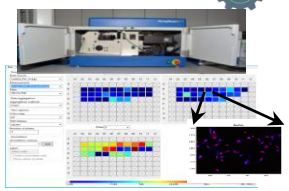
Décrire les données de microscopies en biologie: vers un consensus communautaire ou comment créer un standard de metadonnées

How to describe bioimage data? Toward a community consensus, or how to create a metadata standard?

Perrine Paul-Gilloteaux
UAR BioCore MicroPICell
FBI.data, Bretagne Loire Node of France BioImaging

France BioImaging DMP in a nutshell

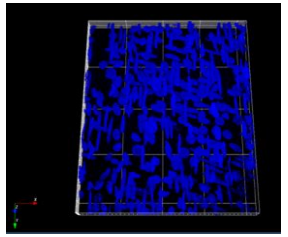


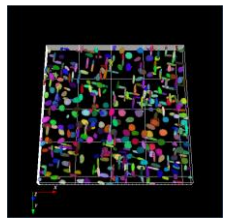
Metadata and history of processing (workflows)

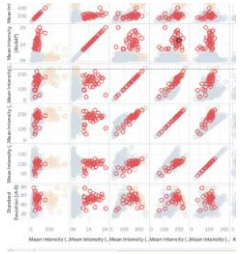
Metadata and history of processing (workflows)

Metadata and history of processing (workflows)

Labelled mesocenters

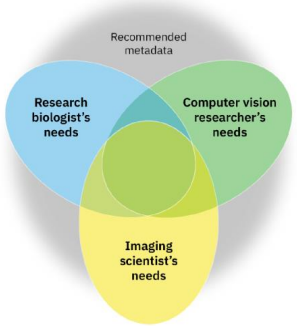






Publication

Metadata and history of processing (workflows)



Data repository with certifications

BioImage Archive

IDR

EMPIAR

FRANCE-BIOIMAGING

Why Metadata?
They are everywhere in FAIR principles…

Findable

- F1 Data are assigned a globally unique and persistent identifier
- F2 Data are described with rich metadata (R1)
- F3 Metadata clearly and explicitly include the identifier of the data they describe
- F4 Data are registered or indexed in a searchable resource

Accessible

- A1 Data are retrievable by their identifier using a standardised communications protocol
- A2 Metadata are accessible, even when the data are no longer available

Interoperable

- I1 Data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2 Data use vocabularies that follow FAIR principles
- I3 Data include qualified references to other (meta)data

Reusable

- R1 Data are richly described with a plurality of accurate and relevant attributes (license, provenance, relevant community standards)

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship.
Sci Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Why Metadata?
They are everywhere in FAIR principles…

Findable

- F1 (Meta)data are assigned a globally unique and persistent identifier
- F2 Data are described with rich metadata (R1)
- F3 Metadata clearly and explicitly include the identifier of the data they describe
- F4 (Meta)data are registered or indexed in a searchable resource

Accessible

- A1 (Meta)data are retrievable by their identifier using a standardised communications protocol
- A2 Metadata are accessible, even when the data are no longer available

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

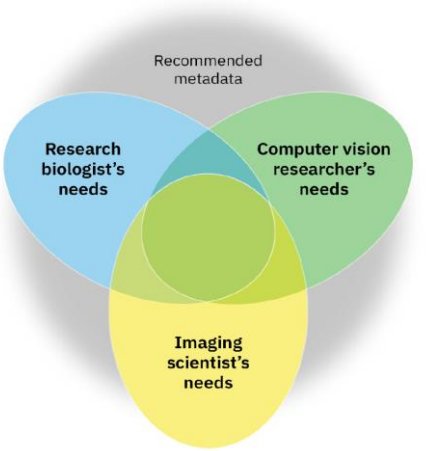
Reusable

- R1 (Meta)data are richly described with a plurality of accurate and relevant attributes (license, provenance, relevant community standards)

FRANCE-BIOIMAGING

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship.
Sci Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

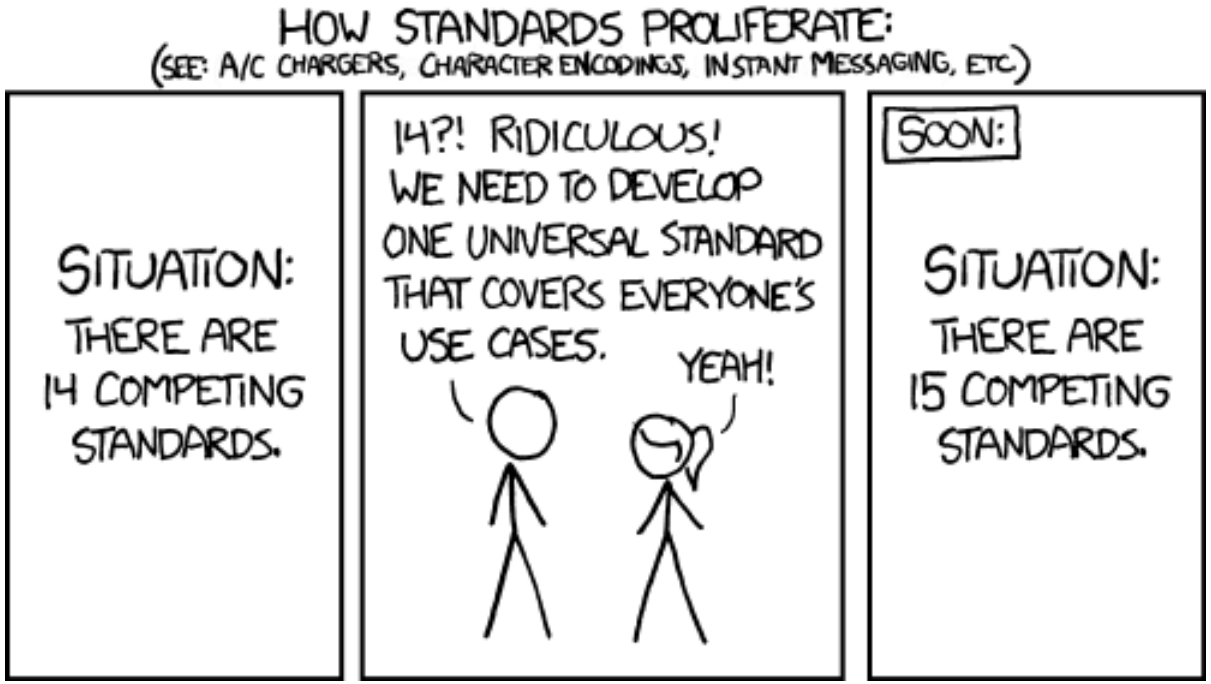
Which [standard] metadata?



