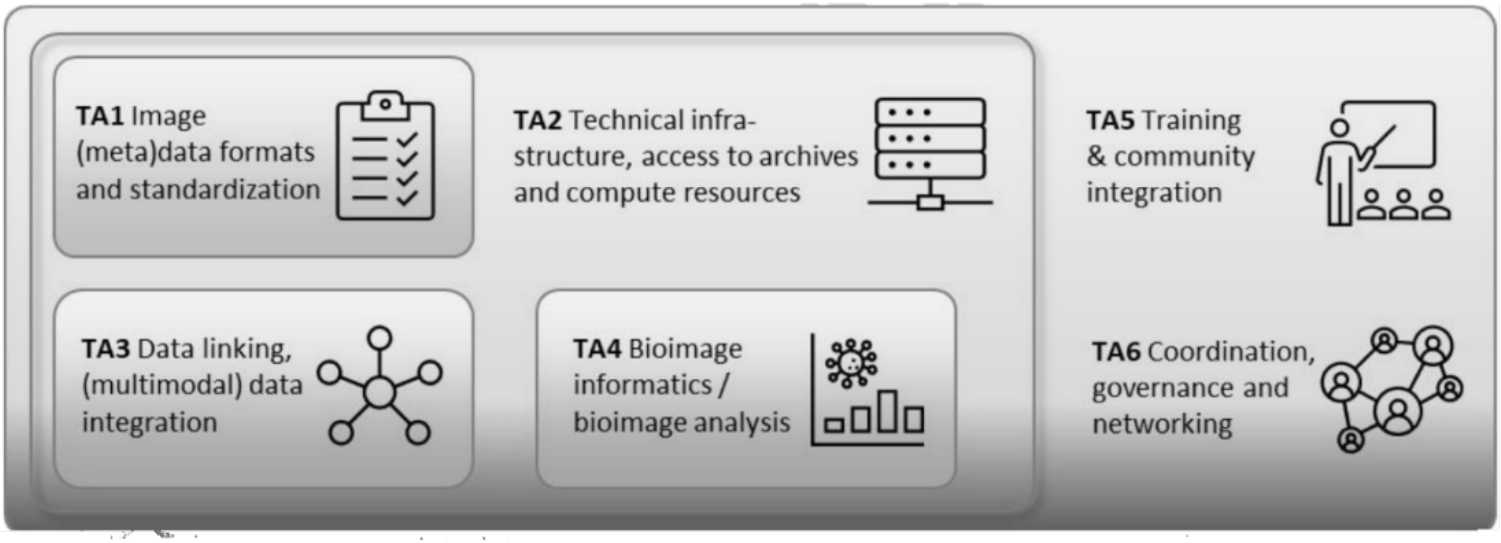
NFDI4BIOIMAGE

Carsten Fortmann-Grote[1]

April 18 2024

[1] carsten.fortmann-grote@evolbio.mpg.de

The NFDI4BIOIMAGE Consortium



- Harmonization of bioimage data and metadata formats and standards
- FAIR Image Objects
- Multimodal metadata vocabularies
- Standard Operating Procedures for FAIR image processing
- Proliferation of best practices in image data management and processing

All Hands Meeting Oct. 2023

No Milestones due yet

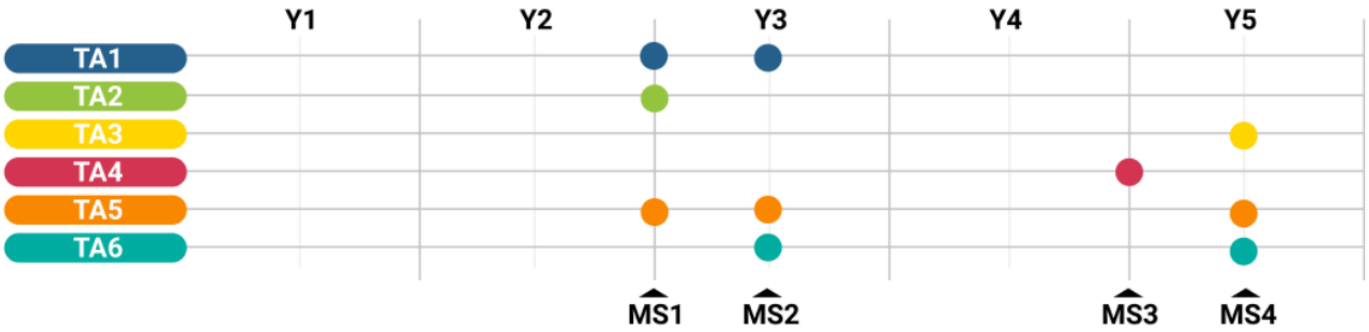


Figure 13: Overview of the project milestones, consisting of the indicated milestone parts from the Task Areas.

Milestone 1: White paper on the FAIR-IO concept published (MS1-TA1), production services for NFDI4BIOIMAGE available (MS1-TA2) and search index for bioimage RDM training materials implemented (MS1-TA5)

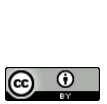
Milestone 2: Release of RFC process document (MS2-TA1), first NFDI4BIOIMAGE conference (MS2-TA6), and publication of survey results and search index statistics (MS2-TA5)

Milestone 3: SOPs for FAIR usage of image analysis workflows available (MS3-TA4)

Milestone 4: Second NFDI4BIOIMAGE conference (MS4-TA6), preview of over-arching database for multimodal metadata (MS4-TA3), and publication of the search index statistics and survey results (MS4-TA5)

Outline

- FAIRification of image data: 5 star scheme
- Teaching & Training material
- BIDS, ARC, OMERO interfaces
- Data integration at MPI-EvolBio

