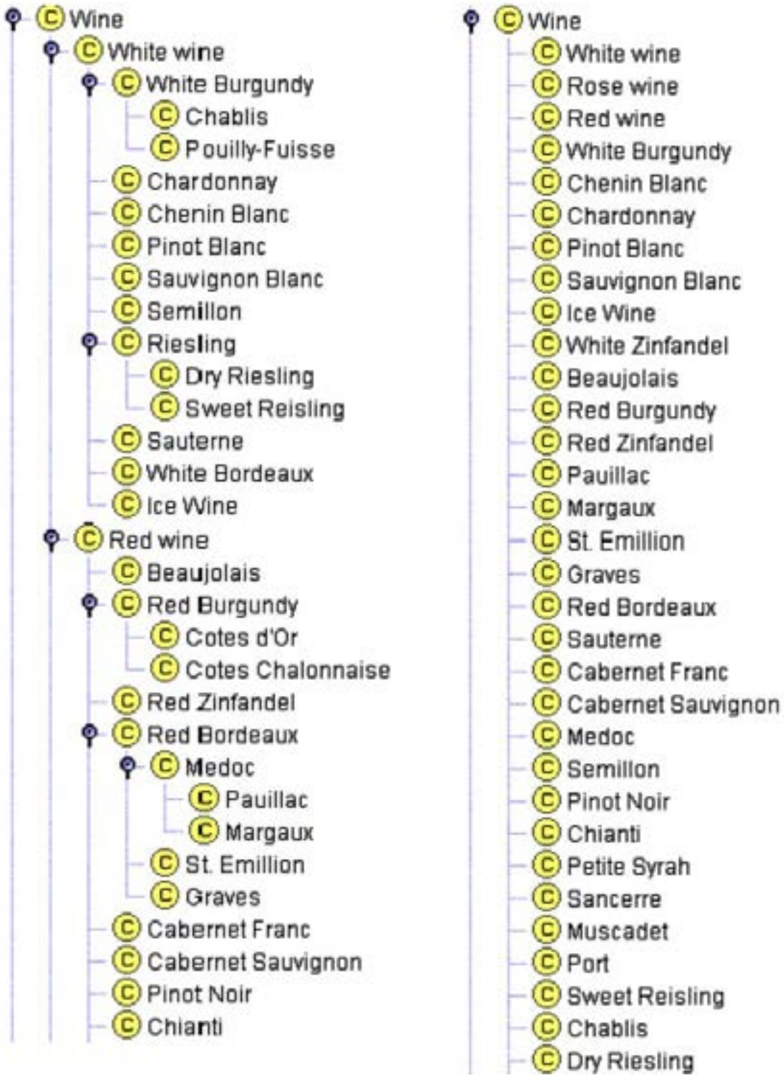
1. Determine the **domain and scope** of the ontology.

