

Open image data at scale

From molecules to organisms

Matthew Hartley

Team Leader, BioImage Archive, EMBL-EBI

OME Community Meeting, May 2024



Uh-oh...



“Deposit data in a trusted repository and provide open access to it.”

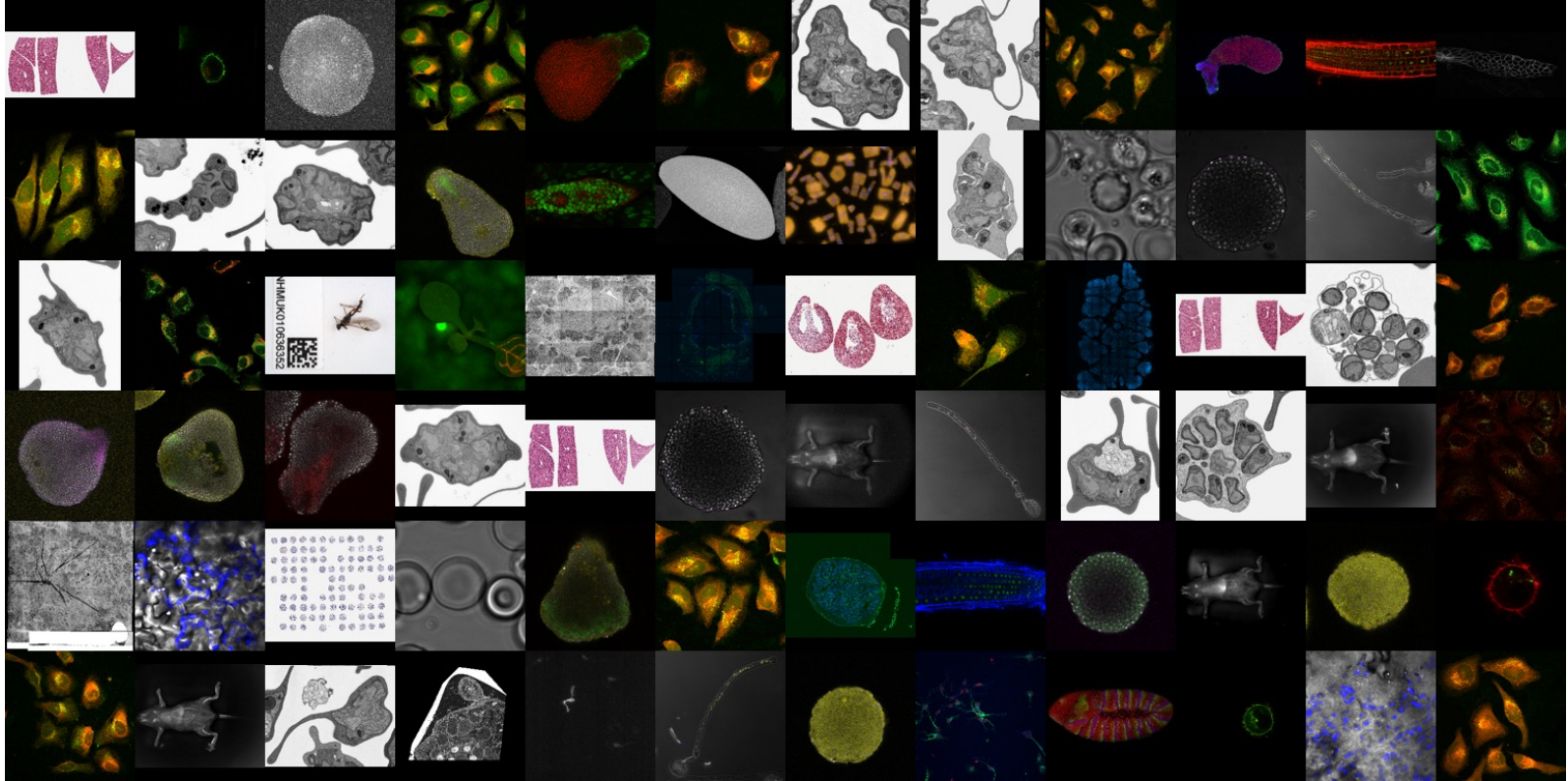
NEWS | 16 February 2022 | Correction [16 February 2022](#)

NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

...we can expect that most data supporting published results will eventually need to be open

Across a very diverse domain!



The BioImage Archive @ EMBL-EBI

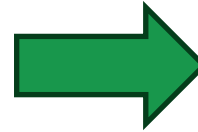
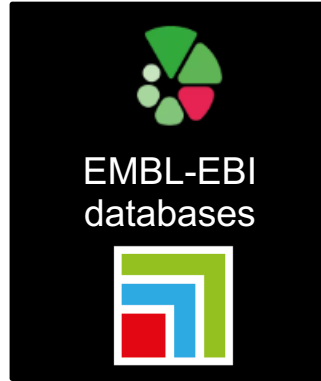
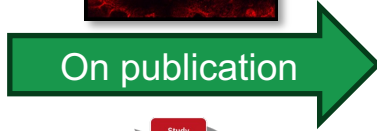
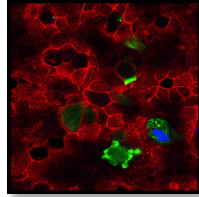
A deposition database for biological images from any modality to support:

- **Reproducibility** of results
- **Reference** imaging datasets
- **Reuse** of images for new discovery



BioImage Archive

A deposition database



Added-value data resources, e.g:

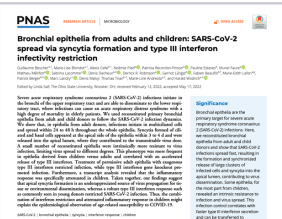
- Knowledgebases
- Data portals
- Data reprocessors



Methods development

New discovery

Training



The BioImage Archive May 2024

628 accessions

42 imaging modalities

20 submissions per month

15% AI related submissions

Expecting:

1 submission / day
later 2024

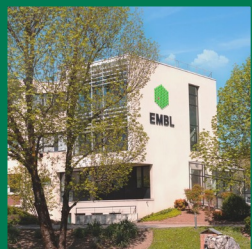
Entry 1000
mid 2025

The European Molecular Biology Laboratory



EMBL-EBI

Bioinformatics



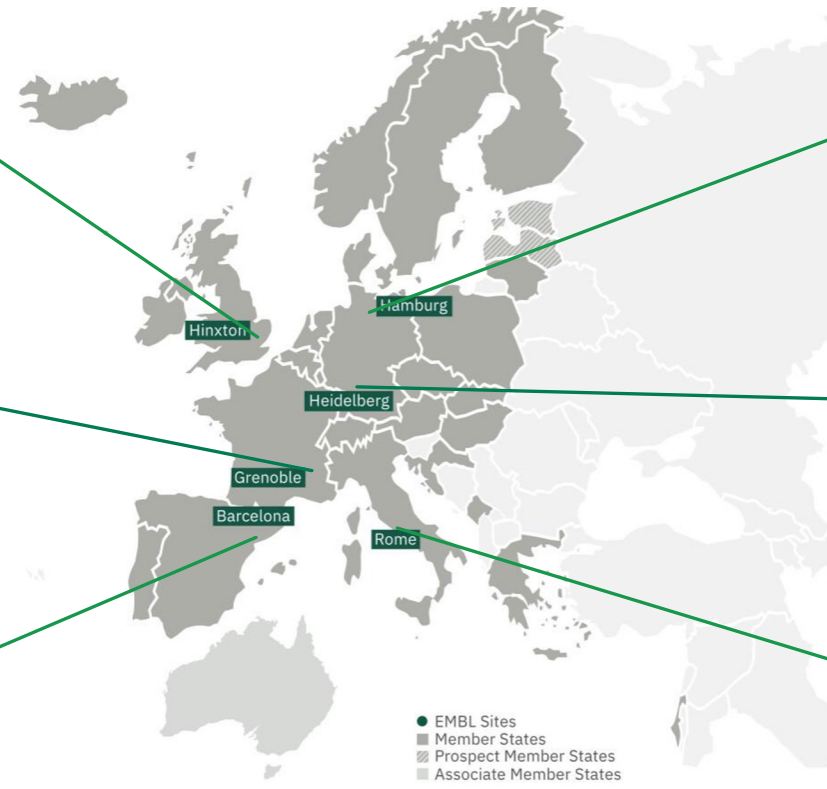
Grenoble

Structural biology



Barcelona

Tissue biology
and disease
modelling



Hamburg

Structural biology



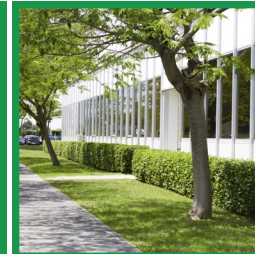
Heidelberg

Life sciences



Rome

Epigenetics
and neurobiology



Data resources at EMBL-EBI



Chemicals, molecules and drug discovery

ChEBI
ChEMBL
MetaboLights
Open Targets
SureChEMBL



Genes, genomes and RNA

Ensembl
European Nucleotide Archive
Expression Atlas
HGNC
MGnify
Rfam
RNACentral
VEuPathDB
WormBase