- Metadata are domain dependant:
  - What does the related community need?
  - Minimally … « The more the better » does not always apply here.
  - They should be « easy » to get and **has a purpose**

- What is a standard? « widely accepted norm », « agreed-upon way of doing something »

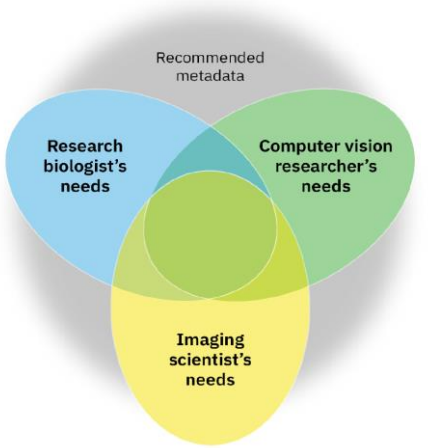
Standards?



Source: xkcd a webcomic of Romance; Sarcasm, math and Language [xkcd: Standards]

FRANCE-BIOIMAGING

Which [standard] metadata?

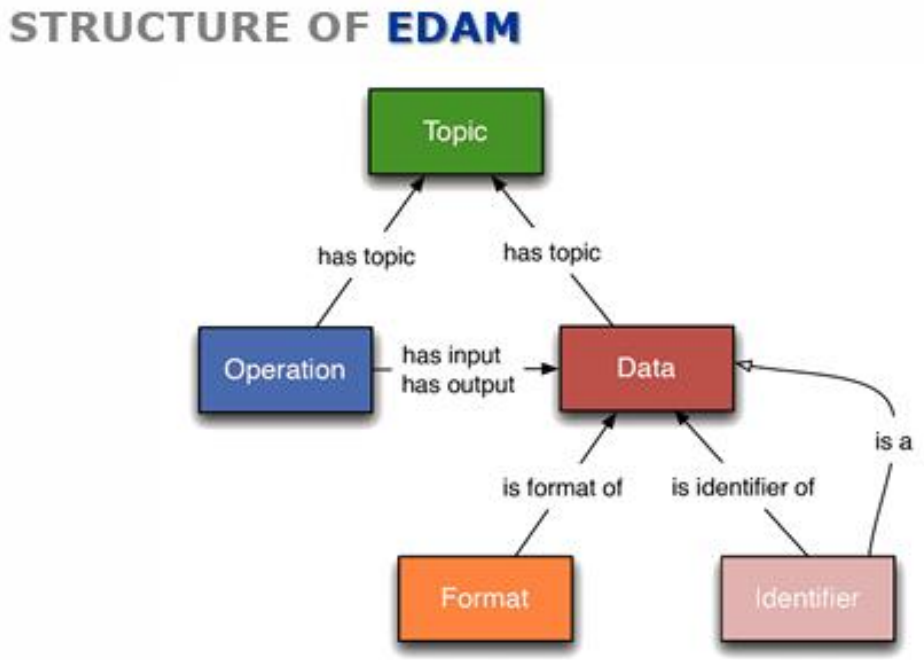


- Metadata are domain dependant:
  - What does the related community need?
  - Minimum Information Model… « The more the better » does not always here.
  - They should be « easy » to get and **has a purpose**
- What is a standard? « widely accepted norm », « agreed-upon way of doing something »
- Controlled Vocabulary VS Ontology
  - CVs is a list of terms that have been enumerated explicitely (to choose from). Each term has a definition.
  - Ontology is a Controlled Vocabulary express with a « grammar »

What are ontologies?

- an ontology is a formal explicit description of concepts in a domain of discourse
  - **classes** (sometimes called concepts),
  - **properties** of each concept describing various features

An ontology together with a set of individual instances of classes constitutes a **knowledge base**.
(simplified definition from WebProtégé)

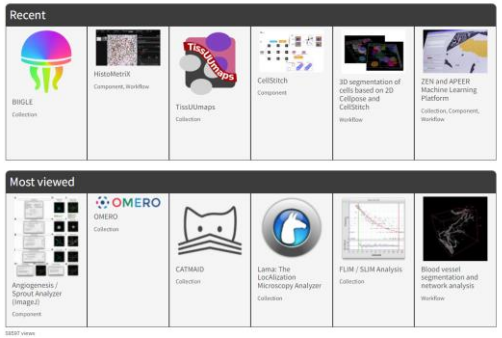


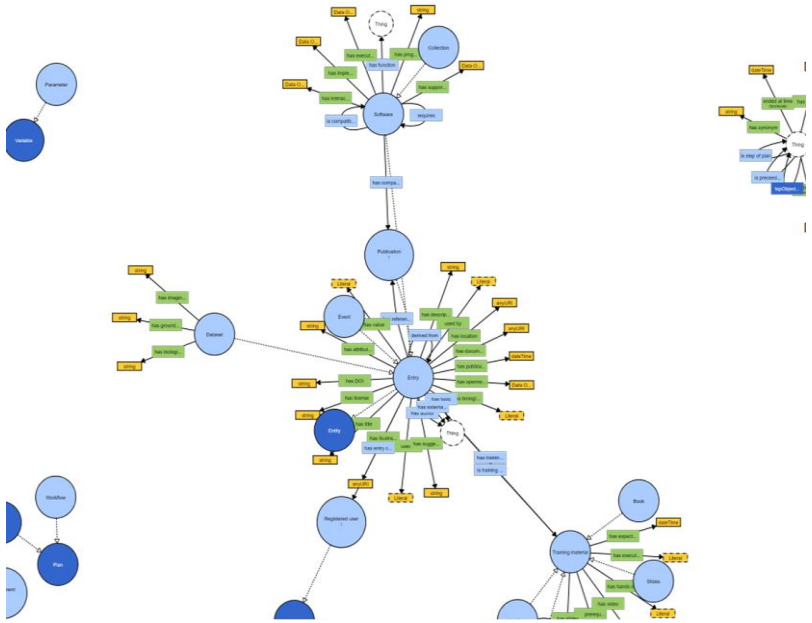
What are ontologies?