2. **Consider reusing** existing ontologies. At least to start with

3. Enumerate **important terms**.

4. Define the **classes & class hierarchy**.

5. Define the **properties of classes**.



from http://www.ksl.stanford.edu/people/dlm/papers/ontology101/

If you are planning to create your own ontology

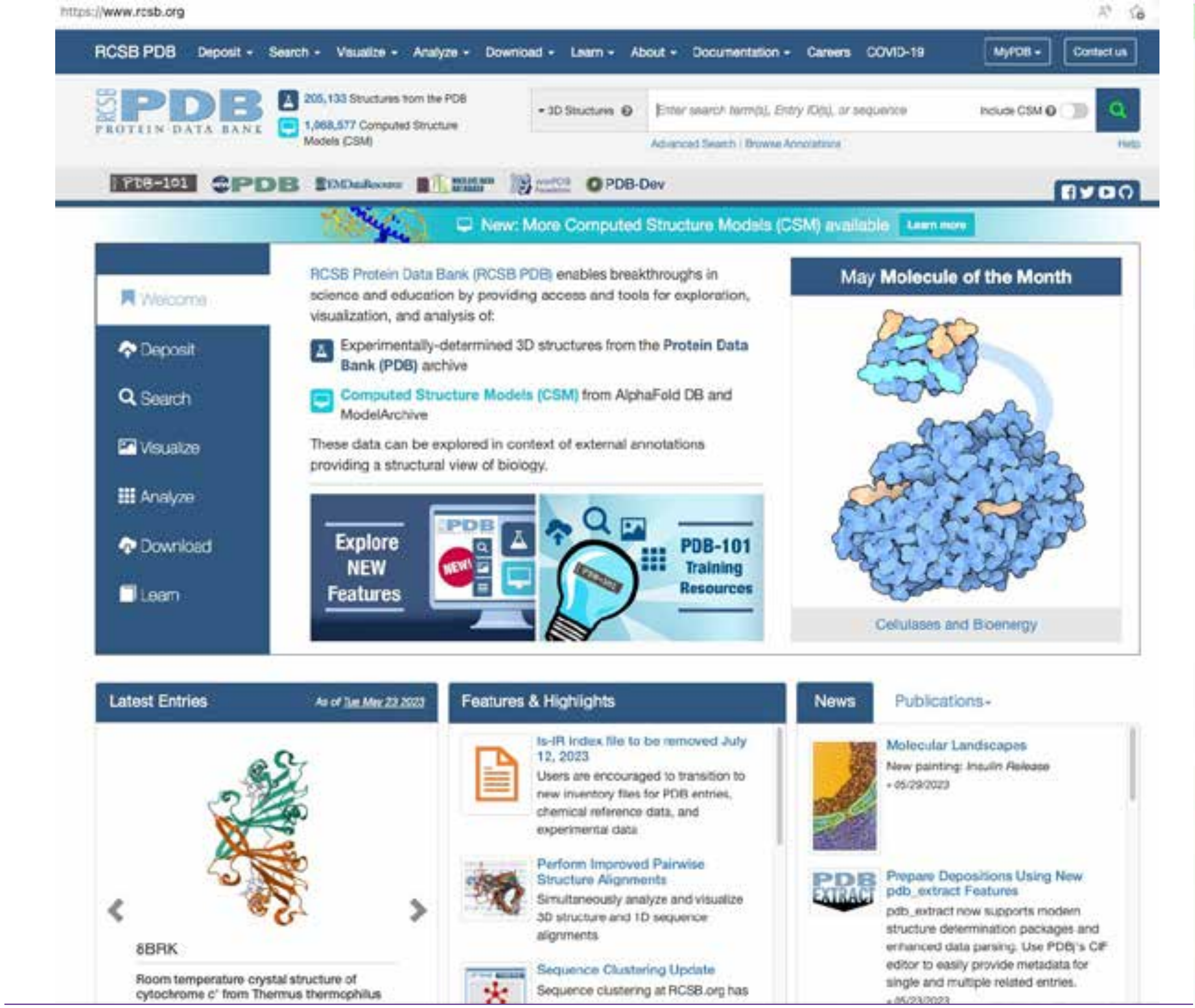
- Try to **avoid individual group level solutions** at all costs, as they are almost never sustainable and F.A.I.R.

- Try to **reach out to the largest possible community** and get them all involved (if possible).

  This usually makes the process longer and more tedious, but in the end it **results in an ontology that is used by everyone** in the field (see PRIDE, PDB, etc).

“We just do the right thing to the best of our ability and that happens to be F.A.I.R.”

On of the success stories for a database, metadata, ontology



205,133 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) available

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

May Molecule of the Month

Cellulases and Bioenergy

Latest Entries As of Tue May 23 2023

8BRK

Room temperature crystal structure of cytochrome c' from Thermus thermophilus

Features & Highlights

Is-IR index file to be removed July 12, 2023
Users are encouraged to transition to new inventory files for PDB entries, chemical reference data, and experimental data