Proteins

AlphaFold DB
Enzyme Portal
InterPro
PDBe
PDBe-KB
Pfam
PRIDE
UniProt



Imaging and cellular structure

BioImage Archive
Electron Microscopy Data Bank
Electron Microscopy Public Image Archive



Genetic variation and disease data

COVID-19 Data Platform
DECIPHER
European Genome-phenome Archive
European Variation Archive
Mouse informatics



Literature and knowledge management

BioModels
BioSamples
BioStudies
Complex Portal
Europe PMC
GWAS Catalog
IntAct
OmicsDI
Ontologies
Reactome



EMBL-EBI in 2022

794

members
of staff

471,000

unique IP
addresses
accessed our
online training

390

petabytes
of raw data
storage

193

journal
papers and
preprints
published

197

active grants

158

collaborative
grants with
researchers in
53 countries
from 753
institutes

3700

people
participated in
our 27 public
engagement
events

978

participants in
Industry Programme
knowledge-exchange
workshops

107

million requests
to our data
resource
websites on an
average day

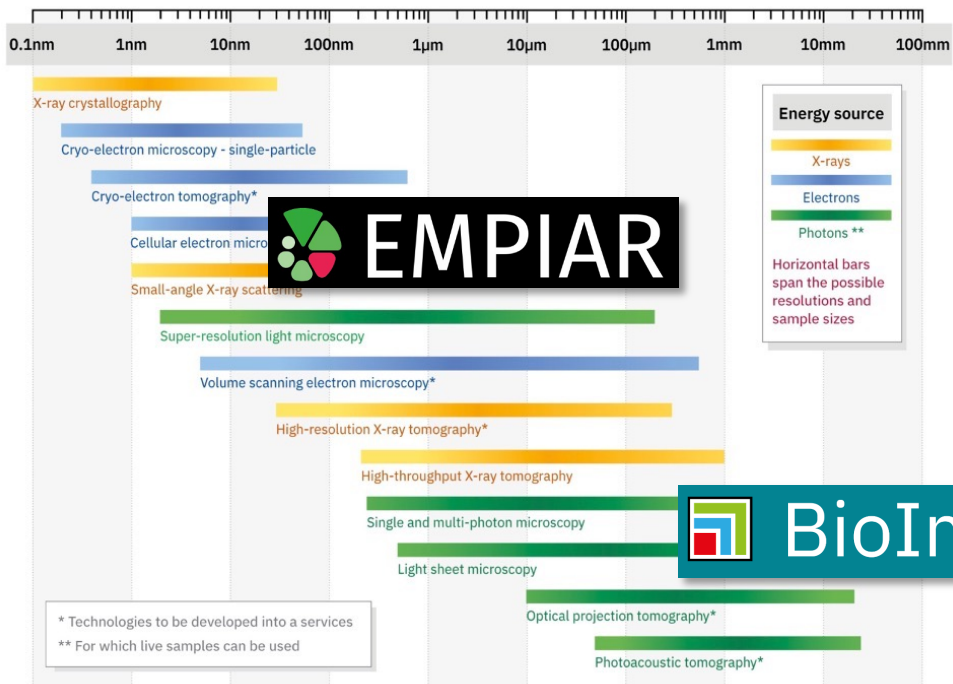
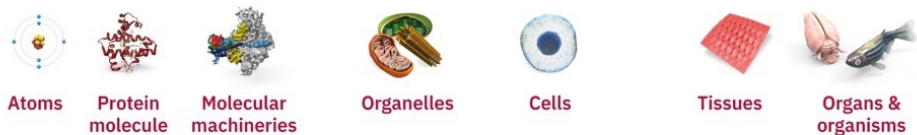
from

41

million unique
IP addresses



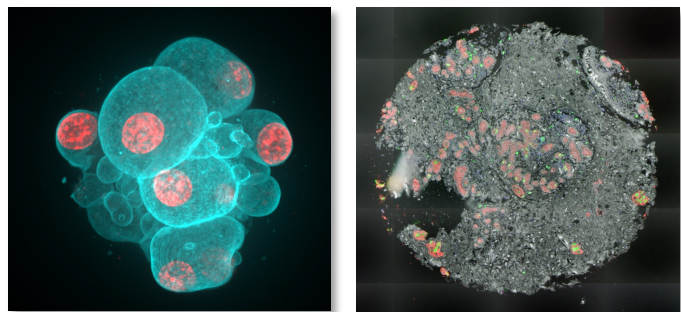
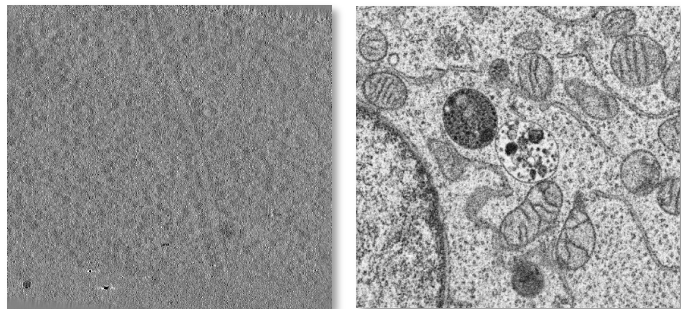
Images and resources across and between scales



Comprehensively annotated reference data

Deposition databases for images and annotations...

..across all the scales of life



Images



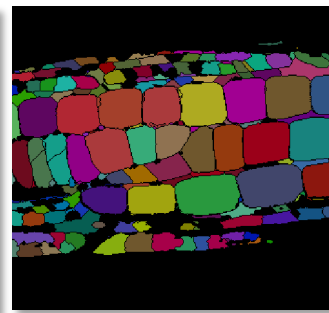
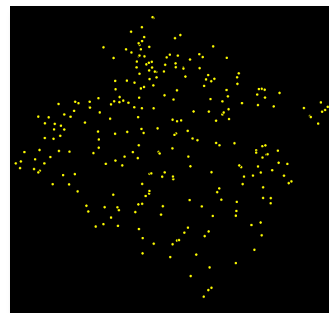
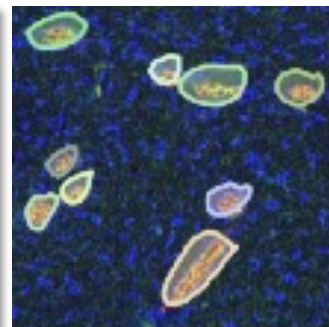
Imaging
scientists