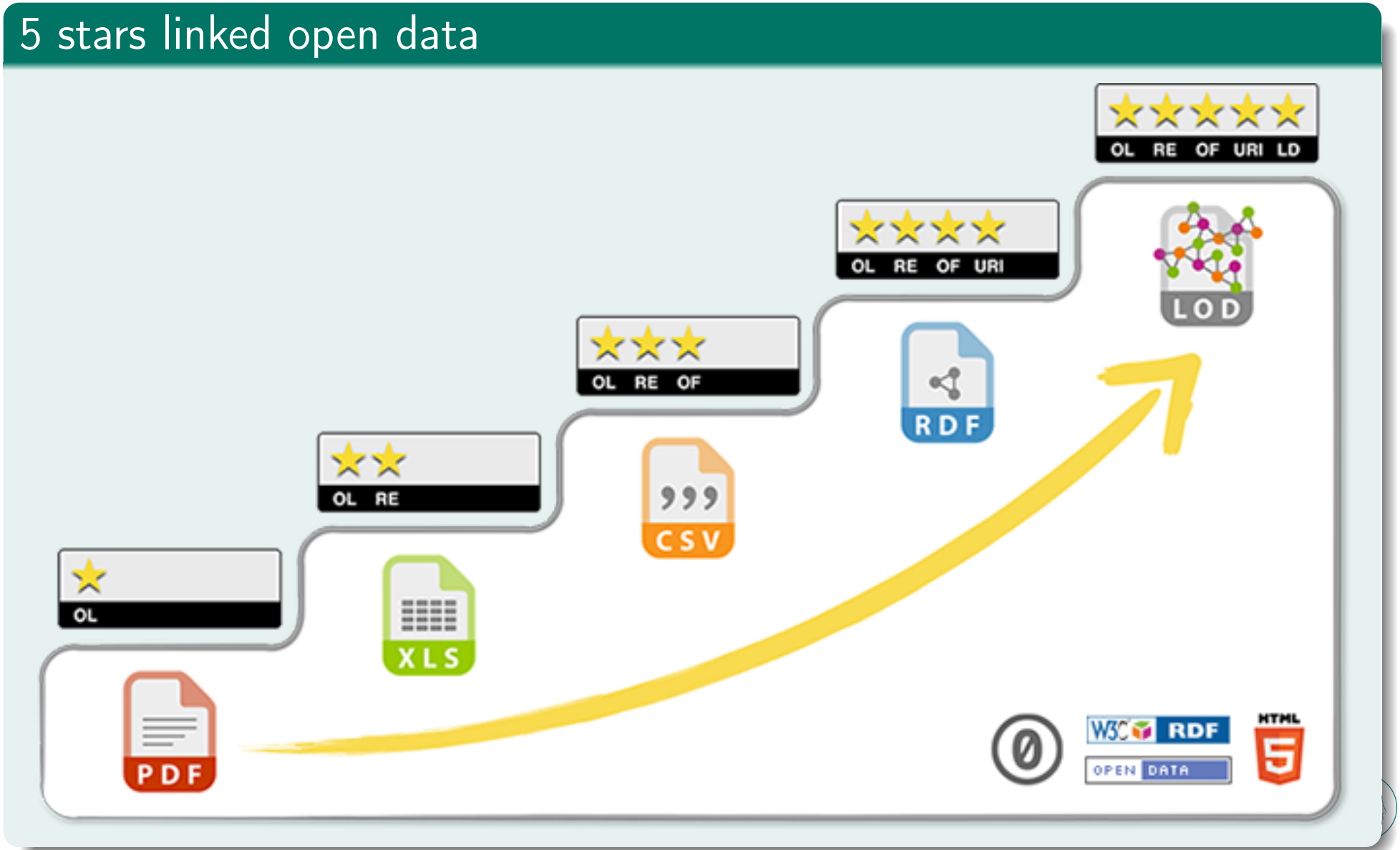



FAIRification of Image (meta)data

Quantify the FAIRness of containerized Bioimage research objects

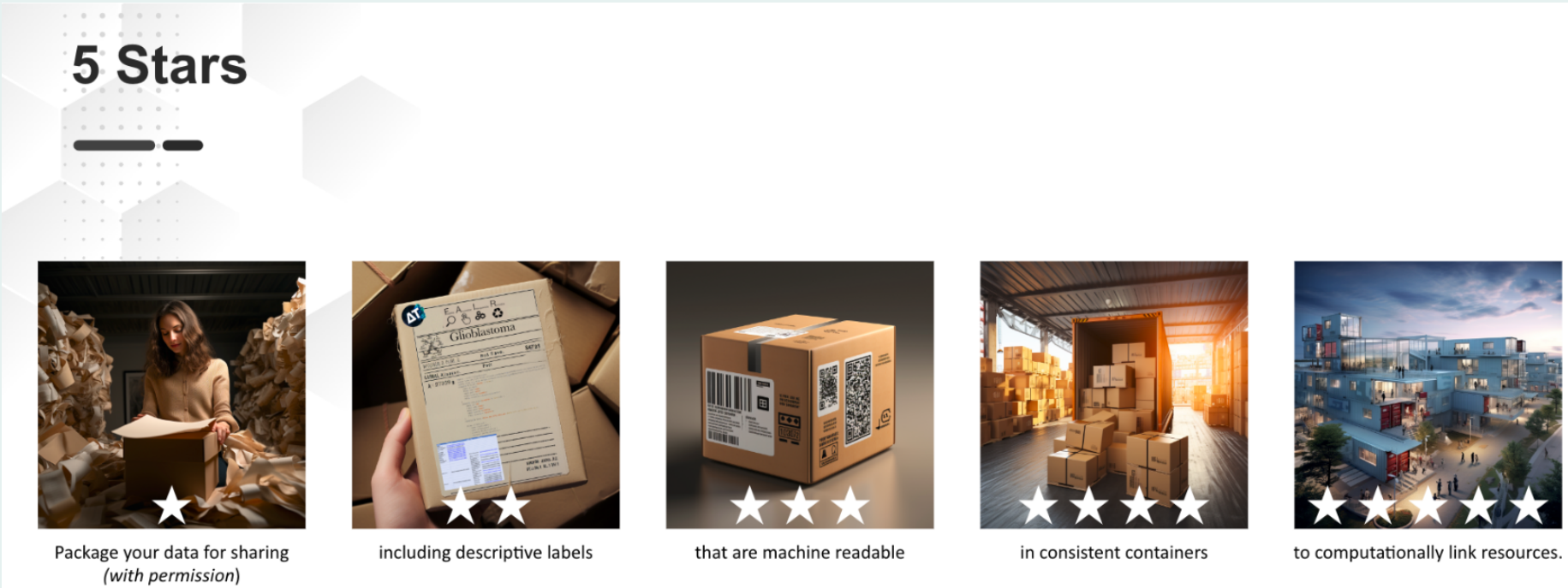
FAIRification of Image (meta)data

5 stars linked open data



FAIRification of Image (meta)data

5 star bioimage containers

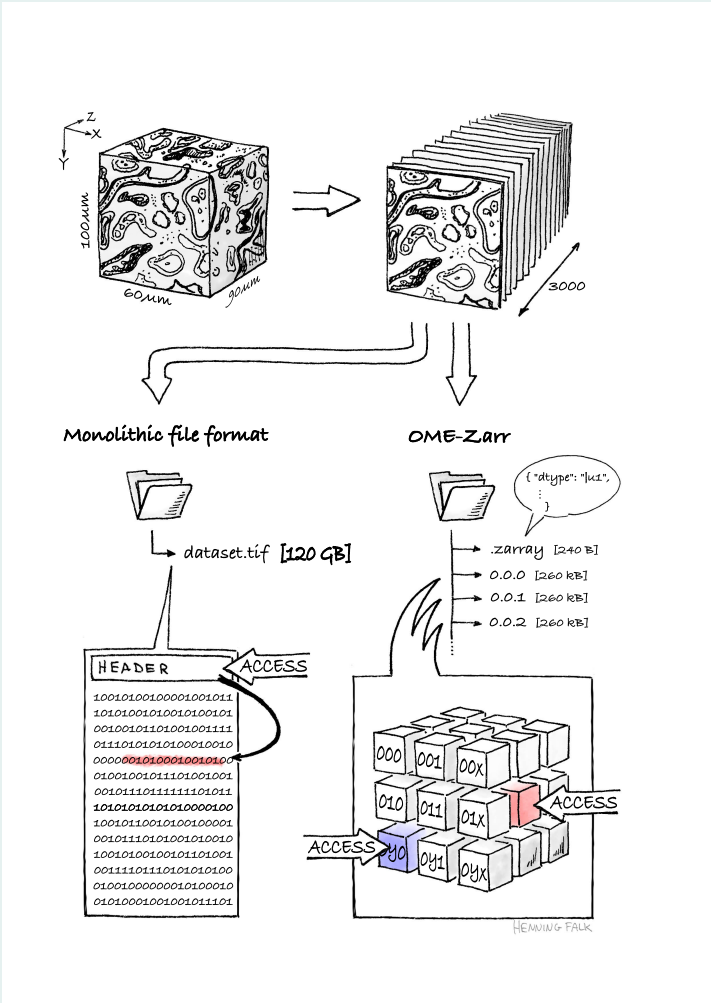


Images generated by Andra Waagmeester & Josh Moore through AI generation using Midjourney (Sep. 2023)

Zarr is the technical backbone of next generation (bio)image file formats

What is Zarr?