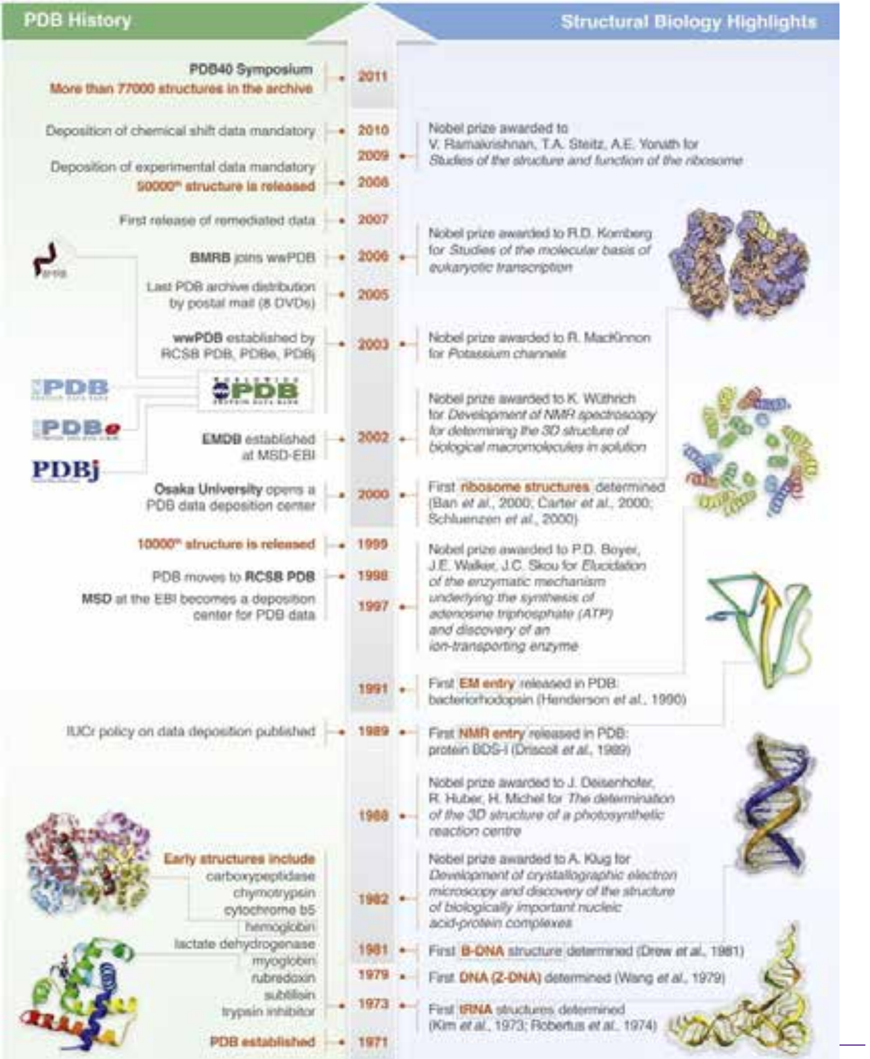
Perform Improved Pairwise Structure Alignments
Simultaneously analyze and visualize 3D structure and 1D sequence alignments