Research
biologists

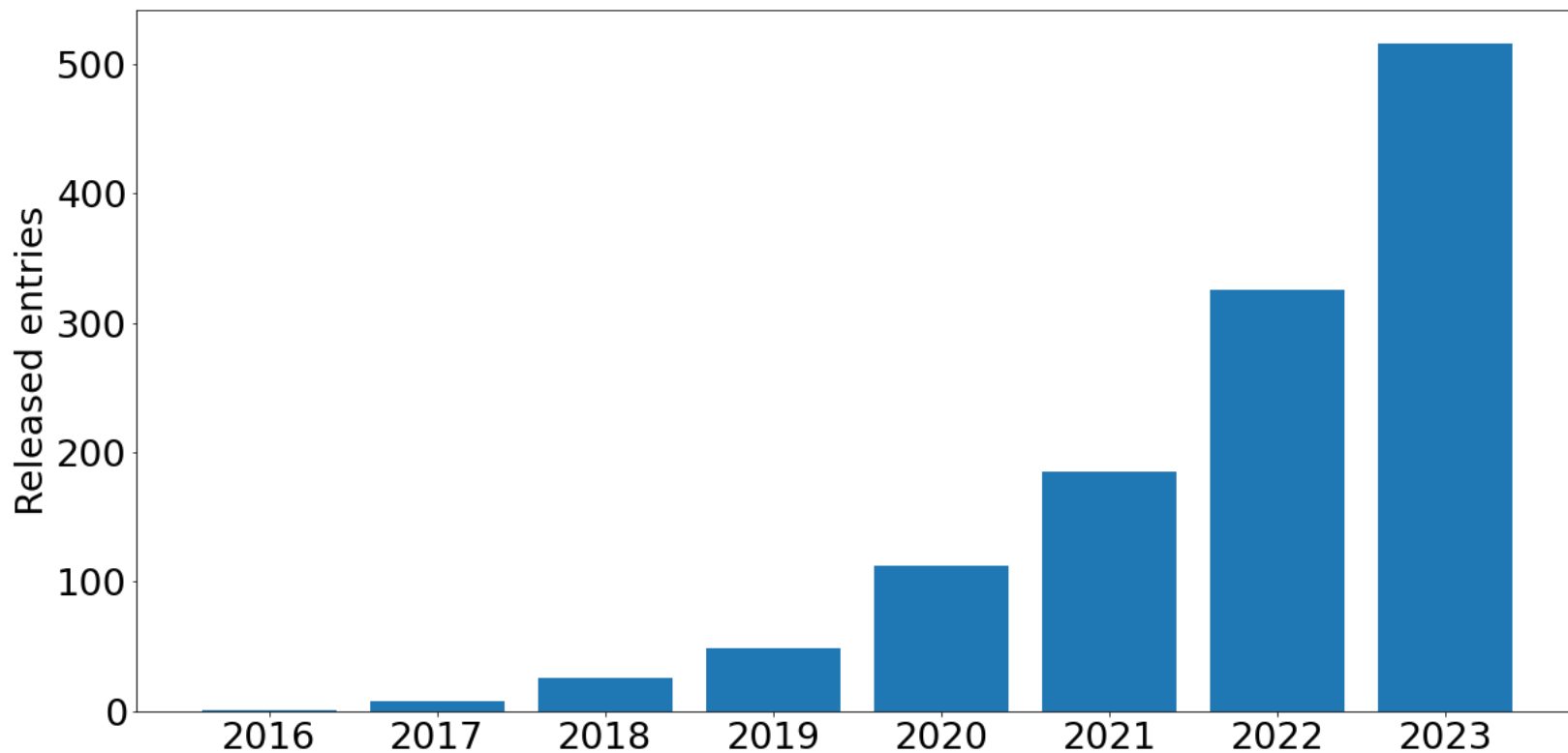


Computer
vision
researchers

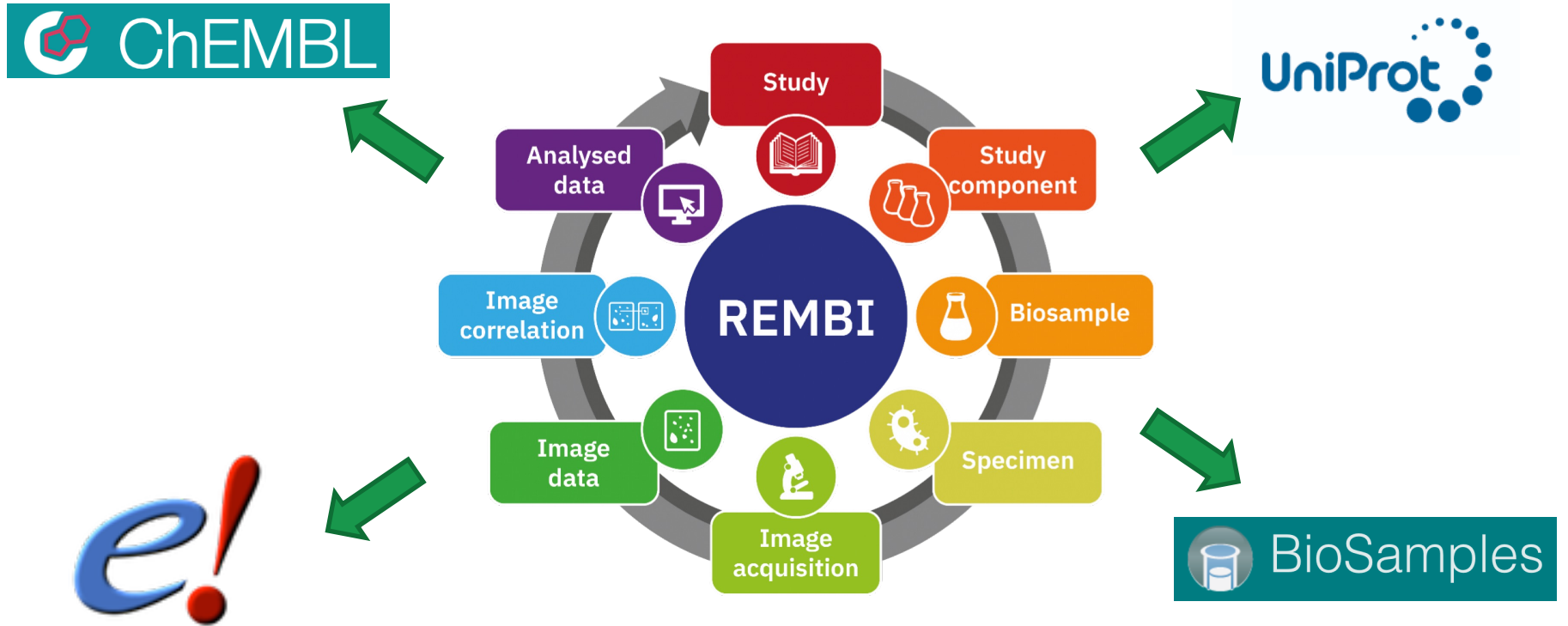


Annotations
(segmentations, particle stacks etc...)

Growth in available datasets



REMBI: Recommended Metadata for Biological Images



Anatomy of a submission



BIOSTUDIES / BIOIMAGES / S-BIAD677

Release Date: 2 May 2023 • Modified: 2 May 2023

[Cite] [JSON] [PageTab] [FTP] [Globe]

Bronchial epithelia from adults and children: SARS-CoV-2 spread via syncytia formation and type III interferon infectivity restriction

Nicolas Landrein ¹, Harald Wodrich ¹, Guillaume Beucher ¹, Thomas Trian ¹, Marie-Line Andreola ¹, Isabel Kemmer ²

¹ University of Bordeaux ² Euro-BioImaging

Accession S-BIAD677

Description Here, we record SARS-CoV-2 infections spread fast, resulting in syncytia into the apical lumen of bronchial epithelia. We revealed an interferon signaling defect in bronchial epithelia also epithelial respiratory syncytia formation. License Keyword Acknowledgement the FranceBioBioImaging EF Funding state receives funding Publication Esteves, Murie Fabien Beaufrere Harald Wodrich type III interferon

Data files

Show 5 entries

<input type="checkbox"/>	Name	Size	Section	data type	donor age	donor ID	infection	time [days post infection]	microscope type	magnification	X,y resolution (nm)	view	cf
<input type="checkbox"/>	A3 SARS Day 1 N 3G9 488 phall 560 5AC 647 dapi X63.lif	717.4 MB	Study Component	raw data	Adult	A3	yes	1	confocal	63X	901,876	Hyperstack	4.
<input type="checkbox"/>	composite A3 SARS Day 1 N 3G9 488 phall 560 5AC 647 dapi X63.tif	896.7 MB	Study Component	processed image	Adult	A3	yes	1	confocal	63X	901,876	Z-section	4.

then fixed with 4% paraformaldehyde for 30min using **Specimen** epithelia were then washed in PBS and permeabilized with 0.5% TritonX-100 in PBS for 1h and washed again before being blocked in IF buffer (PBS containing 10% SVF and 0.05% saponin) for 1h at room temperature. Primary antibody was diluted in IF buffer and applied to inserts for 1h at room temperature. Samples were washed three times under agitation with PBS and incubated with secondary antibody, fluorescently labeled phalloidin to stain the actin cytoskeleton and 2µg/mL of DAPI (4',6-diamidino-2-phenylindole), diluted in IF buffer and incubated for 2h at room temperature. Inserts were then washed extensively in PBS, desalted in H2O milliQ and rinsed in 100% Ethanol and air-dried. Membranes were then removed from inserts and mounted in DAKO Fluorescence Mounting Medium prior to microscopy analysis.

[microtubule dynamics at adherens junctions](#)

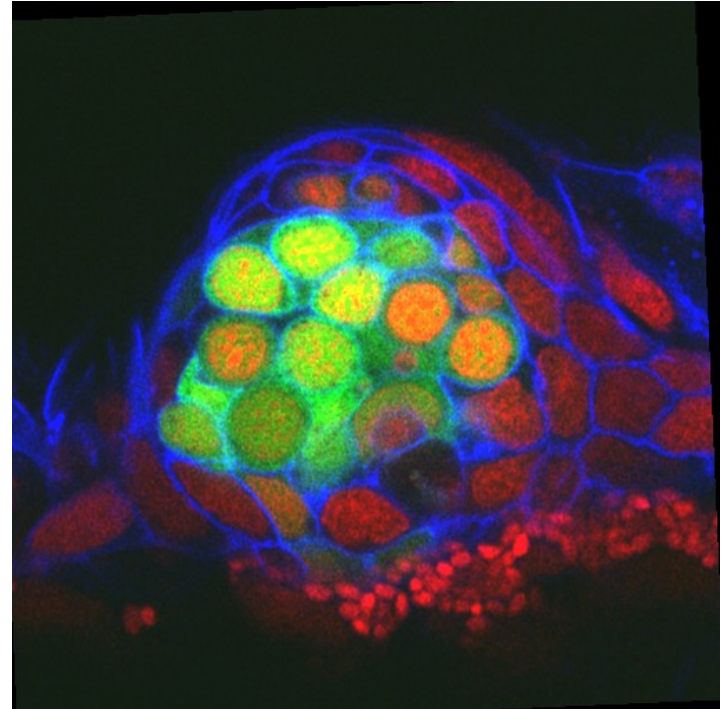
[S-JCBD-201306019]