Example from BIII

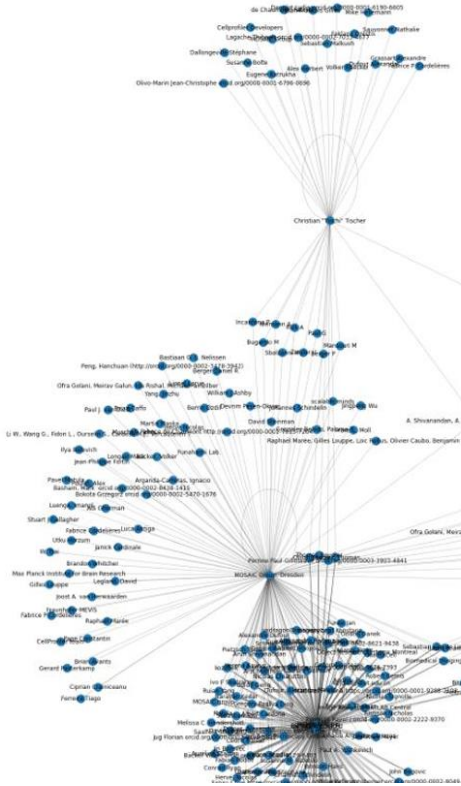
www.biii.eu

https://doi.org/10.48550/arXiv.2312.11256





Ontologies



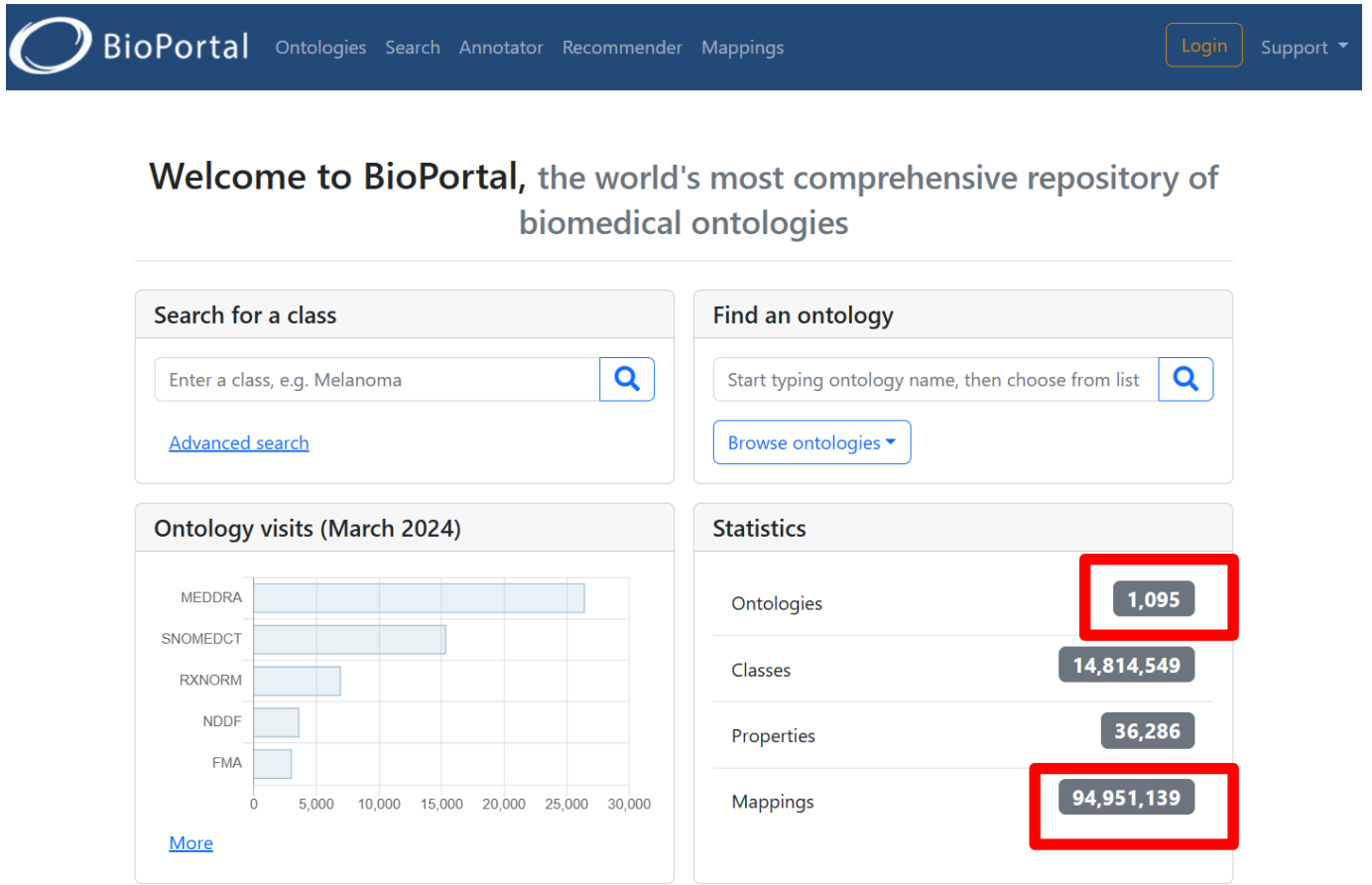
Collaborating software authors based on image analysis operations

Example of queries from the knowledge model

FRANCE-BIOIMAGING

Which metadata?