- hdf5 like hierarchical layers
- chunks stored in separate "files" (blobs), linked by json file (vs. monolithic binary blob in hdf5)
- suitable for object storage

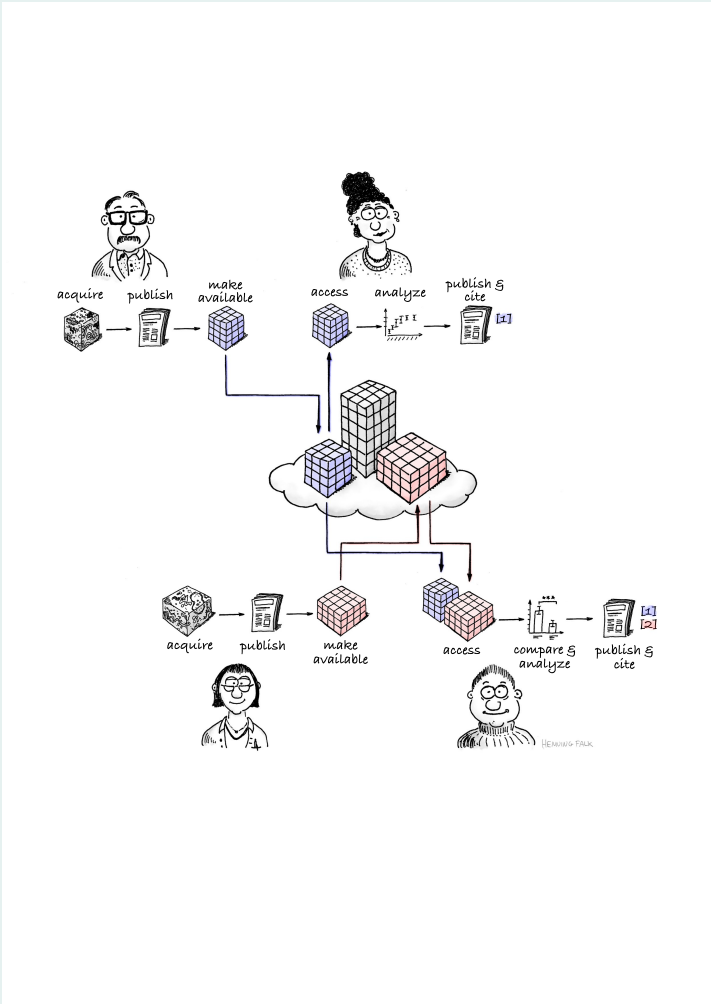


10.5281/zenodo.7037679

Zarr is the technical backbone of next generation (bio)image file formats

What is Zarr?

- IO latency issues
- NFDI4BI leads community efforts for specification and implementation
- BioImage Archive (BIA), NASA, Open Geospatial Consortium (OGC) have committed to adopting Zarr as community standard

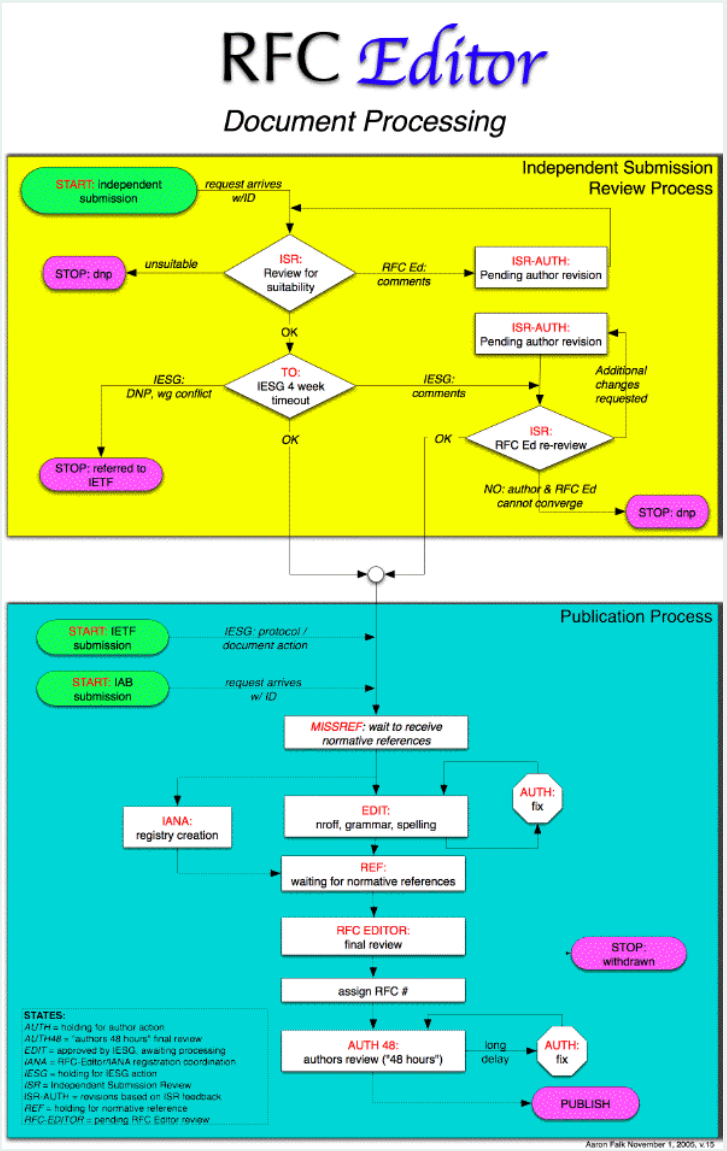


10.5281/zenodo.7037679

The RFC (Request For Comments) process

- Established Internet Engineering Task Force (IETF) procedure for
  - protocols (HTTP, FTP)
  - best practices (RFC)
  - standards (RDF)
  - file formats (XML, HTML)
  - anything that defines the internet

The RFC process



Example: RFC2119

BEST CURRENT PRACTICE
Updated by: 8174 Errata Exist
Network Working Group S. Bradner
Request for Comments: 2119 Harvard University
BCP: 14 March 1997
Category: Best Current Practice

Key words for use in RFCs to Indicate Requirement Levels

Status of this Memo

This document specifies an Internet Best Current Practices for the Internet Community, and requests discussion and suggestions for improvements. Distribution of this memo is unlimited.

Abstract

In many standards track documents several words are used to signify the requirements in the specification. These words are often capitalized. This document defines these words as they should be interpreted in IETF documents. Authors who follow these guidelines should incorporate this phrase near the beginning of their document:

> The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

Note that the force of these words is modified by the requirement level of the document in which they are used.

1. **MUST** This word, or the terms "REQUIRED" or "SHALL", mean that the definition is an absolute requirement of the specification.

2. **MUST NOT** This phrase, or the phrase "SHALL NOT", mean that the definition is an absolute prohibition of the specification.

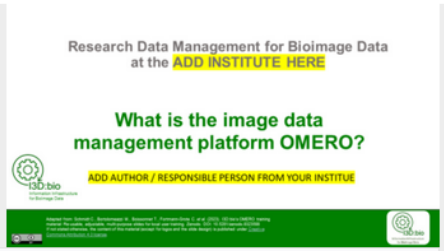
3. **SHOULD** This word, or the adjective "RECOMMENDED", mean that there may exist valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.

4. **SHOULD NOT** This phrase, or the phrase "NOT RECOMMENDED" mean that there may exist valid reasons in particular circumstances when the particular behavior is acceptable or even useful, but the full implications should be understood and the case carefully weighed before implementing any behavior described with this label.