- [E-cadherin mediated Apical Membrane Initiation Site localisation: mESCs cultured in Matrigel. IF and live imaging dataset](#) [S-BIAD473]



Recent & active development

- AI – data standards, deposition of annotations
- In-browser visualisation
- Spatial transcriptomics, and other multimodal approaches



S-BIAD886: Cell shape analysis of zebrafish lateral line neuromasts

Standards for AI data

Reduce number of formats

NGFF/OME-Zarr

COCO

GeoJSON

EMDB-SFF

CSV/TSV



Standardise metadata

Annotators

Annotation type

Annotation method

Annotation coverage

Annotation criteria

Association with source image

Confidence level

Transformations

Spatial information



API

Data browsing

Link to community tools

Metadata search

Dataset curation

Reach out to journals



Preserve annotator credit

Organise annotation events

Acknowledge exceptional contributors

Make data accessible

Encourage data production

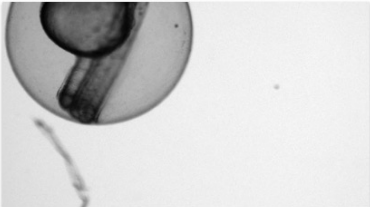
aybuke/work/ebi/code/bia-integrator/static-sitigen/tmp/pages/ai.html

BioImage Archive

BioImage Archive AI datasets

A selection of AI related studies

This is a collection of AI/Machine Learning datasets from the BioImage Archive from which one or more images have been converted to OME-Zarr. It is intended to present the AI related datasets with relevant tags and visualisation of images from the archive's collection, and to provide easy access to AI datasets and encourage tool development.

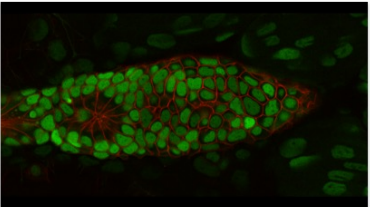


S-BIAD531 →

bright-field microscopy

Danio rerio (zebrafish)

Tags: pixel classification, segmentation, training data, test data, Time series, object classification,

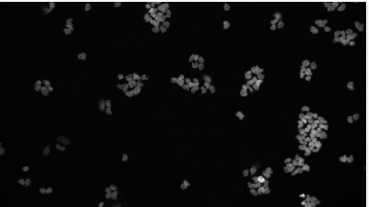


S-BIAD599 →

array-scan confocal microscopy

Danio rerio (zebrafish)

Tags: 3D, instance segmentation,



S-BSST265 →

confocal fluorescence microscopy

Homo sapiens

Tags:

ALPHA

S-BIAD634

An annotated fluorescence image dataset for training nuclear segmentation methods

Released: 2023-03-07

By: Sabine Taschner-Mandl, Inge M. Ambros, Peter F. Ambros, Klaus Beiske, Allan Hanbury, Wolfgang Doerr, Tamara Weiss, Maria Berneder, Magdalena Ambros, Eva Bozsaky, Florian Kromp, Teresa Zulueta-Coarasa

On this page

[Study Information](#) ▾

[Images](#) ▾

[Annotations](#) ▾

[Models used](#) ▾

In a nutshell

388 images

388 annotations

Study size: 472.2MiB

Filetype breakdown:

- .jpg: 79 (16.6MiB)
- .png: 115 (1.5MiB)
- .svg: 79 (4.1MiB)
- .tif: 194 (450.0MiB)
- .txt: 1 (3.1KiB)

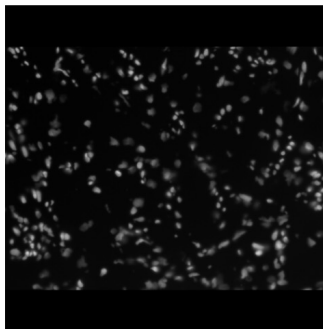
License : CC0

This dataset has

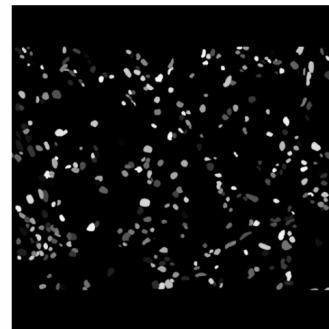
segmentation masks

test data

training data

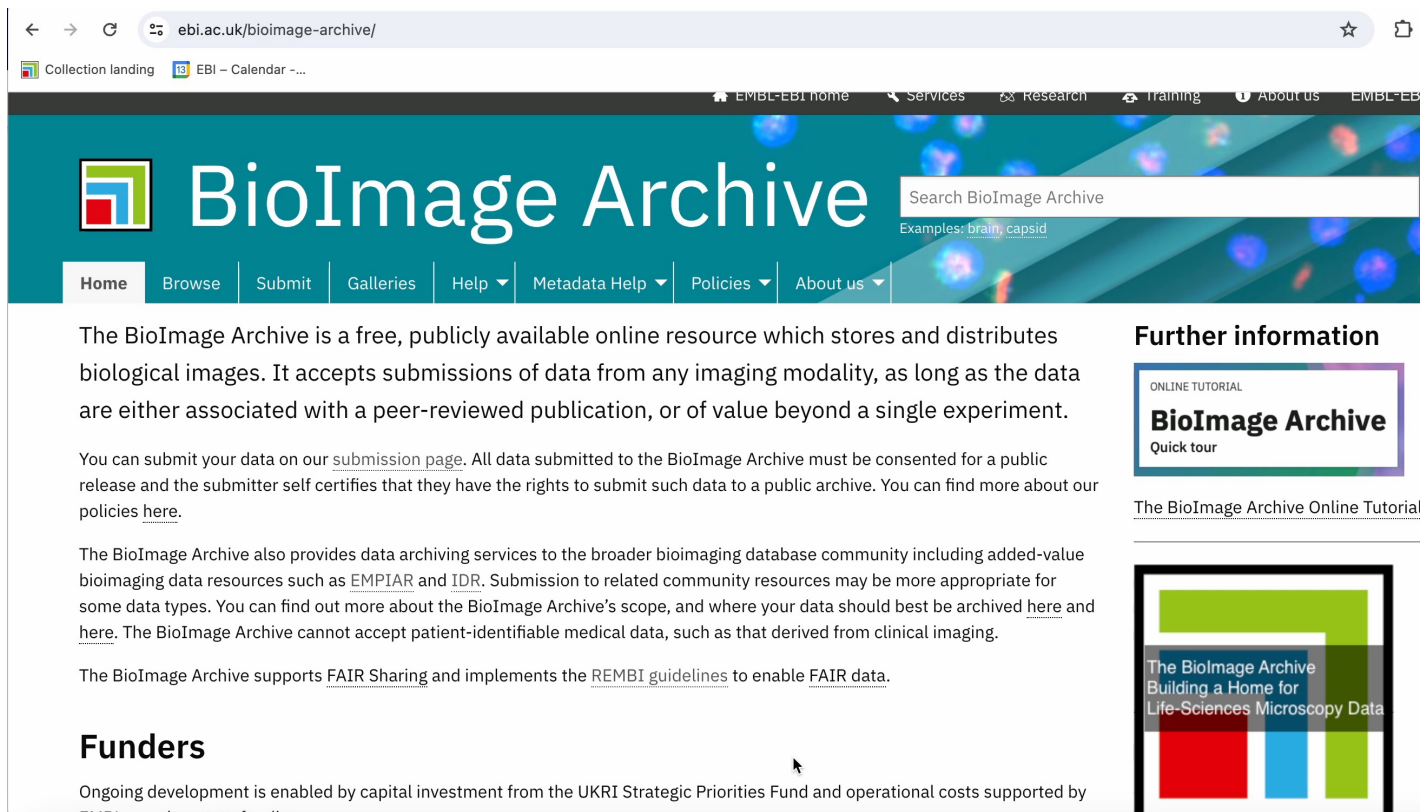


Example image for this dataset



Example annotation for this dataset

In-browser visualisation




The screenshot shows the BioImage Archive website homepage. The browser address bar displays `ebi.ac.uk/bioimage-archive/`. The page features a teal header with the BioImage Archive logo and a search bar containing the text "Search BioImage Archive" and "Examples: brain, capsid". Below the header is a navigation menu with links for Home, Browse, Submit, Galleries, Help, Metadata Help, Policies, and About us. The main content area includes a paragraph describing the archive as a free, publicly available online resource for biological images. It also provides information on submission requirements, data archiving services (EMPIAR and IDR), and support for FAIR data. A "Further information" section highlights an online tutorial and quick tour. The "Fundamentals" section mentions ongoing development supported by the UKRI Strategic Priorities Fund.

ebi.ac.uk/bioimage-archive/

Collection landing EBI - Calendar - ...

EMBL-EBI home Services Research Training About us EMBL-EBI

 **BioImage Archive** Search BioImage Archive
Examples: brain, capsid

Home Browse Submit Galleries Help Metadata Help Policies About us

The BioImage Archive is a free, publicly available online resource which stores and distributes biological images. It accepts submissions of data from any imaging modality, as long as the data are either associated with a peer-reviewed publication, or of value beyond a single experiment.