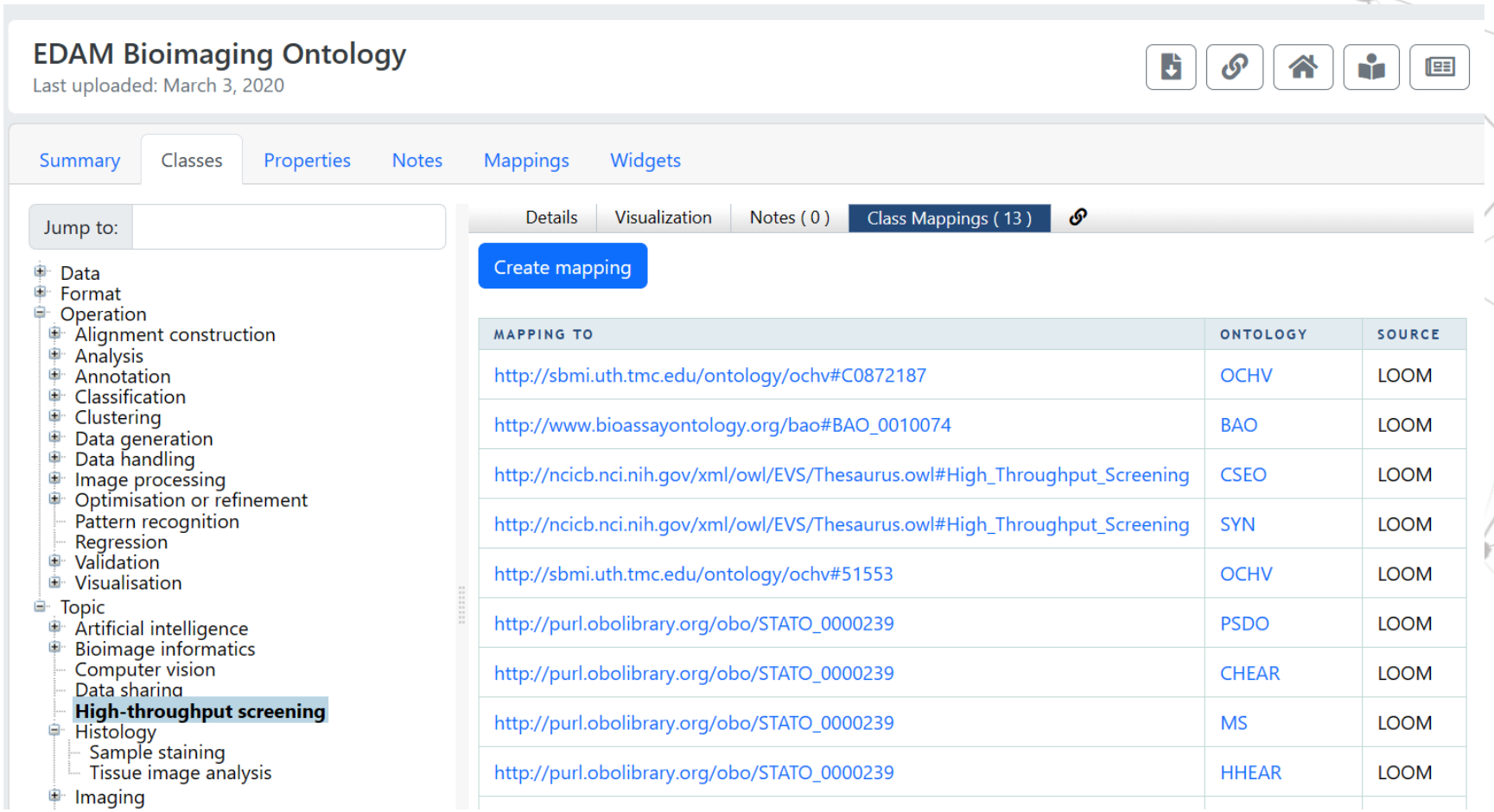
https://bioportal.bioontology.org/



FRANCE-BIOIMAGING

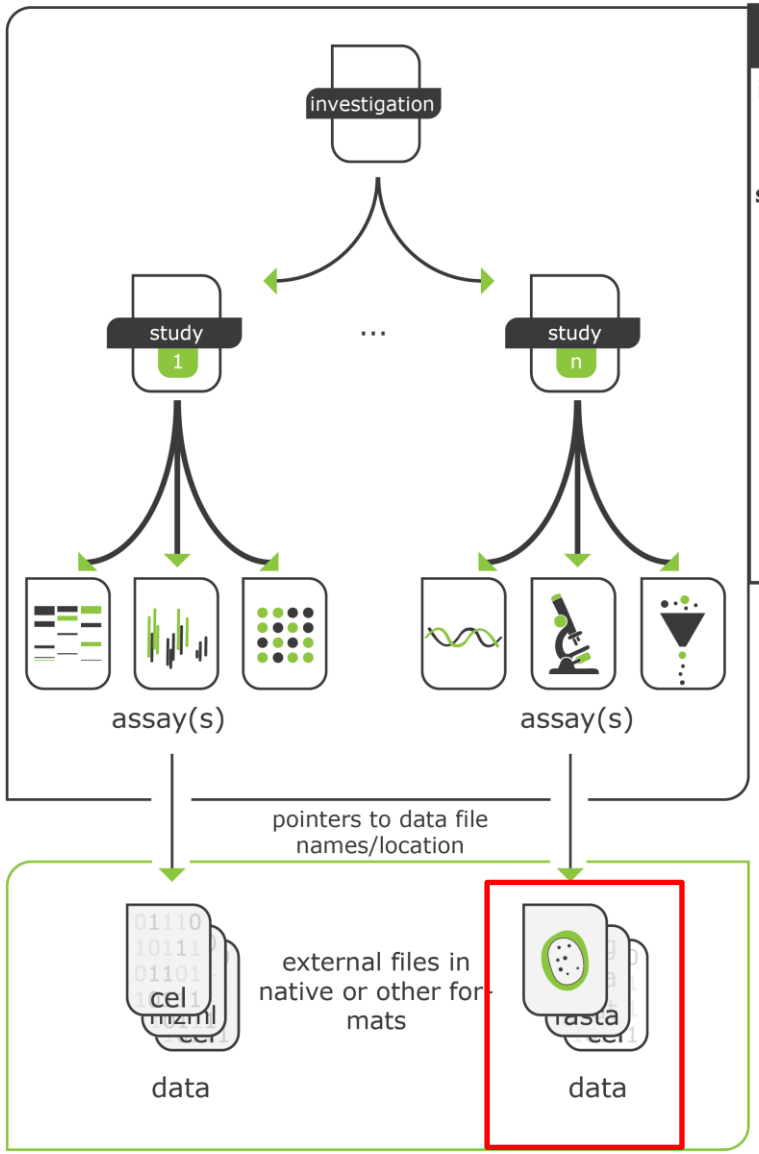
Mapping/Alignment/Matching

Can be automatized by pairing related concepts (BioPortal use LOOM for example based on names and synonyms case-insensitive)



ISA: A generic framework for life science experiments

(Yes, there is more than imaging in life [sciences])



isa

investigation
high level concept to link related studies

study
the central unit, containing information on the subject under study, its characteristics and any treatments applied.
a study has associated ***assays***

assay
test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)

[Specifications | ISA tools (isa-tools.org)]

FRANCE-BIOIMAGING

Recommended Metadata for Biological Images

BioImage Archive

Study is the highest level metadata, describing your project, including funding and publications. ***Dublin Core, Datacite, Schema.org***

Study Component acts as a container that helps you organise your data, based on experiment types or samples etc. A Study Component contains one or more of the following components, or required if relevant).

Biosample is about what it is you have imaged, for example, the species of the organism that you're imaging, if you're using a particular cell line, genetic background etc. ***EFO, NCBI taxonomy***

Specimen metadata describes how your sample was prepared for imaging. ***OME, EFO, EDAM-BIOIMAGING, FBbi***

Image Acquisition is about how your images were captured. ***4DN, QUAREP, Fbbi, EDAM-BIOIMAGING***