I3D-bio customizable slide decks

Training material by the I3D:bio team

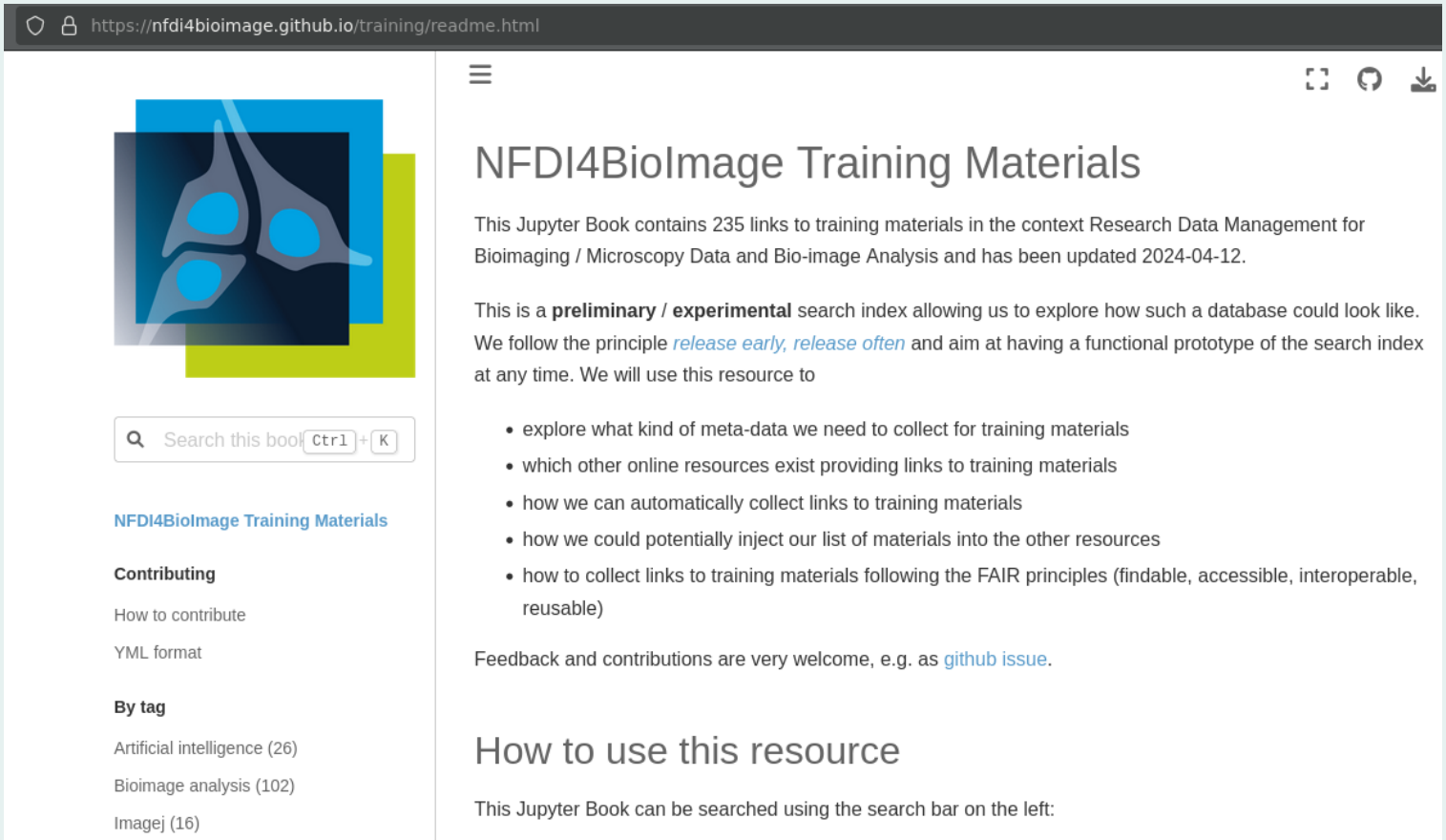
I3D:bio's OMERO training material: Re-usable, adjustable, multi-purpose slides for local user training



Description | Links | License & Citation

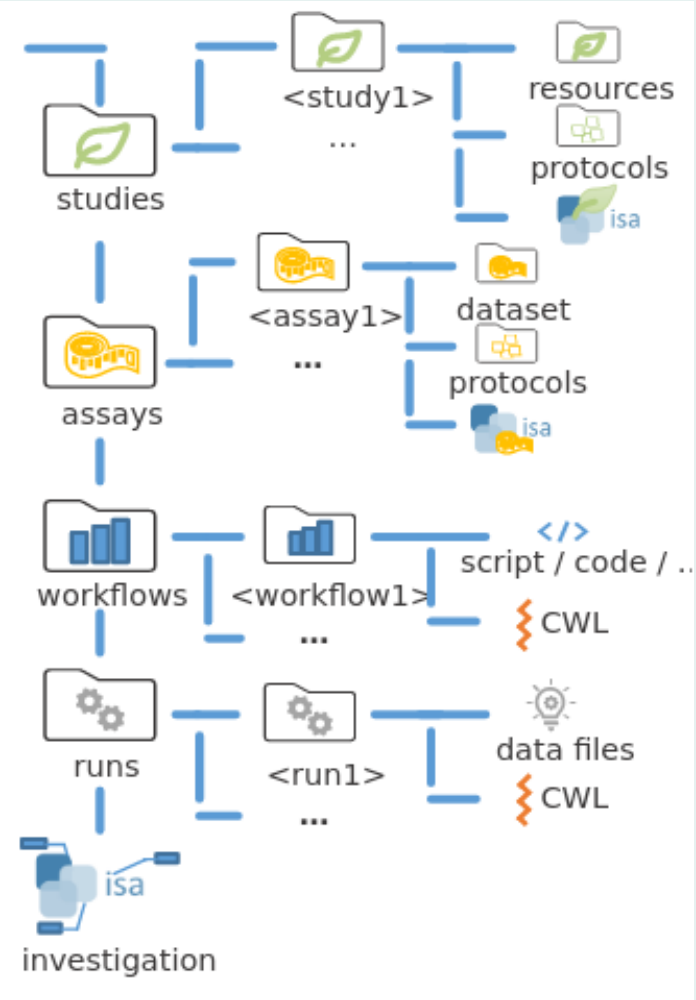
- Record on zenodo with all material for download: https://zenodo.org/records/8323588
- Youtube playlist of the video tutorials: https://www.youtube.com/playlist?list=PL2k-L-zWPoR7SHjG1HhDIwLZj0MB_stIU