You can submit your data on our [submission page](#). All data submitted to the BioImage Archive must be consented for a public release and the submitter self certifies that they have the rights to submit such data to a public archive. You can find more about our policies [here](#).

The BioImage Archive also provides data archiving services to the broader bioimaging database community including added-value bioimaging data resources such as [EMPIAR](#) and [IDR](#). Submission to related community resources may be more appropriate for some data types. You can find out more about the BioImage Archive's scope, and where your data should best be archived [here](#) and [here](#). The BioImage Archive cannot accept patient-identifiable medical data, such as that derived from clinical imaging.

The BioImage Archive supports [FAIR Sharing](#) and implements the [REMBI guidelines](#) to enable [FAIR data](#).


Funders

Ongoing development is enabled by capital investment from the UKRI Strategic Priorities Fund and operational costs supported by

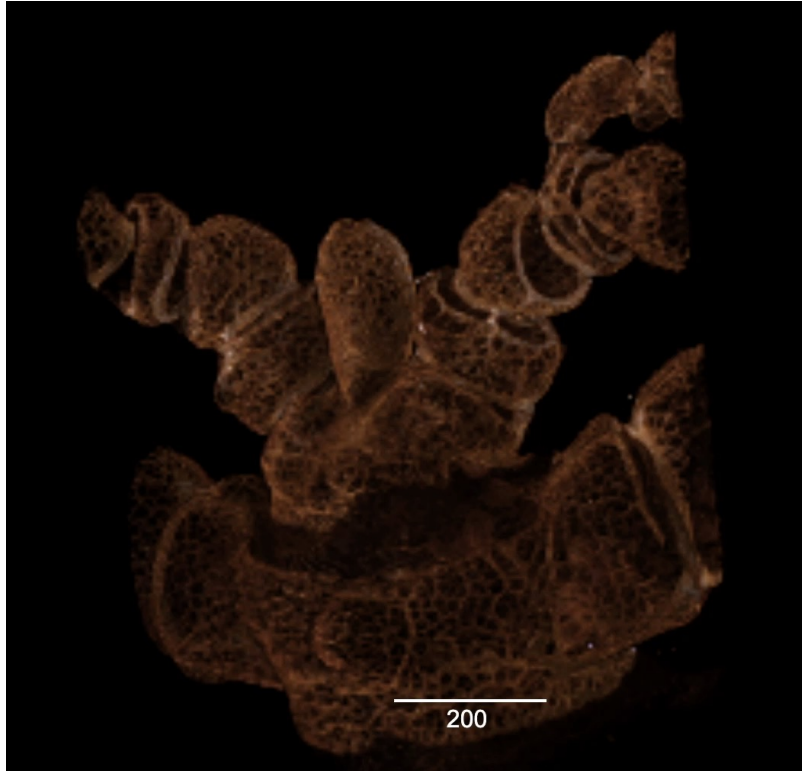
Further information

ONLINE TUTORIAL
BioImage Archive
Quick tour

[The BioImage Archive Online Tutorial](#)

 The BioImage Archive
Building a Home for
Life-Sciences Microscopy Data

Visualisation in 3D



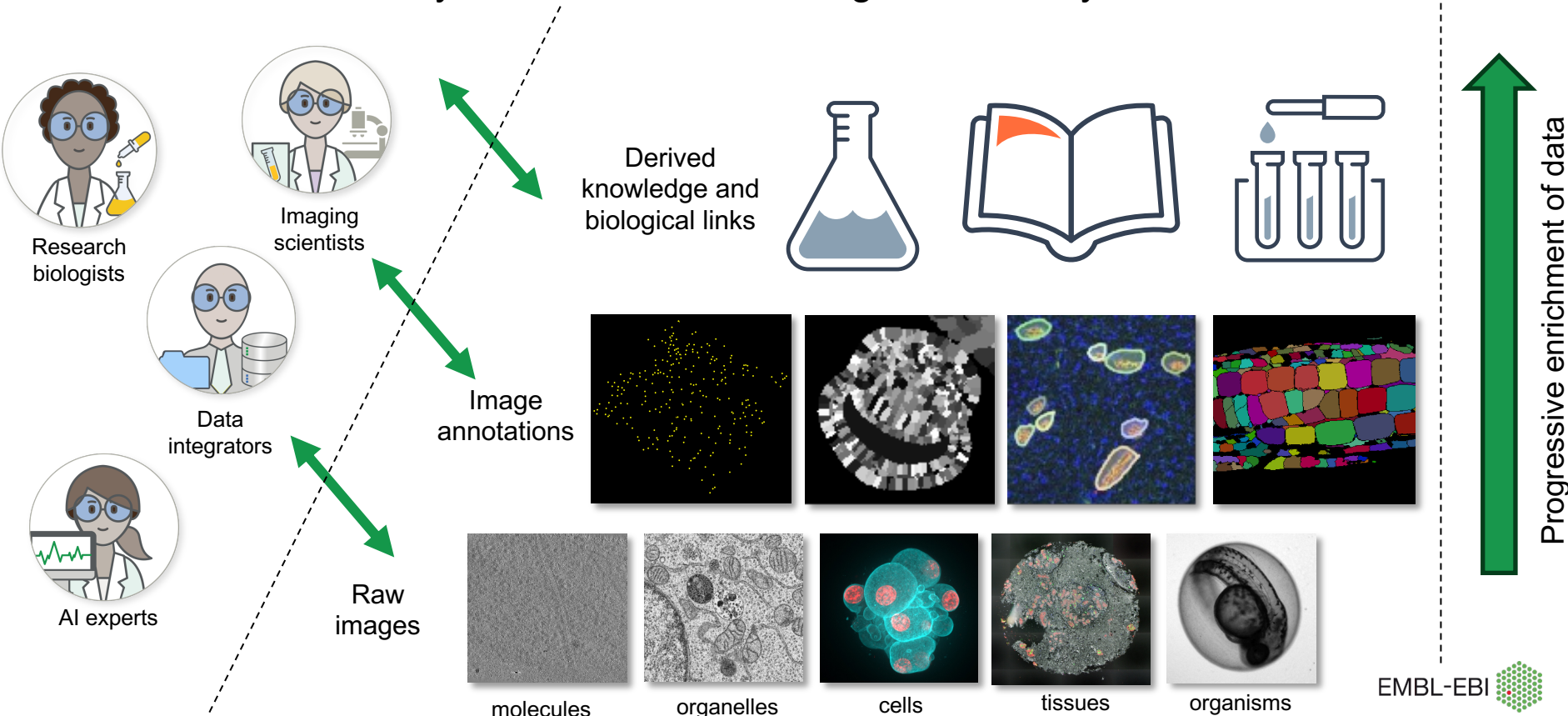
Conversion to a common format, OME-Zarr enables multiple alternative visualisation tools

S-BIAD606: The sea spider *Pycnogonum littorale* overturns the paradigm of the absence of axial regeneration in molting animals

Why: continual enhancement of scientific value

Scientific community

Image data ecosystem



Looking forward – open data & scaling

1

submission / day
2024



10

submissions / day
2027 (?)

NEWS | 16 February 2022 | Correction [16 February 2022](#)

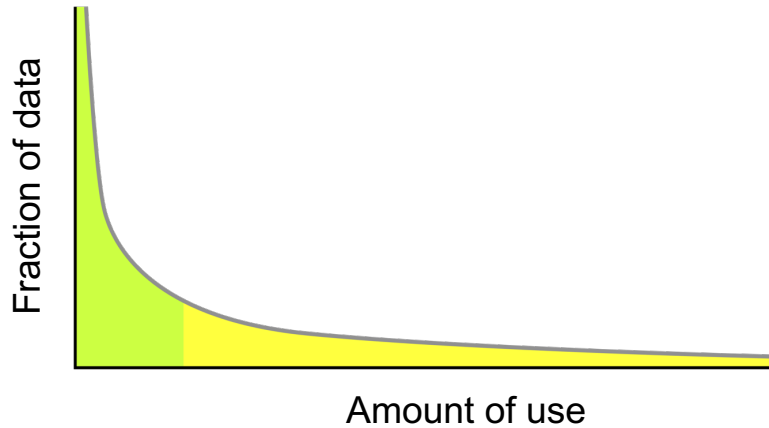
NIH issues a seismic mandate: share data publicly

The data-sharing policy could set a global standard for biomedical research, scientists say, but they have questions about logistics and equity.