Image Data contains image level metadata. This is implemented as the File List for BioImage Archive submissions, instead of a seperate component in the submission form. ***OME, EDAM-BIOIMAGINg***

Image Correlation is optional and contains metadata about correlating images from different modalities. ***EDAM-BIOIMAGING***

Analysed Data, includes **Image Analysis** metadata; information about how you analysed your images, if applicable. ***EDAM BioIMAGING, OME***

Comment | Published: 21 May 2021

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

Ugis Sarkans, Wah Chiu, Lucy Collinson, Michele C. Darrow, Jan Ellenberg, David Grunwald, Jean-Karim Hériché, Andrii Iudin, Gabriel G. Martins, Terry Meehan, Kedar Narayan, Ardan Patwardhan, Matthew Robert Geoffrey Russell, Helen R. Saibil, Caterina Strambio-De-Castillia, Jason R. Swedlow, Christian Tischer, Virginie Uhlmann, Paul Verkade, Mary Barlow, Omer Bayraktar, Ewan Birney, Cesare Catavitello, Christopher Cawthorne, Stephan Wagner-Conrad, Elizabeth Duke, Perrine Paul-Gilloteaux, Emmanuel Gustin, Maria Harkiolaki, Pasi Kankaanpää, Thomas Lemberger, Jo McEntyre, Josh Moore, Andrew W. Nicholls, Shuichi Onami, Helen Parkinson, Maddy Parsons, Marina Romanchikova, Nicholas Sofroniew, Jim Swoger, Nadine Utz, Lenard M. Voortman, Frances Wong, Peijun Zhang, Gerard J. Kleywegt & Alvis Brazma — Show fewer authors

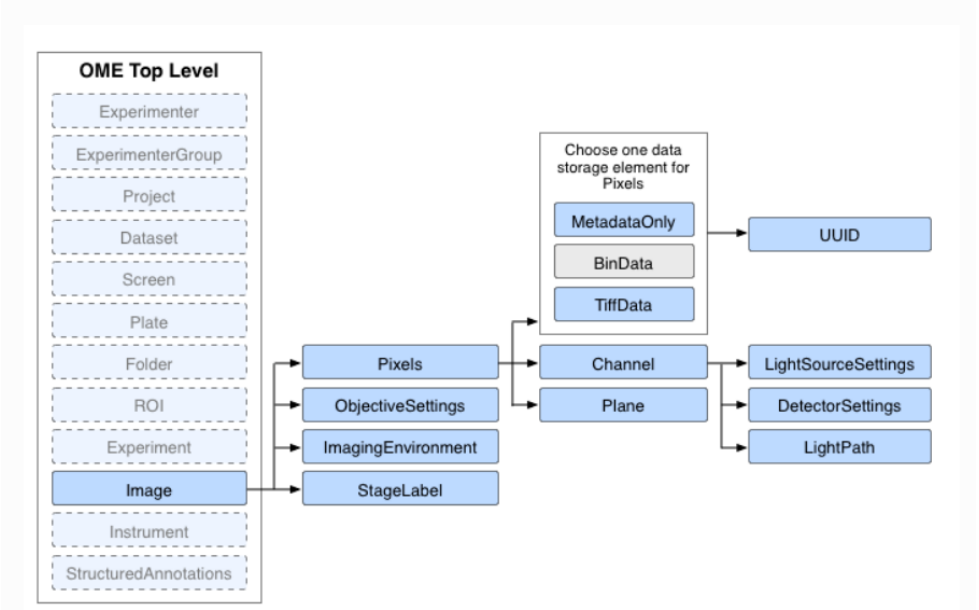
Nature Methods **18**, 1418–1422 (2021) | Cite this article

