Most of the world's bioimaging data **lacks a clear path** to being shared.

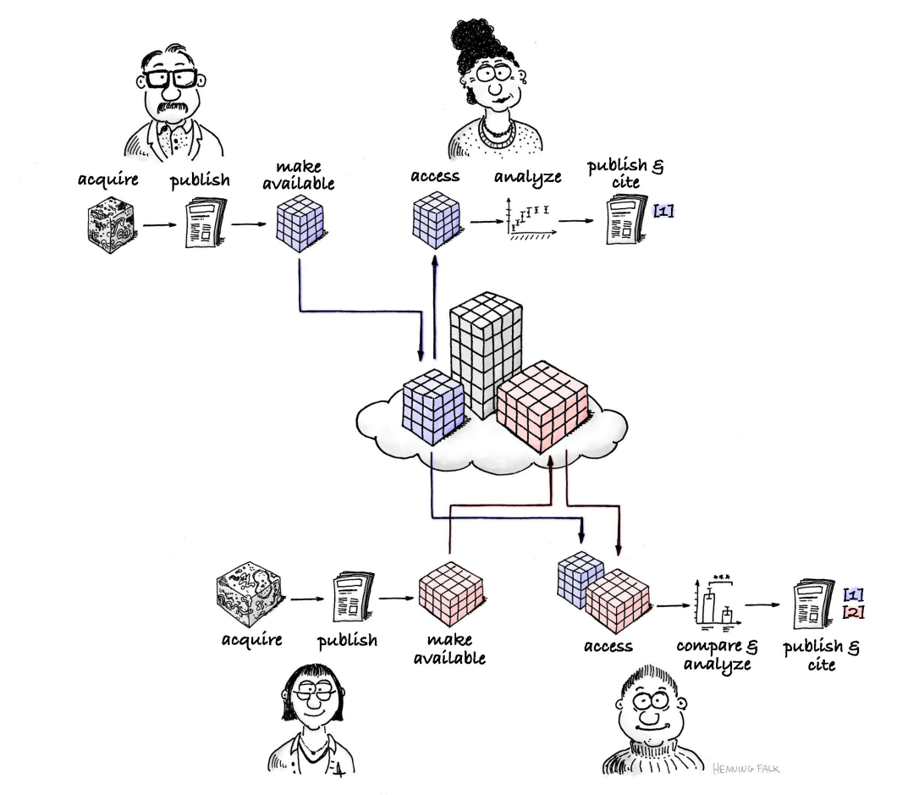
Susanne Kunis
Department of Biology/Chemistry, Center for Cellular Nanoanalytics, University Osnabrück, Germany
0000-0001-6523-7496

Josh Moore
German BioImaging e.V, Society for Microscopy and Image Analysis, Konstanz, Germany
0000-0003-4028-811X

*AI's Dirty Little Secret: **Without FAIR Data**, It's Just Fancy Math*

Restricted access or a **limited** number of biological images and their metadata inhibits researchers from developing robust and generalizable models, potentially **decreasing** the **accuracy** and **performance** of AI applications in bioimaging.

Community Repositories are available to **share** and **find valuable data.** Make use of them! ❶❷❸❻



"FAIR re-use" by @DrHenningFalk, ©2022 @NumFOCUS is used under a CC-BY 4.0 license.

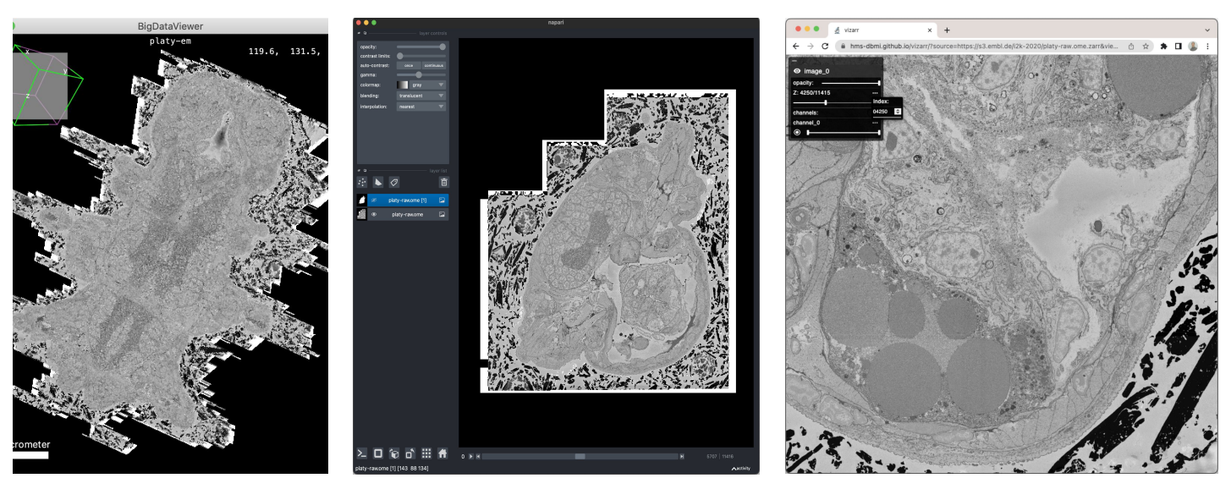
EMDB Electron Microscopy Data Bank
IDR
EMPIAR Electron Microscopy Public Image Archive
BioImage Archive

FAIR Image Objects: FDO-compatible datatype for bioimaging to ensure data is **open** and **web-accessible** to enable **efficient distributed processing.** ❶❹❺❻

Terabytes of pixels as well as analytical results can be made shareable, linkable, browsable, re-usable, archivable. A **pyramidal** structure allows Google Maps-style zooming. A cloud-optimized ("**chunked**") format allows referencing individual regions of an image in parallel.