Jupyter Book + Search Index

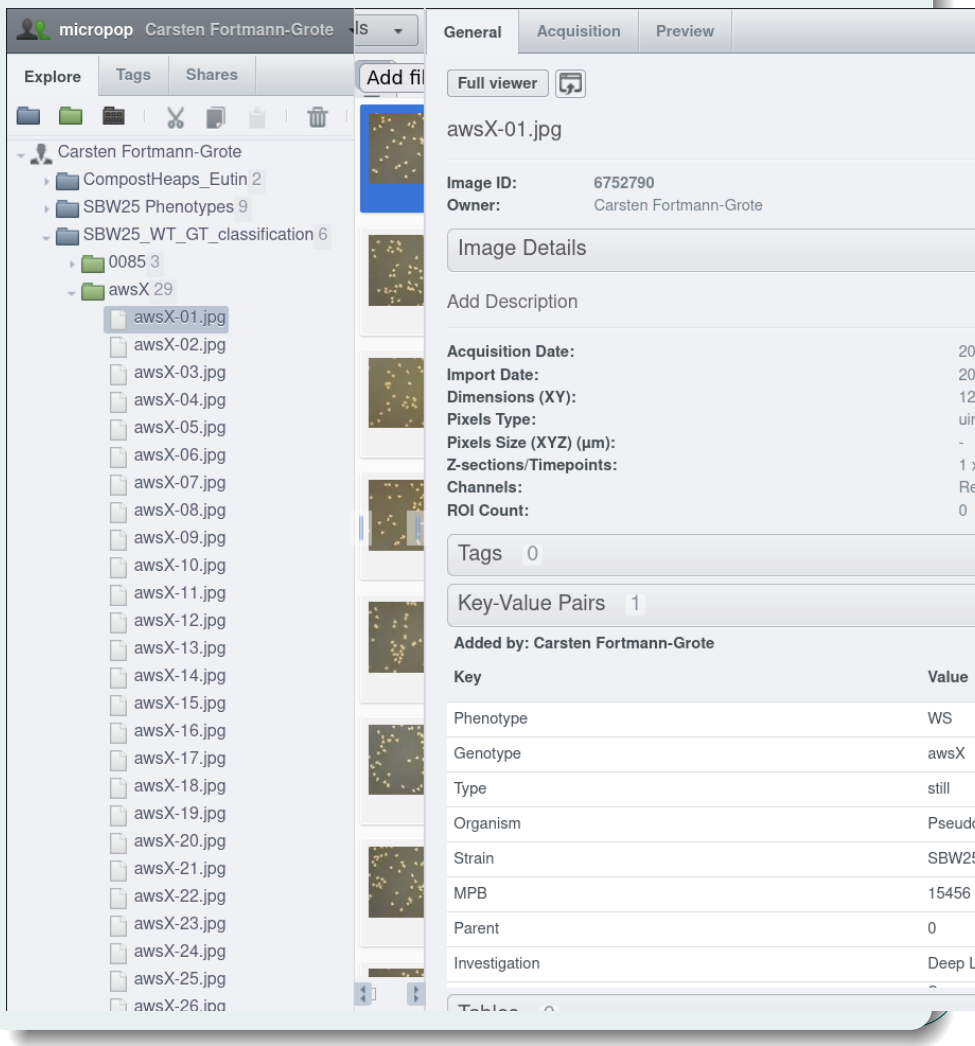


BIDS-ARC-OMERO (TA1, TA3)

ARC (Annotated Research Context, NFDI4Plants)



OMERO



Many open questions

- One-to-one mapping of ARC entities to/from OMERO?
- What about non-image data in ARC?
- OMERO as metadata manager on top of ARC?
- round trip possible?
- large image files (think TB) in git?


Mapping (attempts)

Electronic labbook

Clearing protocol with TO-PRO-1 and tert-butanol
- transcradial perfusion with 1x PBS (phosphate-buffered saline) (pH 7.4) as pre-rinse with 4% PFA (perfluoroalkoxy) (pH 7.4) as fixative
- 4h post-fixation in 4% PFA at 4°C
- storage of the tissue in 1x PBS until the start of clearing
- wash in 1x PBS --> 3x 30min

Permeabilization solution (2 d)
- (1.5% Goat serum,
- 0.5% Triton X-100,
- 0.5 mM methyl-betacyclodextrin,
- 0.2% trans-1-acetyl-4-hydroxy-proline,
- 0.05% sodium acide in 1x PBS)
at 37°C, slow rotation (in 20 ml rolled rim tubes)

Incubation solution (5 d)
- (= permeabilization solution plus TO-PRO-1, 1:1000)
at 37°C, slow rotation, dark!

Washing in wash solution (3x 2 h)
- (1.5% Goat serum,
- 0.5% Triton-X-100,
- 0.05% Sodium acide in 1x PBS)
at 37°C, slow rotation, dark

OME-XML Metadata (from ome.tif File)


```
ImageDescription: <?xml version="1.0" encoding="UTF-8"?><OME xmlns="http://
www.openmicroscopy.org/Schemas/OME/2008-02" xmlns:ca="http://www.
openmicroscopy.org/Schemas/CA/2008-02" xmlns:xsi="http://www.w3.org/2001/
XMLSchema-instance" xsi:schemaLocation="http://www.openmicroscopy.org/
Schemas/OME/2008-02 http://www.openmicroscopy.org/Schemas/OME/2008-02/ome.
xsd" UUID="urn:uuid:251F06D7-A0F4-471C-84E5-A908DE0AFE37">
<Experimenter ID="Experimenter:Tobias_Gottschall">
<FirstName>Tobias</FirstName>
<LastName>Gottschall</LastName>
<Email>tobias.gottschall@                    </Email>
<Institution>LIN</Institution>
</Experimenter>
<Group ID="urn:lsid:imspector.group.id:Group:SP-ELMI" Name="SP-ELMI"/>
<Image ID="Image:100DD12B-DD11-4E7C-AACD-4C14334C7D88"
Name="13-05-56_Standard_gerbil_5418_TOPRO-3_AF_4x_2L5_2mue_nDF_4x5_20percen
t_150ms_UltraII[00 x 00]_C00_xyz-Table 20000.ome.tif"
DefaultPixels="Pixels:100DD12B-DD11-4E7C-AACD-4C14334C7D88">
<CreationDate>2020-12-12T12:33:26</CreationDate>
<Description>not_specified</Description>
<Pixels ID="Pixels:100DD12B-DD11-4E7C-AACD-4C14334C7D88"
DimensionOrder="XYZCT" PixelType="uint16" BigEndian="false" SizeX="2048"
SizeY="2048" SizeZ="3585" SizeT="1" SizeC="2" PhysicalSizeX="1.625000"
PhysicalSizeY="1.625000" PhysicalSizeZ="2.000000">
<TiffData FirstZ="0" FirstC="0">
<UUID
FileName="13-05-56_Standard_gerbil_5418_TOPRO-3_AF_4x_2L5_2mue_nDF_4x5_20pe
rcent_150ms_UltraII[00 x 00]_C00_xyz-Table 20000.ome.
tif">urn:uuid:251F06D7-A0F4-471C-84E5-A908DE0AFE37</UUID>
</TiffData>
..
<Filter0 ExcitationWL="500" EmissionWL="535"/>
<Filter1 ExcitationWL="642" EmissionWL="661"/>
</OME>
```


BIDS microscopy full example

```
atlas_transform_2d-to-3d
├── dataset_description.json
├── README
├── samples.tsv
├── sourcedata
│   ├── protocol
│   │   ├── clearing_protocol_SD832_full.json
│   │   └── Miltenyi_UltraMicroscopeII_manuel.pdf
│   ├── sub-200928
│   │   └── micr
│   │       ├── 200928_[00 x 00]_C00.ome.tif
│   │       ├── 200928_[00 x 00]_C01.ome.tif
│   │       ..
│   ├── sub-THG5418
│   │   └── micr
│   │       ├── THG5418_[00 x 00]_C00.ome.tif
│   │       ├── THG5418_[00 x 00]_C01.ome.tif
│   │       ..
├── sub-200928
│   └── micr
│       ├── sub-200928_sample-tissue_FLUO.json
│       └── sub-200928_sample-tissue_FLUO.ome.zarr
├── sub-THG5418
│   └── micr
│       ├── sub-THG5418_sample-tissue_FLUO.json
│       └── sub-THG5418_sample-tissue_FLUO.ome.zarr
├── derivatives
│   ├── 3D-gerbilIN-atlas_alpha
│   │   ├── 3D-gerbilIN-atlas_alpha_template.nii
│   │   ├── 3D-gerbilIN-atlas_alpha_annotations.nii
│   │   └── 3D-gerbilIN-atlas_alpha_labels.csv
│   ├── downsampled
│   │   ├── sub-200928
│   │   │   └── micr
│   │   │       ├── sub-200928_sample-tissue_res-17x_FLUO.json
│   │   │       └── sub-200928_sample-tissue_res-17x_FLUO.ome.zarr
│   │   └── sub-THG5418
│   │       └── micr
│   │           ├── sub-THG5418_sample-tissue_res-17x_FLUO.json
│   │           └── sub-THG5418_sample-tissue_res-17x_FLUO.ome.zarr
│   ├── metricCalc
│   │   ├── 201205_LSstandardMouse
│   │   ├── 20201005_072782_Rat
│   │   └── 5679_LSgerbil
│   └── publication
│       ├── Gottschall_2022_MSC_gerbil-atlas_2d-to-3d.pdf
│       └── gerbil-3d-atlas-development_2021_pre-registration.pdf
└── code
    ├── example
    │   └── reworked_gerbil_itk_snap_label.label
    ├── imgs
    │   └── ImageJ_background_grey-value.png
    ├── README.md
    ├── run_main.py
    └── src
        └── 2D_multi-registration_transformation_v.py
```