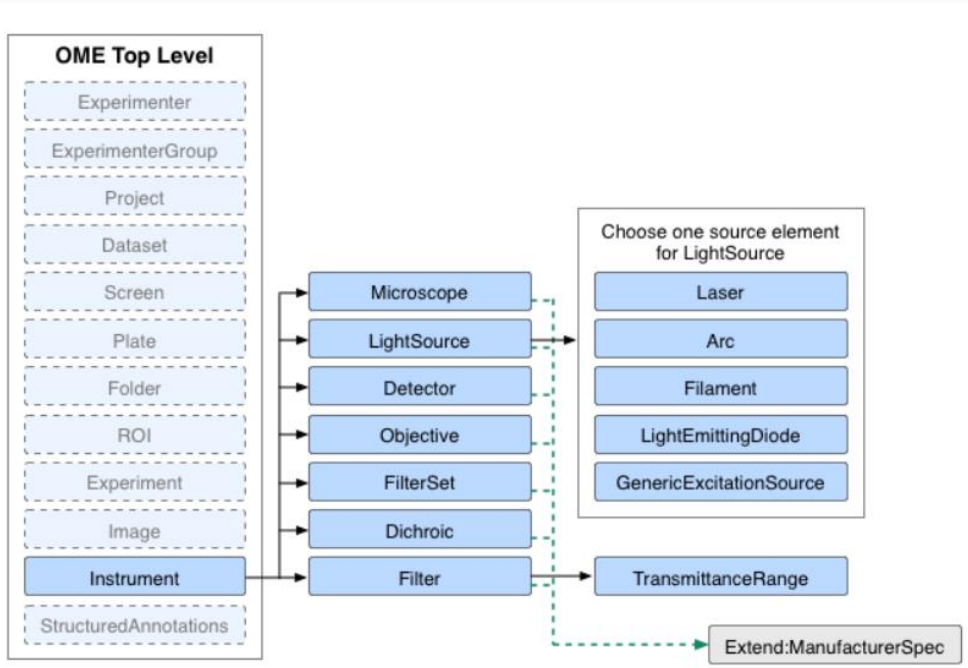
12k Accesses | **55** Citations | **43** Altmetric | Metrics

FRANCE-BIOIMAGING

OME : standard metadata encoded in image file format (our closest to « DICOM » for Connoisseurs…) BioFormat/Ome-Tif/Ome-zarr/Ngff

Used in most bio image software





Instrument branch of the OME Schema

[Current Data Model overview — OME Data Model and File Formats documentation (ome-model.readthedocs.io)]

FRANCE-BIOIMAGING

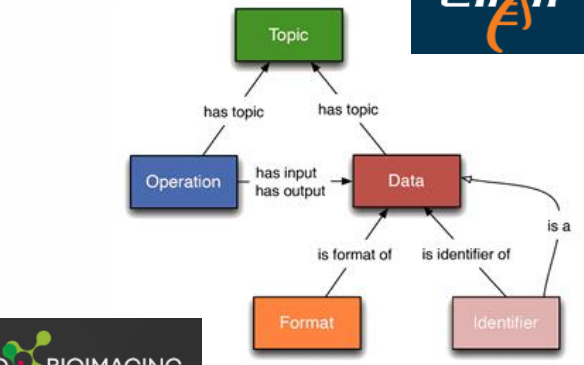
Edam Bioimaging

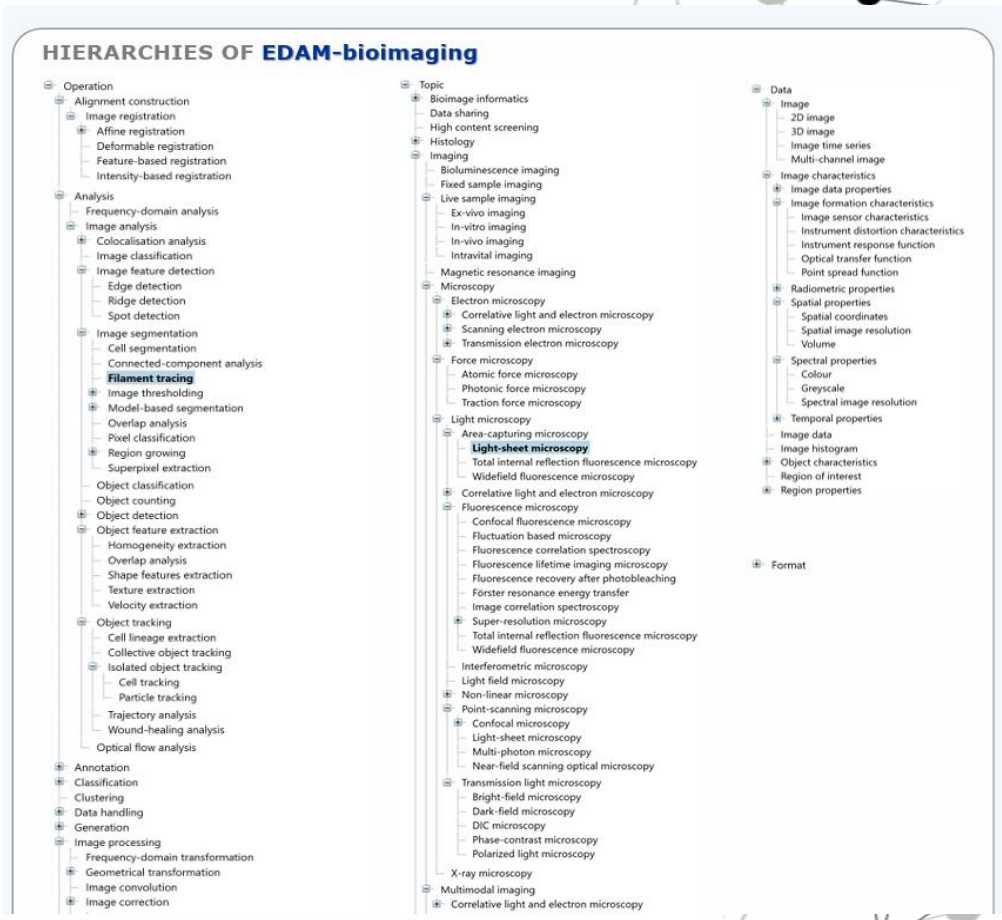
What? EDAM-bioimaging is an extension of the EDAM ontology, dedicated to bioimage analysis, bioimage informatics, and bioimaging.