Sequence Clustering Update
Sequence clustering at RCSB.org has

News Publications

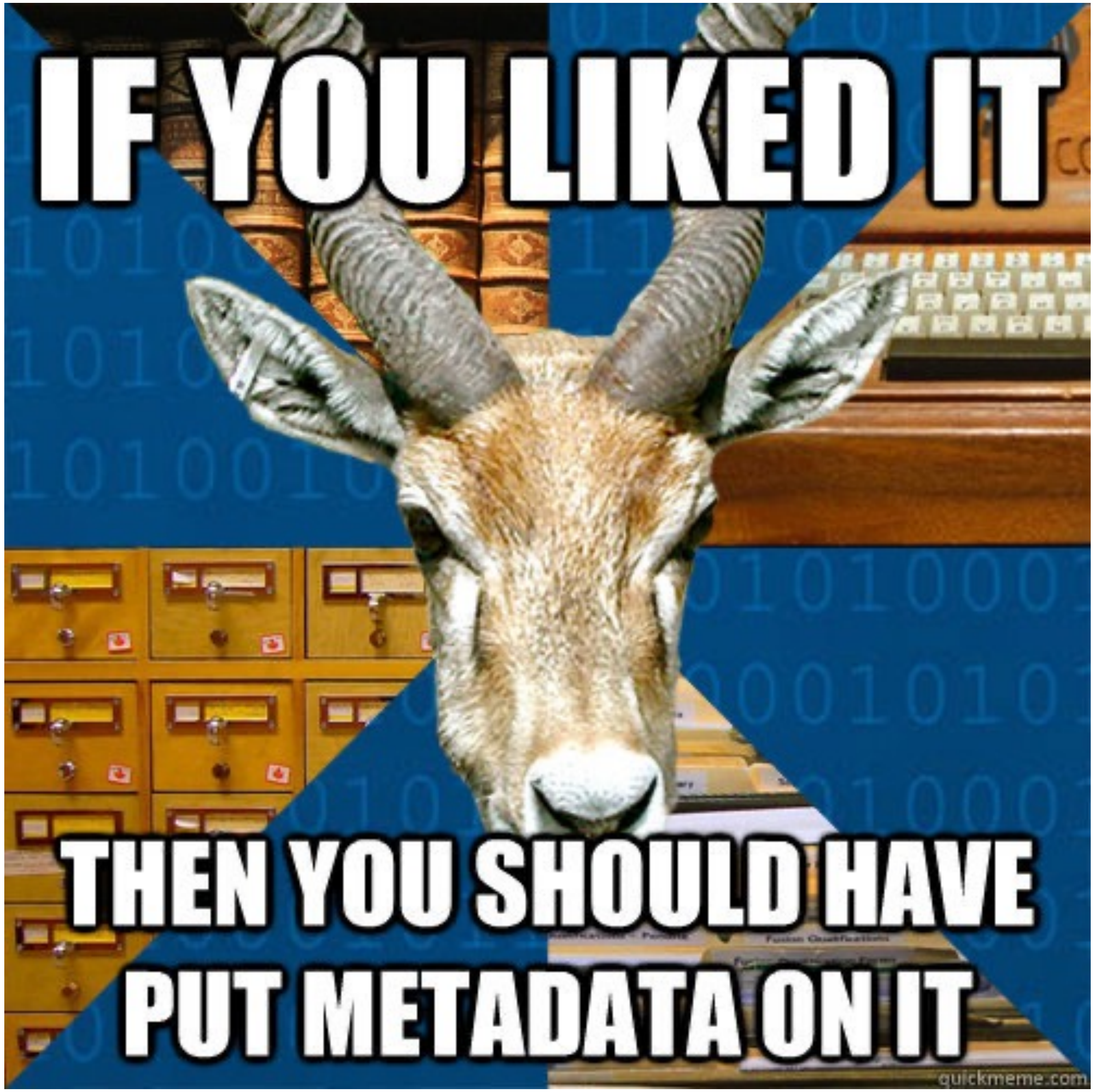
Molecular Landscapes
New painting: *Insulin Release*
• 05/29/2023

Prepare Depositions Using New pdb_extract Features
pdb_extract now supports modern structure determination packages and enhanced data parsing. Use PDBj's CIF editor to easily provide metadata for single and multiple related entries.
• 05/23/2023