FIGURE 3 (bottom-left) export of the OME-XML metadata from an ome.tif file automatically created by the control software of BioTec UltraMicroscope II; (right) manually created dataset as complete as possible of the development of a 2D to 3D Mongolian gerbil brain atlas using light sheet microscopy [7]

LEGEND FIGURE 3 & 4 (blue background) dataset components related to the creation, transformation, or analysis of the data, such as source code or parameters necessary for the analysis; (light red background) study data itself as well as its descriptive metadata; (yellow background) analysis results, figures created from the data, and components essential for the publication of the results; (triple color background) file content must be cut and reassigned to multiple other files; (part of figures below grey bar) these contents are part of the dataset but not exclusive or mandatory for the BIDS standard; (part of figures below blue bar) valid recommended or mandatory structures of the BIDS standard; (part of figures below red bar) valid recommended or mandatory structures of the ARC standard

ARC conversion of BIDS example

```
atlas_transform_2d-to-3d
├── isa.investigation.xlsx
├── assays
│   ├── atlas_transform_2d-to-3dDownsampled
│   │   ├── dataset
│   │   │   ├── sub-200928
│   │   │   │   └── micr
│   │   │   │       ├── sub-200928_sample-tissue_res-17x_FLUO.json
│   │   │   │       ├── sub-200928_sample-tissue_res-17x_FLUO.ome.zarr
│   │   │   │       ..
│   │   │   └── sub-THG5418
│   │   │       └── micr
│   │   │           ├── sub-THG5418_sample-tissue_res-17x_FLUO.json
│   │   │           └── sub-THG5418_sample-tissue_res-17x_FLUO.ome.zarr
│   │   ├── isa.assay.xlsx
│   │   ├── protocols
│   │   │   ├── clearing_protocol_SD832_full.json
│   │   │   └── Miltenyi_UltraMicroscopeII_manuel.pdf
│   │   └── README.md
│   ├── atlas_transform_2d-to-3dRegular
│   │   ├── dataset
│   │   │   ├── sub-200928
│   │   │   │   └── micr
│   │   │   │       ├── sub-200928_sample-tissue_FLUO.json
│   │   │   │       ├── sub-200928_sample-tissue_FLUO.ome.zarr
│   │   │   │       ..
│   │   │   └── sub-THG5418
│   │   │       └── micr
│   │   │           ├── sub-THG5418_sample-tissue_FLUO.json
│   │   │           └── sub-THG5418_sample-tissue_FLUO.ome.zarr
│   │   ├── isa.assay.xlsx
│   │   ├── protocols
│   │   │   ├── clearing_protocol_SD832_full.json
│   │   │   └── Miltenyi_UltraMicroscopeII_manuel.pdf
│   │   └── README.md
│   └── atlas_transform_2d-to-3dSourceData
│       ├── dataset
│       │   ├── sub-200928
│       │   │   └── micr
│       │   │       ├── 200928_[00 x 00]_C00.ome.tif
│       │   │       ├── 200928_[00 x 00]_C01.ome.tif
│       │   │       ..
│       │   └── sub-THG5418
│       │       └── micr
│       │           ├── THG5418_[00 x 00]_C00.ome.tif
│       │           ├── THG5418_[00 x 00]_C01.ome.tif
│       │           ..
│       ├── isa.assay.xlsx
│       ├── protocols
│       │   ├── clearing_protocol_SD832_full.json
│       │   └── Miltenyi_UltraMicroscopeII_manuel.pdf
│       └── README.md
├── runs
│   ├── 3D-gerbilIN-atlas_alpha
│   │   ├── 3D-gerbilIN-atlas_alpha_annotations.nii
│   │   ├── 3D-gerbilIN-atlas_alpha_labels.csv
│   │   └── 3D-gerbilIN-atlas_alpha_template.nii
│   ├── metricCalc
│   │   ├── 201205_LSstandardMouse
│   │   ├── 20201005_072782_Rat
│   │   └── 5679_LSgerbil
├── studies
│   └── atlas_transform_2d-to-3dStudy
│       ├── isa.study.xlsx
│       ├── protocols
│       │   └── clearing_protocol_SD832_full.json
│       ├── README.md
│       ├── resources
│       │   ├── example
│       │   │   └── reworked_gerbil_itk_snap_label.label
│       │   ├── imgs
│       │   │   └── ImageJ_background_grey-value.png
│       │   └── publication
│       │       ├── Gottschall_2022_MSC_gerbil-atlas_2d-to-3d.pdf
│       │       └── gerbil-3d-atlas-development_2021_pre-registration.pdf
└── workflows
    ├── README.md
    ├── run.cwl
    ├── run.yml
    └── src
        ├── 2D_multi-registration_transformation_v.py
        ├── run_main.py
        └── stats_testing.py
```

isa.assay.xlsx SWATE spreadsheet

Source Name	Characteristic [sample type]	Parameter [Instrument Model]	Component [Perfusion]	Characteristic [fixative]	Term Source REF (CHEBI:50913
			phosphate buffered	perfluoroalkyl	
sub-200928	Tissue	Light Microscopy	saline solution	substance	CHEBI
			phosphate buffered	perfluoroalkyl	
sub-THG5418	Tissue	Light Microscopy	saline solution	substance	CHEBI

Term Accession Number (CHEBI:50913	Characteristic [Time]	Characteristic [temperature day]	Characteristic [ratio]
library.org/obo/		Celsius based	percentage
CHEBI_172397	4.00 hour	unit	based unit
ibrary.org/obo/		Celsius based	percentage
CHEBI_172397	4.00 hour	unit	based unit

isa.study.xlsx SWATE spreadsheet

Source Name	Sample Name	Characteristic [Animal]	Characteristic [Age]
sub 200928	sample-tissue	Mouse	14.00 week
sub-THG5418	sample-tissue	Mongolian Gerbil	16.00 week

isa.investigation.xlsx ARC structure metadata

ONTOLOGY SOURCE REFERENCE	
INVESTIGATION	
Investigation Identifier	atlas_transform_2d-to-3dInvestigation
Investigation Title	atlas_transform_2d-to-3d
Investigation Description	
Investigation Submission Date	
Investigation Public Release Date	
INVESTIGATION PUBLICATIONS	
INVESTIGATION CONTACTS	
Investigation Person Last Name	Gottschall
Investigation Person First Name	Tobias
Investigation Person Mid Initials	
Investigation Person Email	tobias.gottschall@lin-magdeburg.de
Investigation Person Phone	
Investigation Person Fax	
Investigation Person Address	
Investigation Person Affiliation	Leibniz Institute for Neurobiology
Investigation Person Roles	
Investigation Person Roles Term Accession Number	
Investigation Person Roles Term Source REF	
STUDY	
Study Identifier	atlas_transform_2d-to-3dStudy
Study Title	
Study Description	
Study Submission Date	
Study Public Release Date	
Study File Name	atlas_transform_2d-to-3dStudy/isa.study.xlsx
STUDY DESIGN DESCRIPTORS	
STUDY PUBLICATIONS	
STUDY FACTORS	
STUDY ASSAYS	
Study Assay Measurement Type	
Study Assay Measurement Type Term Accession Number	
Study Assay Measurement Type Term Source REF	
Study Assay Technology Type	
Study Assay Technology Type Term Accession Number	
Study Assay Technology Type Term Source REF	
Study Assay Technology Platform	
Study Assay File Name	atlas_transform_2d-to-3dAssay/isa.assay.xlsx
STUDY PROTOCOLS	
STUDY CONTACTS	