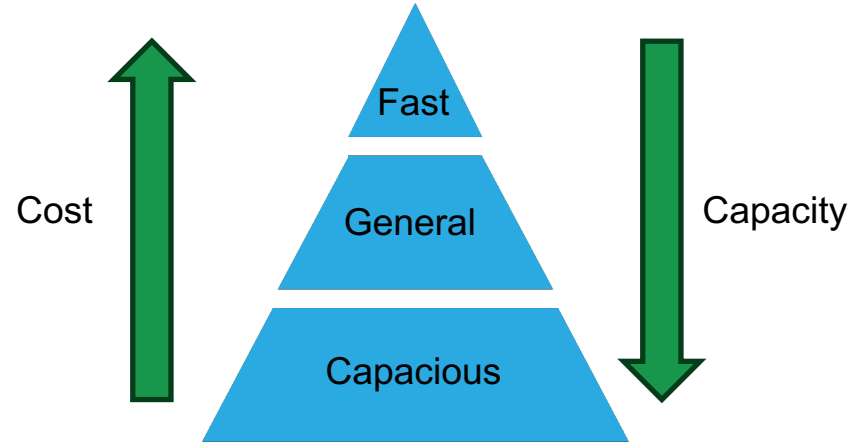


Horizon Europe
2021-2027

Future – data use & storage cost

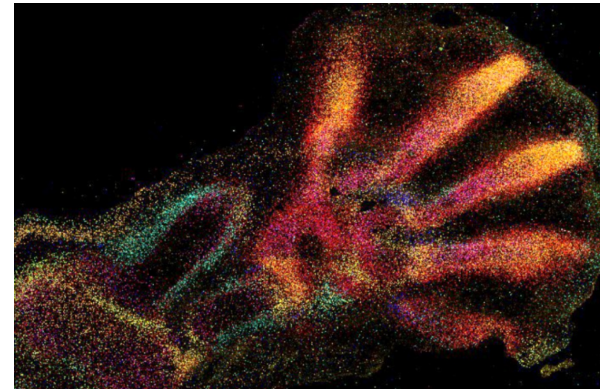
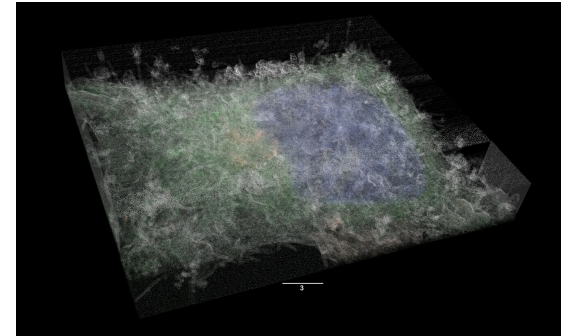
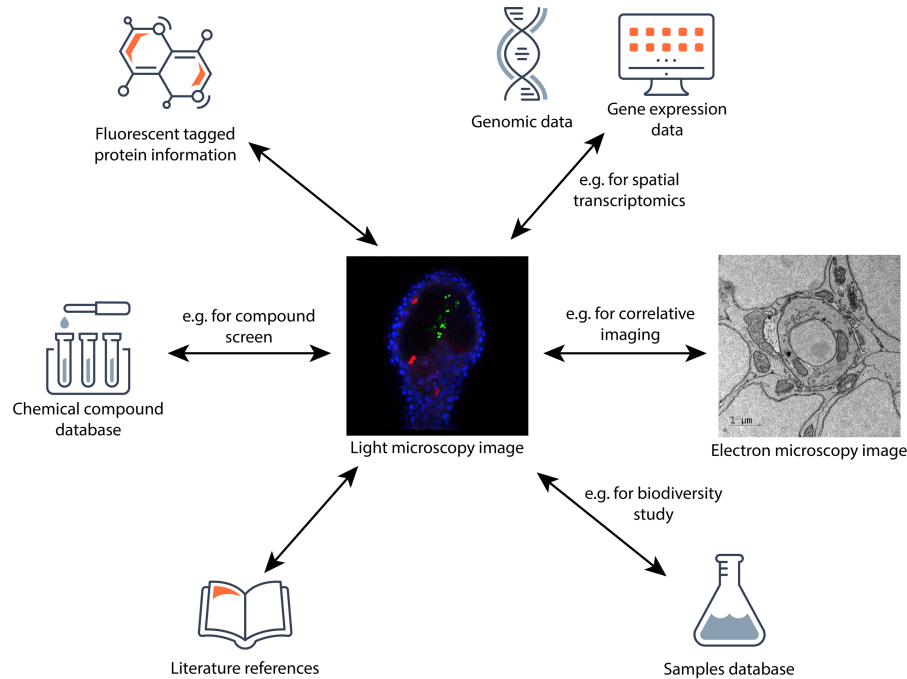


Data access is a long-tailed distribution



Storage has cost/capacity trade-offs

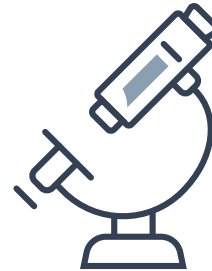
Data integration, correlative and multimodal data



Data integration –complex projects



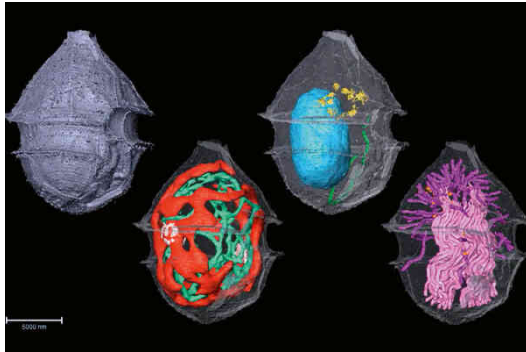
Samples



In-situ imaging



Delayed imaging



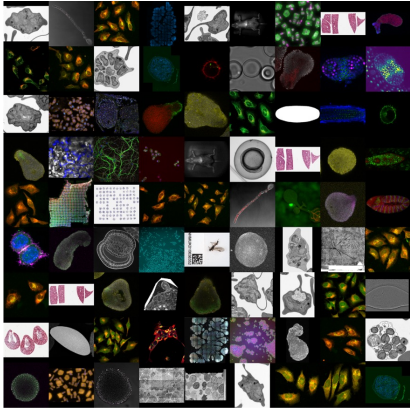
Chemical profiling



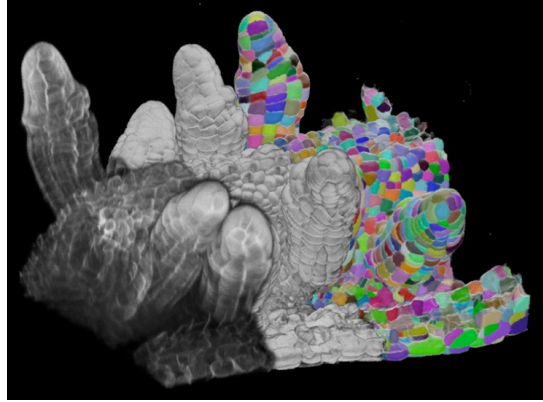
Sequencing



AI – challenges and opportunities



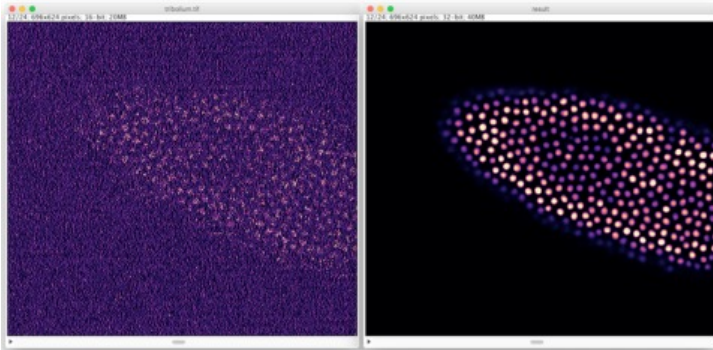
Data for model training
(towards foundation models)