PDB History		Structural Biology Highlights
PDB40 Symposium More than 77000 structures in the archive	2011	
Deposition of chemical shift data mandatory	2010	
	2009	Nobel prize awarded to V. Ramakrishnan, T.A. Steitz, A.E. Yonath for *Studies of the structure and function of the ribosome*
Deposition of experimental data mandatory 50000th structure is released	2008	
First release of remediated data	2007	
BMRB joins wwPDB	2006	Nobel prize awarded to R.D. Kornberg for *Studies of the molecular basis of eukaryotic transcription*
Last PDB archive distribution by postal mail (8 DVDs)	2005	
wwPDB established by RCSB PDB, PDBe, PDBj	2003	Nobel prize awarded to R. MacKinnon for *Potassium channels*
EMDB established at MSD-EBI	2002	Nobel prize awarded to K. Wüthrich for *Development of NMR spectroscopy for determining the 3D structure of biological macromolecules in solution*
Osaka University opens a PDB data deposition center	2000	First ribosome structures determined (Ban *et al.*, 2000; Carter *et al.*, 2000; Schluenzen *et al.*, 2000)
10000th structure is released	1999	
PDB moves to RCSB PDB	1998	
MSD at the EBI becomes a deposition center for PDB data	1997	Nobel prize awarded to P.D. Boyer, J.E. Walker, J.C. Skou for *Elucidation of the enzymatic mechanism underlying the synthesis of adenosine triphosphate (ATP) and discovery of an ion-transporting enzyme*
	1991	First EM entry released in PDB: bacteriorhodopsin (Henderson *et al.*, 1990)
IUCr policy on data deposition published	1989	First NMR entry released in PDB: protein BDS-I (Driscoll *et al.*, 1989)
	1988	Nobel prize awarded to J. Deisenhofer, R. Huber, H. Michel for *The determination of the 3D structure of a photosynthetic reaction centre*
	1982	Nobel prize awarded to A. Klug for *Development of crystallographic electron microscopy and discovery of the structure of biologically important nucleic acid-protein complexes*
	1981	First B-DNA structure determined (Drew *et al.*, 1981)
	1979	First DNA (Z-DNA) determined (Wang *et al.*, 1979)
Early structures include carboxypeptidase chymotrypsin cytochrome b5 hemoglobin lactate dehydrogenase myoglobin rubredoxin subtilisin trypsin inhibitor	1973	First tRNA structures determined (Kim *et al.*, 1973; Robertus *et al.*, 1974)
PDB established	1971	

from [Berman et al 2012]


IF YOU LIKED IT
THEN YOU SHOULD HAVE PUT METADATA ON IT
quickmeme.com

Workshop - LEGO group exercise
(5 groups à 5-6 persons)

- Group into 5 groups à 5-6 persons

LEGO group exercise (5 groups à 5-6 persons)

- **Build the structure assigned to the respective building station**
  - **Document your structure** using available tools
  - **Take a picture** of your build (or tutor will take a picture)

- **Disassemble** build

modified from LEGO® Metadata for Reproducibility game pack by University of Glasgow

LEGO group exercise (5 groups à 5-6 persons)

- **Build #1** - Build the structure assigned to the respective building station
  - **Document your structure** using available tools at the station
  - **Take a picture** of your build (or tutor will take a picture)
- **Disassemble** build

- **Build #2** – Rebuild the previous structure
  using the documentation of the previous group
- **Compare**
- **Discuss your experience**

modified from LEGO® Metadata for Reproducibility game pack by University of Glasgow

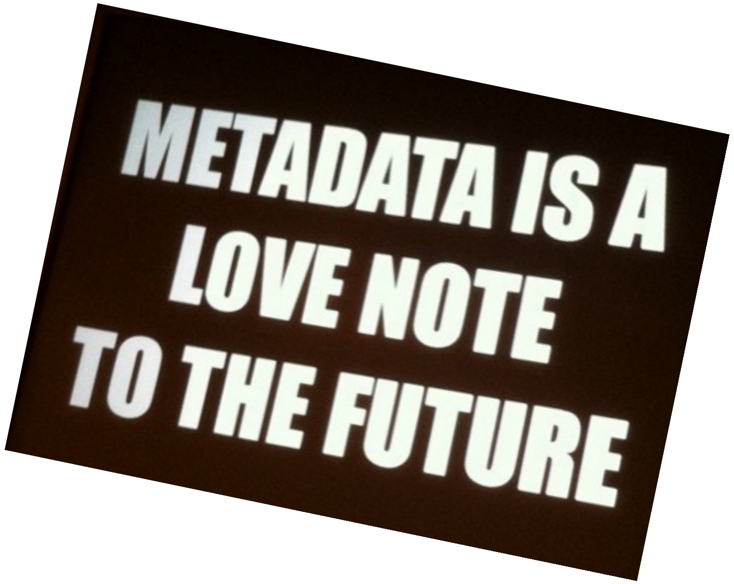
LEGO group exercise- discussion of results

- ➢Did you find this a simple way to document your process?
- ➢Was there anything you found difficult to capture?
- ➢Did those replicating the builds find it straightforward to follow?
- ➢Did you encounter any ambiguity in the instructions?

modified from LEGO® Metadata for Reproducibility game pack by University of Glasgow

We hope you liked the workshop

And always remember:



METADATA IS A LOVE NOTE TO THE FUTURE