FIGURE 4 (left) representation of the directory structure of a valid dataset in the ARC standard manually converted from BIDS directory (compare **FIGURE 3** right); (top-right) small section of clearing_protocol_SD832_full.json transfered from electronic labbook and OME-XML metadata to isa.assay.xlsx file using the SWATE tool [8]; section of BIDS samples.tsv and clearing_protocol_SD832_full.json file transfered to isa.study.xlsx

DOI:10.5281/zenodo.8349563

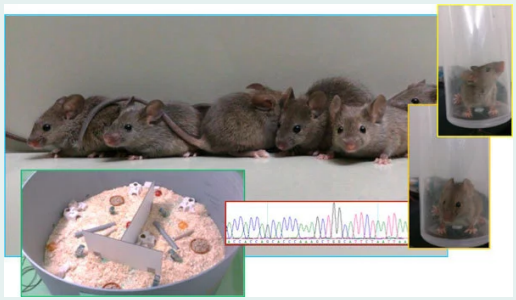
Conclusions so far

- conversion OMERO $\rightarrow$ BIDS $\rightarrow$ ARC seems possible
- only for already specified measurements (BIDS standard)
- requires constant reformatting (e.g. fastq to csv)
- $\Leftrightarrow$ Support by format specific processing clients


MPI-EvolBio's activities

Evolutionary Genetics (D. Tautz (Emeritus))

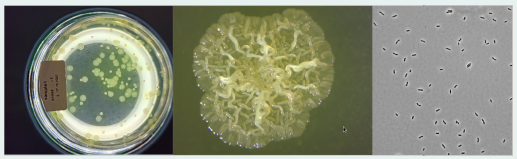
- Model organisms: Mus domesticus, mus musculus
- Behavioural genomics
- Population genetics



credits:
https://www.evolbio.mpg.de/3039130/group_evolanimalbehpers

Microbial Population Biology (P. Rainey)

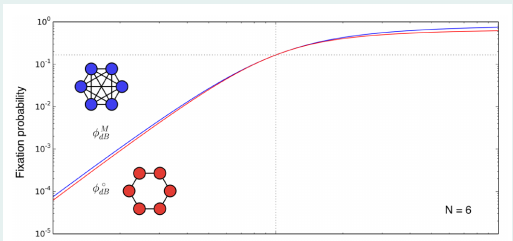
- Model organism: Pseudomonas fluorescens, Bacillus sub.
- Evolution of communities
- Host-microbe interactions
- Genetics



credits: Theodosiou (left), Schwarz (middle), Grote (right)

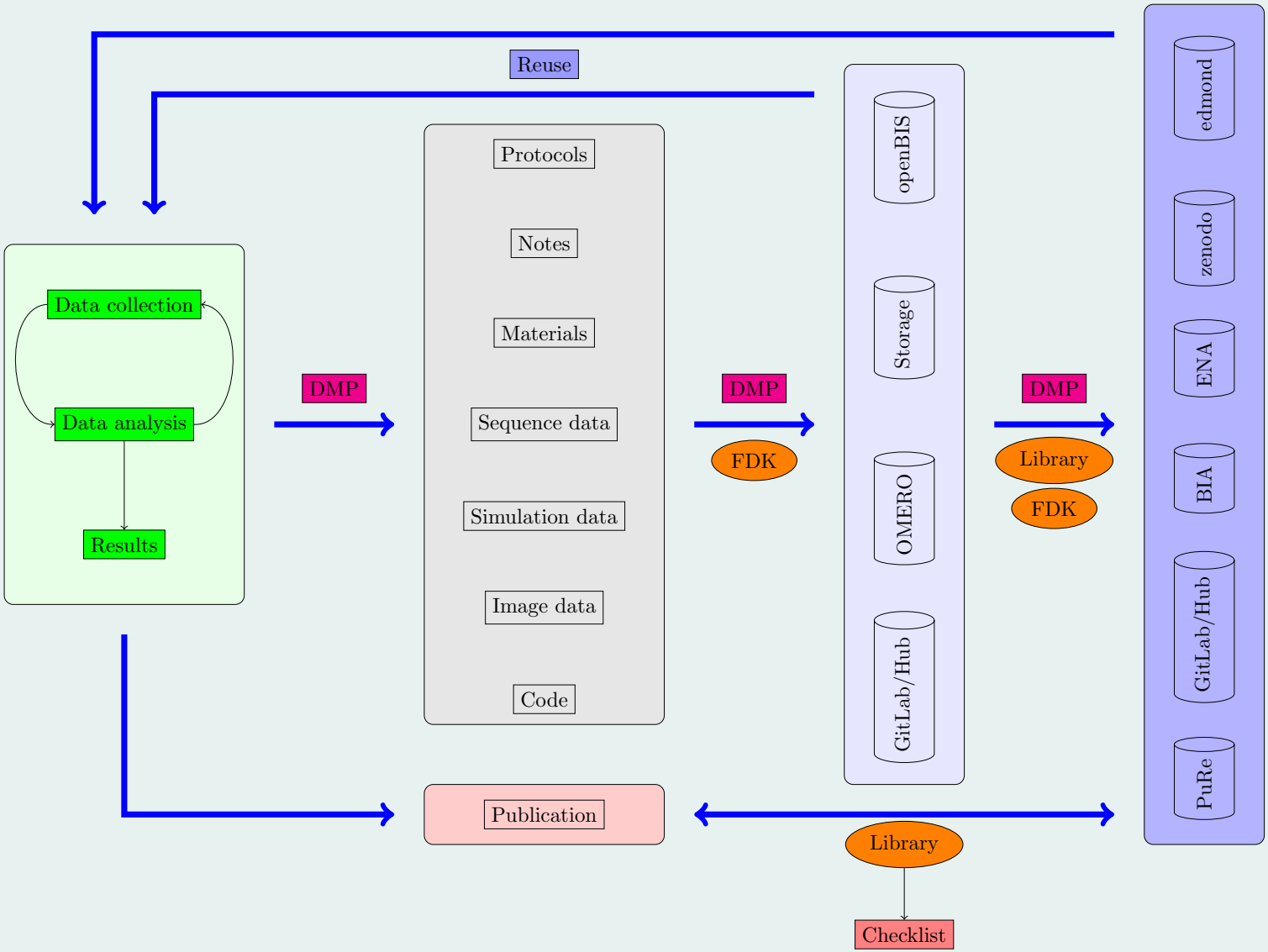
Theoretical Biology (A. Traulsen)

- Population Structure and Game Theory
- Metaorganisms
- Cancer Evolution



Hindersin et al, PLOS Comp. Bio (2019)
10.1371/journal.pcbi.1004437

Data Management Policy enforces deposition of all research data



Multimodal, multidimensional data in experimental evolution research

- Wildtype clone and/or genetically modified strain(s)
- Timelapse microscopy from growing microbial communities
- Time resolved whole genome / core gene NGS data
- Transcription profiles
- Optical density measurements from plaque assays
- Functional annotations

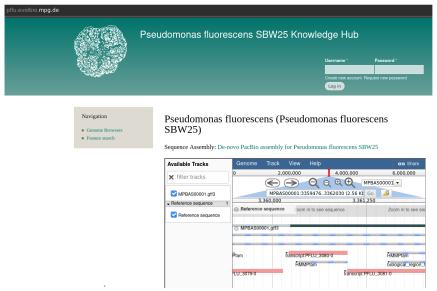
Need for integrated analysis of multiomics-multimodal data

"Find images, ELN entries for Δ mreB strains and annotations for mreB"



Integrating data the SPARQLing way

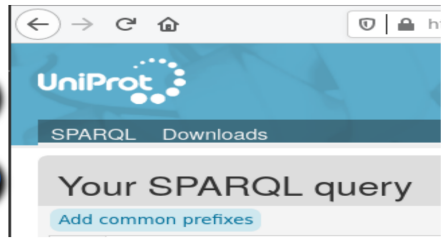
Pseudomonas fluorescens SBW25 genome database (Tripalv3)