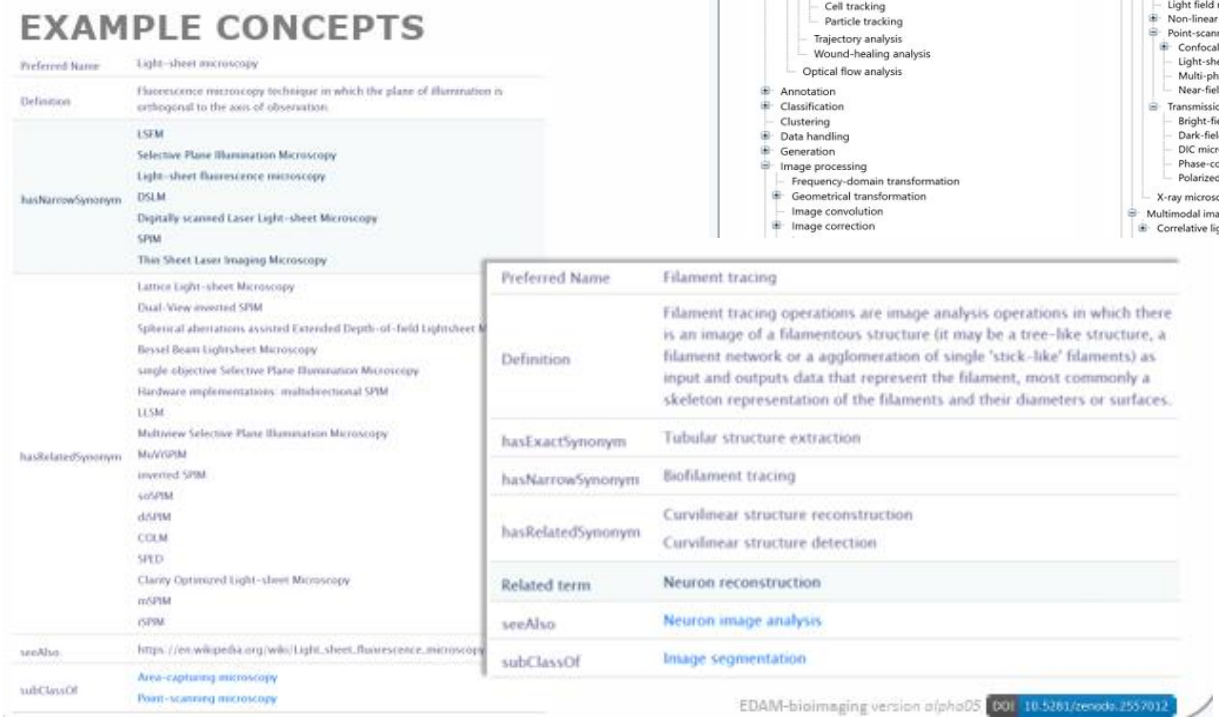
Why? EDAM-bioimaging enables interoperable descriptions of software, publications, data, and workflows, fostering reliable and transparent science.

How? EDAM-bioimaging is developed in a community spirit, in a welcoming collaboration between numerous bioimaging experts and ontology developers.

How can I contribute? We need your expertise! **You can help by reviewing parts of EDAM-bioimaging, posting comments with suggestions, requirements, or needs for clarification**, or participating in a **Taggathon or another hackathon**. Please see https://github.com/edamontology/edam-bioimaging#contributing.