AI for labelling, search,
curation



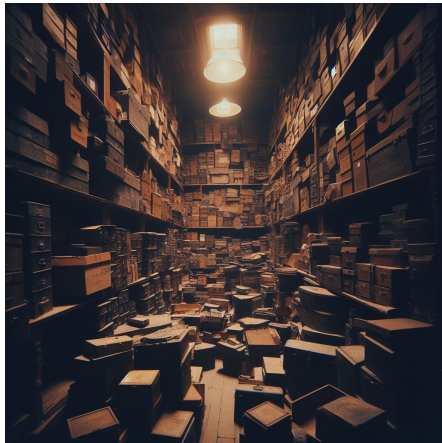
AI supported helpdesk



AI for data enhancement

Moving forward: from files to well described images

Data packed in boxes



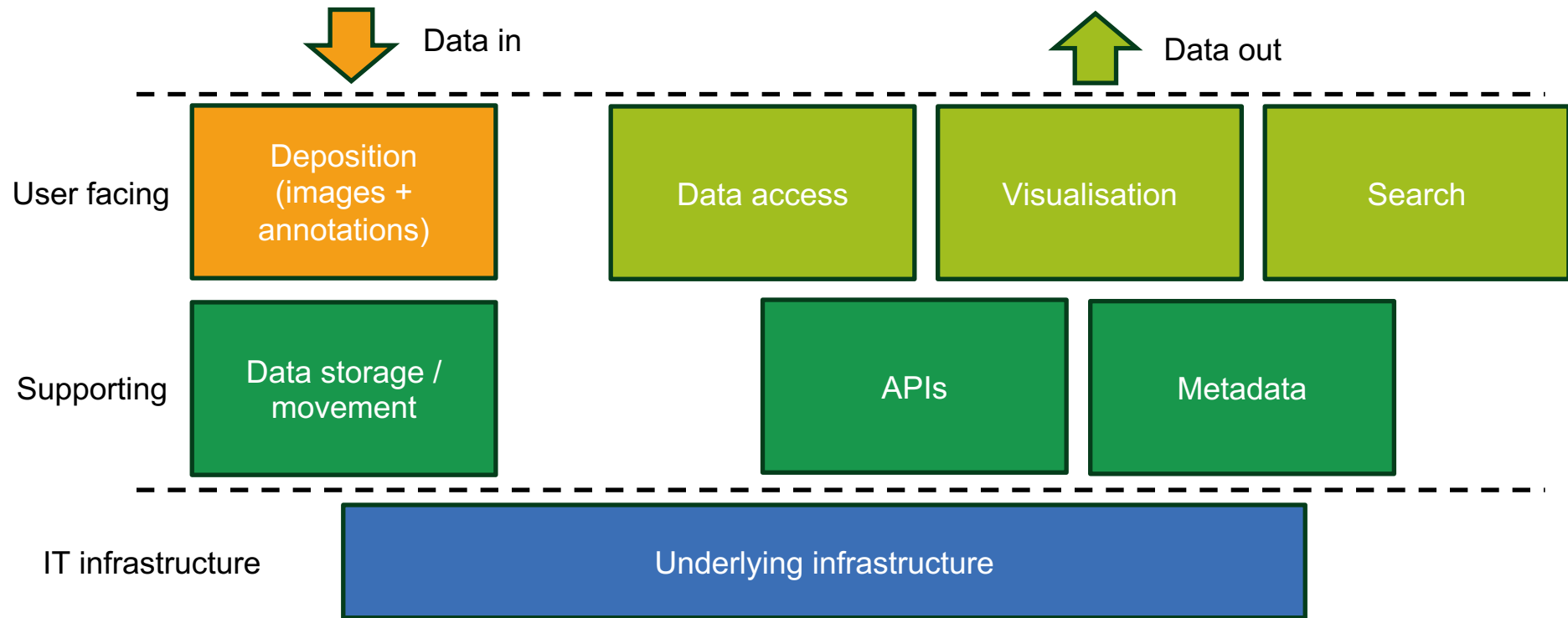
Well labelled data
(in boxes!)

Selective
presentation

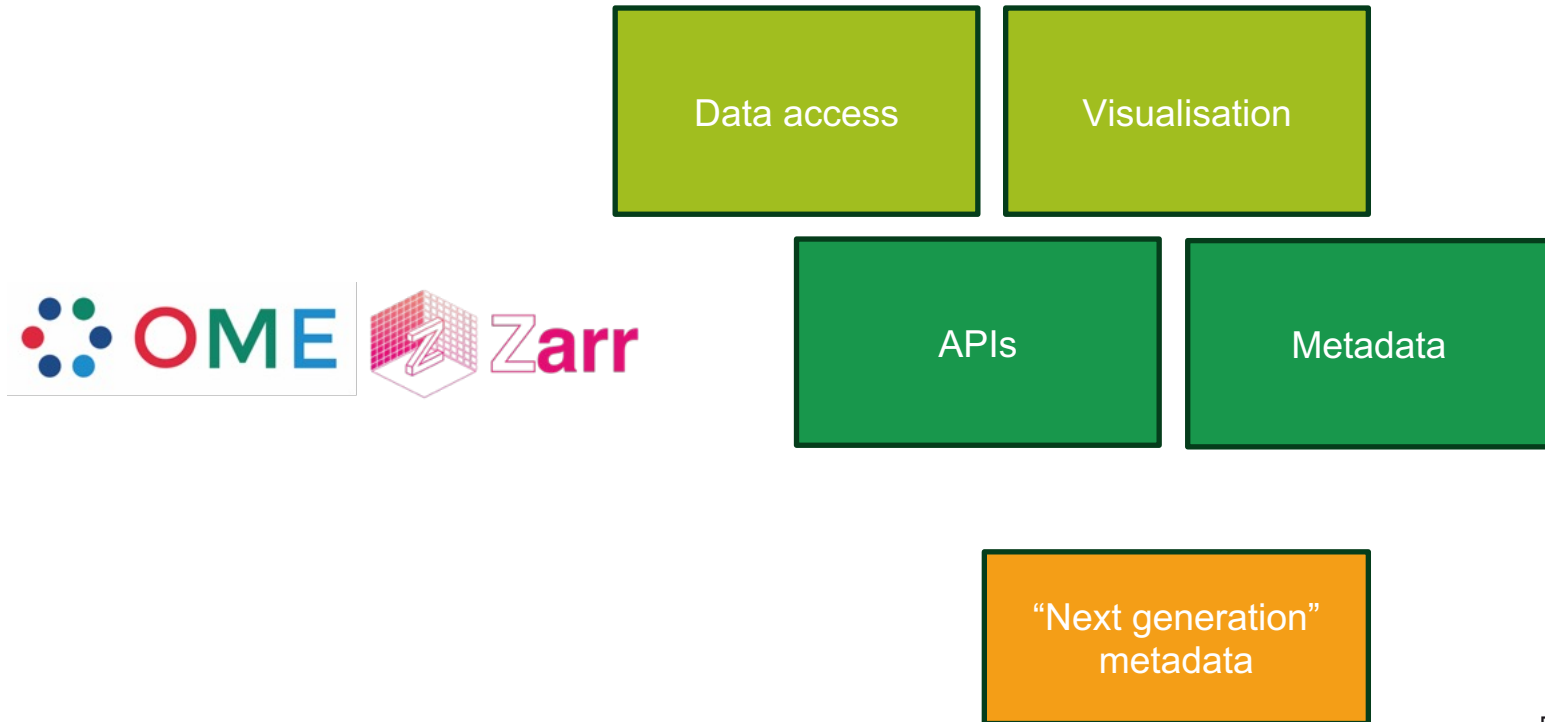


Beautiful, usable
image data

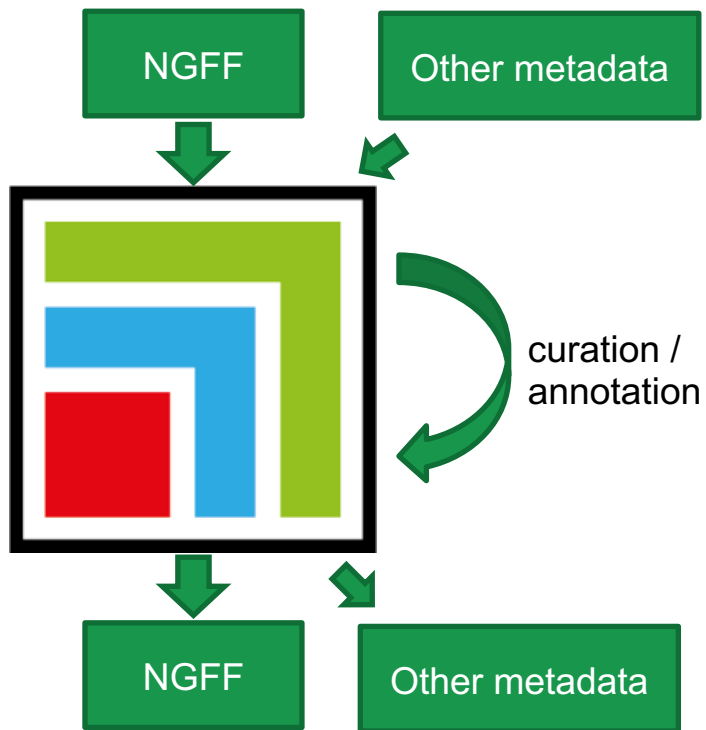
Building a repository...