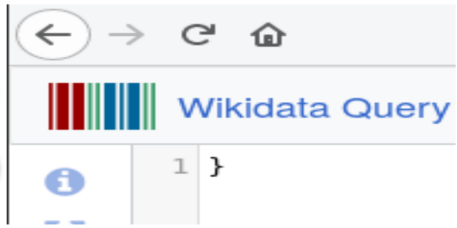


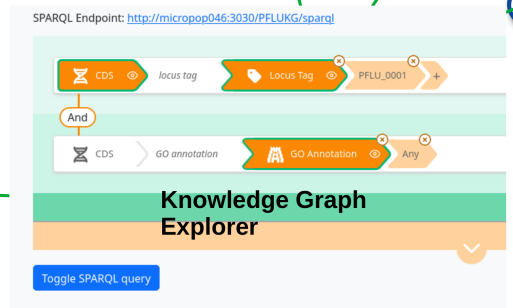
JSON API wrapper

RDF

Remote SPARQL endpoints







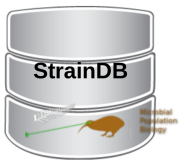
Internal DBs & web resources

API ???

RDF

omero-rdf

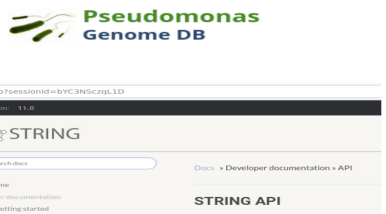
RDF



csv2rdf

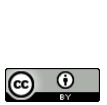
RDF

DBs with or without dedicated API



Data integration of internal sources

DB	API	LOD ready?	RDF conversion
Tripal Genome DB	JSON-LD	yes	SPARQL SPARQL-anything virtualization JSON to RDF serialization
OMERO	JSON-LD	yes	omero-rdf
OpenBIS	JSON-LD	yes	??? ("semantic annotation"
StrainDB	None	no	csv dump → csv2rdf



Flattening a deeply nested JSON-LD graph

[12]:

Client

Client-73a06e67-d796-11ee-8cd8-e454e86b0546

Connection method: Direct

Dashboard: http://172.16.5.46:42575/status

Launch dashboard in JupyterLab

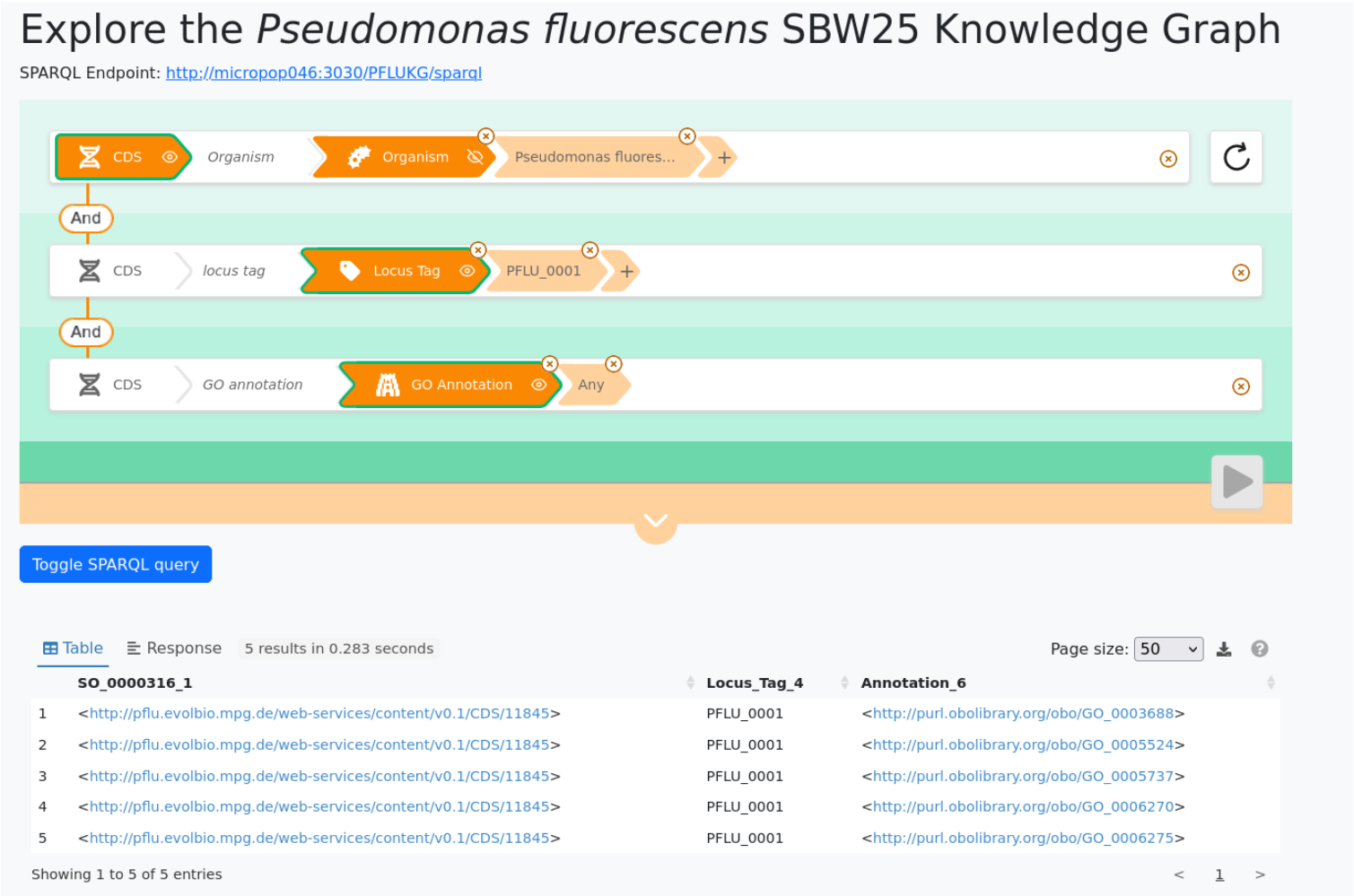
Scheduler Info

[27]:

```python
def get_features(label, serialize=False, distribute=False):
    uri = URIRef('http://pflu.evolbio.mpg.de/web-services/content/v0.1/{}'.format(label))

    logger.info("Getting %s.", uri)
    graph = get_features_graph(uri)
    logger.info("Received %d terms.", len(graph))

    so_id = query_so_id(label)
    so_uri = SO.term(so_id)

    logger.info("SO term for %s is %s", label, so_id)

    subjects =   [s for s in graph.subjects(predicate=rdfs_type,
                        object=so_uri,
                        unique=True)]

    logger.info("Preparing to get %d %s features.", len(subjects), label)

    if distribute:
        logger.info("Running distributed jobs.")
        for g in client.gather([client.submit(get_feature_graph, s) for s in subjects]):
            graph += g
        logger.info("Distributed jobs ended.")
    else:
        for s in tqdm(subjects):
            graph += get_feature_graph(s)

    logger.info("Distributed job finished. Setting bindings and clean up.")
    set_bindings(graph)
    cleanup(graph)

    if serialize:
        fname = f"{label}_{today}.ttl"
        graph.serialize(fname)
        logger.info("Serialized graph to %s.", fname)

    return graph
```

- rdflib for loading json-ld into graph structure
- iteratively parses linked graph uris
- dask for distributed computing
- serialized to turtle format
- -> 2.1 million Triples in ca. 12hrs

SparNatural for visual query generation



Tasks and contributions to NFDI4BI

- P.flu SBW25 Knowledge Graph as a Use Case
- Controlled vocabularies for microbial population biology and evolutionary cell biology
- Contributions to linked open data software, clients, recommendations