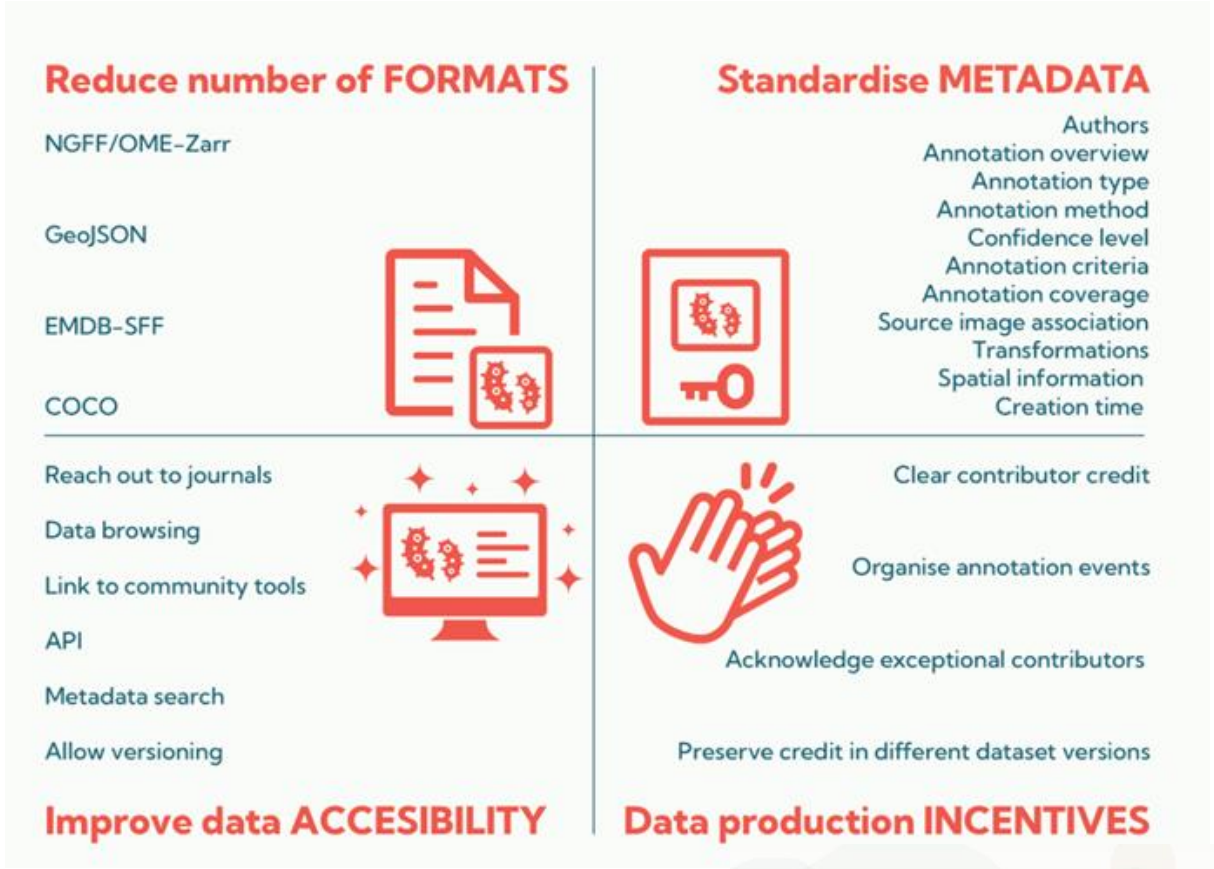


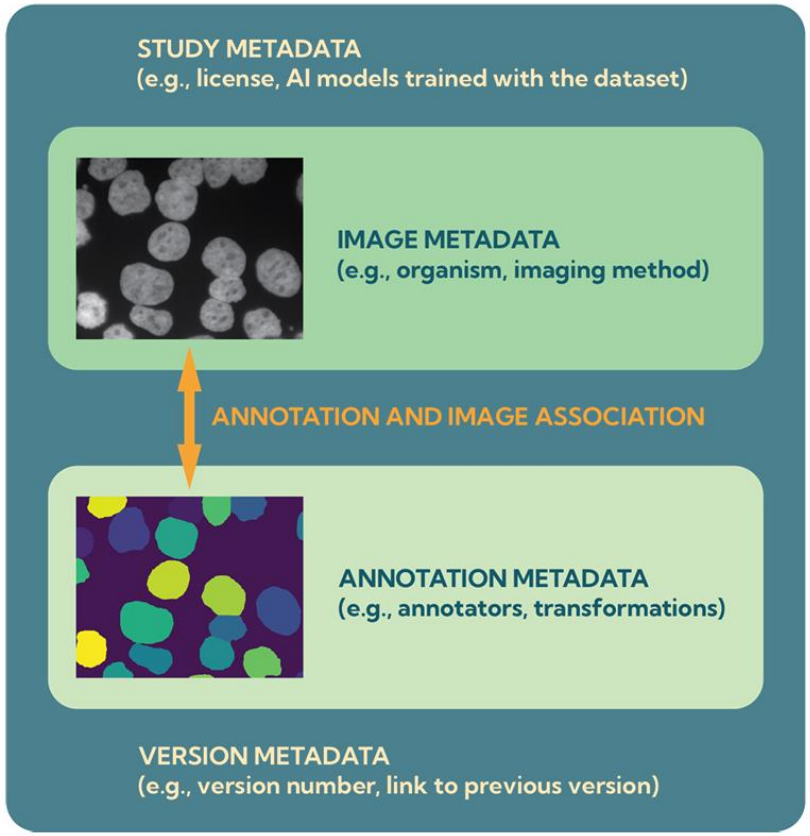




FRANCE-BIOIMAGING

MIFA: Metadata, Incentives, Formats, and Accessibility guidelines





Quantitative Biology > Other Quantitative Biology

[Submitted on 17 Nov 2023 (v1), last revised 22 Nov 2023 (this version, v2)]

MIFA: Metadata, Incentives, Formats, and Accessibility guidelines to improve the reuse of AI datasets for bioimage analysis

Teresa Zulueta-Coarasa, Florian Jug, Aastha Mathur, Josh Moore, Arrate Muñoz-Barrutia, Liviu Anita, Kola Babalola, Pete Bankhead, Perrine Gilloteaux, Nodar Gogoberidze, Martin Jones, Gerard J. Kleywegt, Paul Korir, Anna Kreshuk, Aybüke Küpcü Yoldaş, Luca Marconato, Kedar Narayan, Nils Norlin, Bugra Oezdemir, Jessica Riesterer, Norman Rzepka, Ugis Sarkans, Beatriz Serrano, Christian Tischer, Virginie Uhlmann, Vladimir Ulman, Matthew Hartley

Artificial Intelligence methods are powerful tools for biological image analysis and processing. High-quality annotated images are key to training and developing new methods, but access to such data is often hindered by the lack of standards for sharing datasets. We brought together community experts in a workshop to develop guidelines to improve the reuse of bioimages and annotations for AI applications. These include standards on data formats, metadata, data presentation and sharing, and incentives to generate new datasets. We are positive that the MIFA (Metadata, Incentives, Formats, and Accessibility) recommendations will accelerate the development of AI tools for bioimage analysis by facilitating access to high quality training data.

Comments: 16 pages, 3 figures
Subjects: **Other Quantitative Biology (q-bio.OT)**; Image and Video Processing (eess.IV)
Cite as: arXiv:2311.10443 [q-bio.OT]
(or arXiv:2311.10443v2 [q-bio.OT] for this version)
https://doi.org/10.48550/arXiv.2311.10443

FRANCE-BIOIMAGING

AI4Life

Take Home Messages

- The list of metadata has to be identified from the beginning of a project (Data Management Plans)
- Plan and spend some times curating your data according to where you will deposit them (DMP)
- Try to fit a common effort with no reinvention of the wheel: rather contribute to make existing metadata framework to evolve
- Alignment of ontologies (Mapping) is a key for interoperability and make your initial choice less dramatic

EURO-BIOIMAGING

GLOBAL BIOIMAGING
growing collaboration

Thanks for your attention!

Ome Team
Josh Moore
Jason Swedlow
Kevin Eliceiri

FBI data team
Emmanuel Faure
Guillaume Gay
Marc Mongy
Jean-François Guillaume
Guillaume Maucort
Raphaël Braud-Mussi
Thierry Pecot
Dorian Kauffmann
Théo Barnouin

Elixir /IFB Team
Jon Ison
Matus Kalas
Hervé Ménager
Alban Gaignard

BioImageArchive Team
Matthew Hartley

Neubias taggers
https://biii.eu/show-taggers

FRANCE-BIOIMAGING

ANR