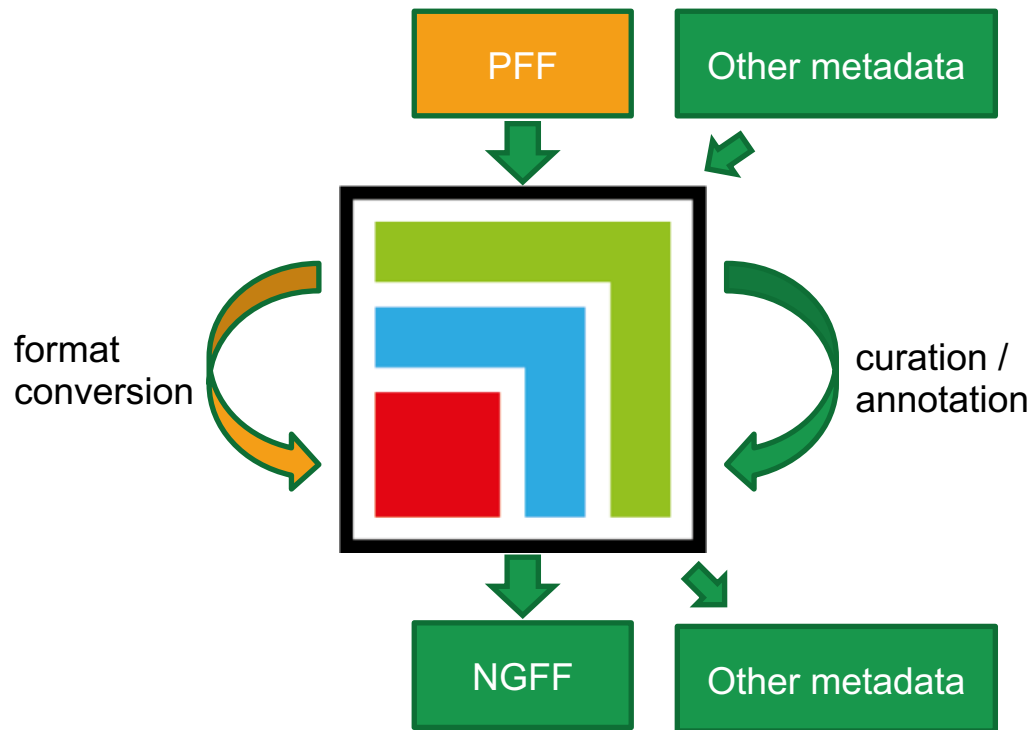
Components of image data resources



How we're using OME-NGFF



Direct deposition



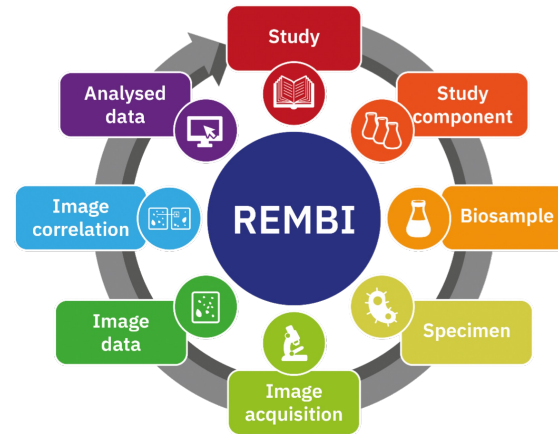
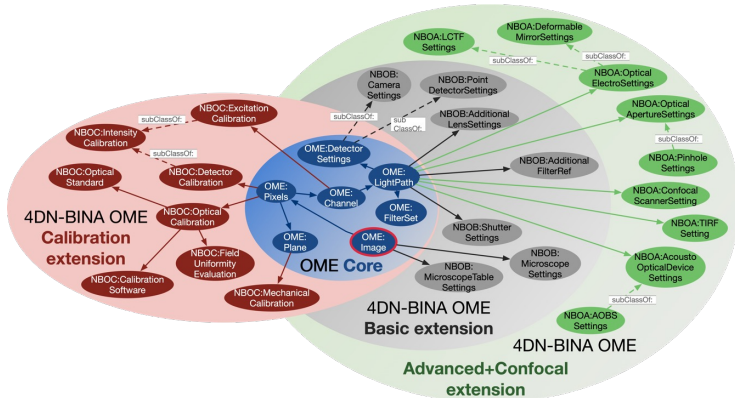
Conversion

Building an API towards...

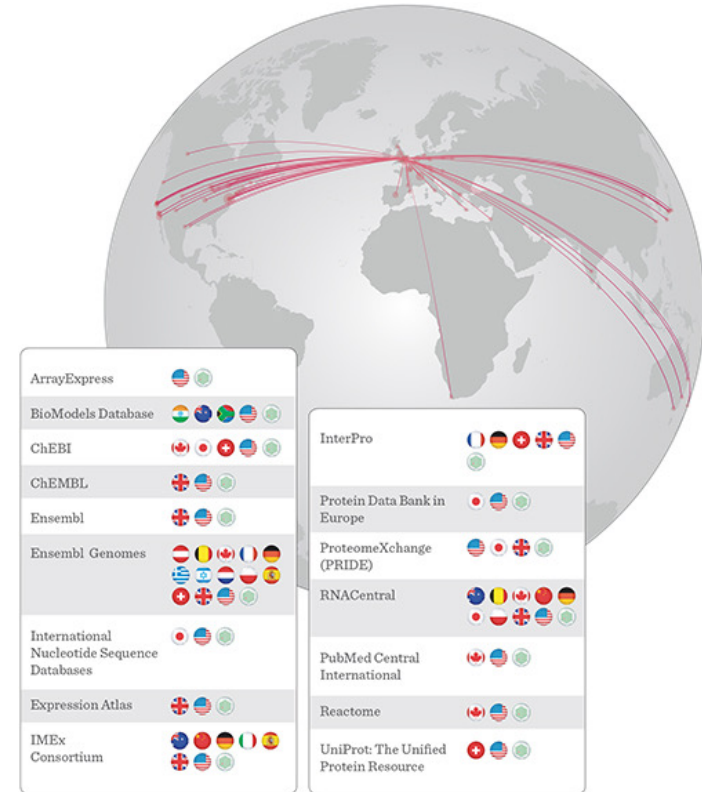
```
% curl -k -s https://45.88.81.209:8080/v1/studies/00000000-0000-0000-0006-09b5d587b158
```

```
{
  "@context":
  "https://raw.githubusercontent.com/BioImage-Archive/bia-
  integrator/main/api/src/models/jsonld/1.0/StudyContext.js
  onld",
  "uuid": "00000000-0000-0000-0006-09b5d587b158",
  "version": 3,
  "title": "Microscopy data for 'RPA shields inherited
  DNA lesions for post-mitotic DNA synthesis'",
  "description": "Single-stranded DNA during DNA
  replication and repair in S/G2 needs protection by
  replication protein A (RPA). In this study we reveal that
  RPA also shields inherited single-stranded DNA in G1,
  representing replication remnants from the previous cell
  cycle, to allow for post-mitotic DNA synthesis.",
  "organism": "Homo sapiens (human)",
  "release_date": "2021-05-31",
  "accession_id": "S-BIAD106",
  "imaging_type": "fluorescence microscopy",
  "file_references_count": 19558,
  "images_count": 19306
}
```


...extensible metadata



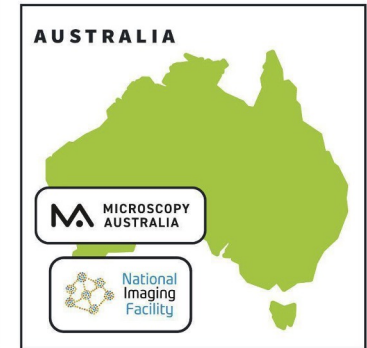
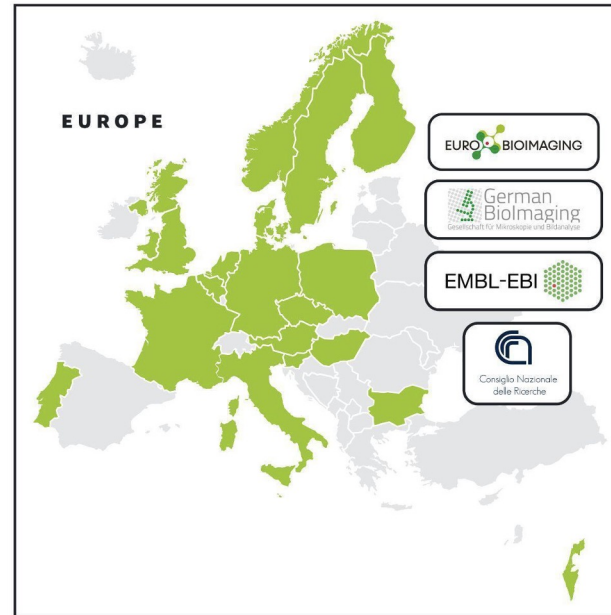
Future – global integration (or: what about 10 to 100?)



FoundingGIDE: Founding a Global Image Data Ecosystem

Laying strong foundation of an ecosystem for image data exchange based on global coordination of technical developments among data infrastructures and communities

- Global coordination among diverse imaging resources & communities
- Concerted development of Ontologies and Metadata models
- Adoption of outputs by global image data resources
- Interoperable solutions for microscopy and pre-clinical data
- Community recommendations FAIR image data management



Thanks!

UK Research
and Innovation



Liviu Anita



Kola Babalola



Teresa Zulueta



Aybuke Yoldas



François Sherwood



Craig Russell



Projects have been funded by the European Union's Horizon 2020 research and innovation programme



EMBL-EBI's data resource survey

- Support EMBL-EBI resources with 10 minutes work!
- [https://www.surveymonkey.com/r/HJKYKTT?channel=\[website\]](https://www.surveymonkey.com/r/HJKYKTT?channel=[website])