https://zarr.dev

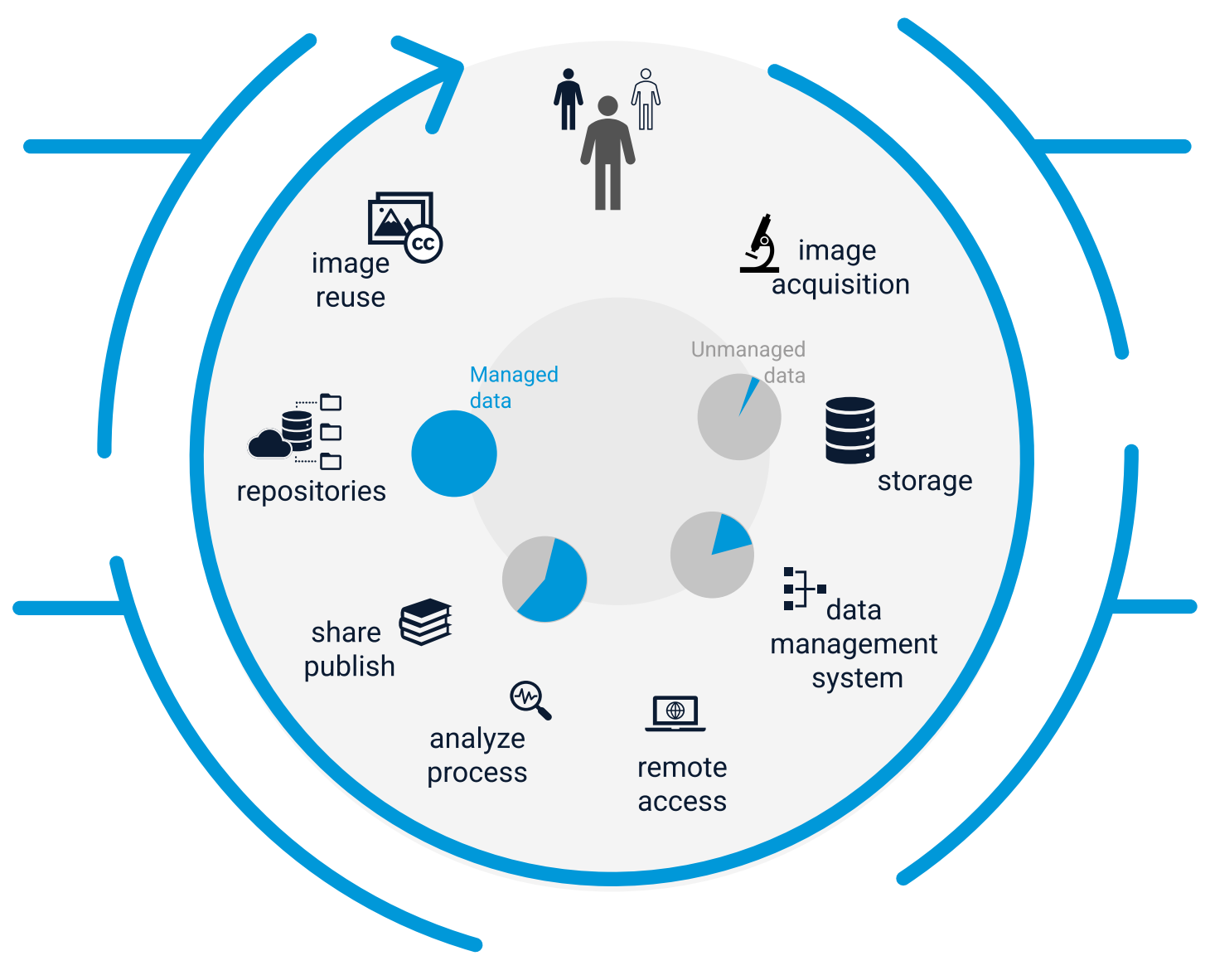
"larger-than-memory, n-dimensional, typed arrays"



8-TB elecon microscopy volume of a 6 day old *Platynereis* larva from Vergara *et al*. 2020 available at: https://s3.embl.de/i2k-2020/platy-raw.ome.zarr

See QR code for demo: https://wklink.org/6422

NFDI4BIOIMAGE is one example of a national endeavor to **organize existing data** for use in AI. ❶❺❻❼



Objective 1
Champion the **standardization** of the „bioimage data" type

Objective 2
Provide scalable **infrastructure** for FAIR image data

Objective 3
Maximize the reach of **reproducible** image analysis workflows in the community

Objective 4
Capacitate researchers for FAIR image data management

Implementing AI...

...Data Requirements for AI

Accessibility ❶
- Foster innovation
- Model improvement

Quality ❷
- AI model accuracy
- Scaling of AI applications

Variety ❸
- Improved generalization
- Enhanced adaptability
- Mitigation of biases
- Facilitation of cross-domain adoption

Context ❹
- Data interpretation
- Ability to learn
- Transparent and comprehensible decisions
- Better handling of unexpected or varying conditions

Structure ❺
- Efficient processing of large amounts of information
- Reusability of data and thus interoperability of AI applications

Formats ❻
- Scalability
- Compatibility
- Storage space optimization
- Efficient processing

Infrastructure ❼

There are **rarely** comprehensive **metadata** associated with image data and a **lack of semantic data integration** means AI can't assign meaning to bioimages.

Community metadata standards and techniques are available to **facilitate and automate** semantic data integration. ❶❷❸❹

OME

REMBI

W3C OWL
W3C RDF
W3C SPARQL

Semantic data integration enables the **integration** and **harmonization of data** from various bioimaging sources and related data. ❶❹❺❻

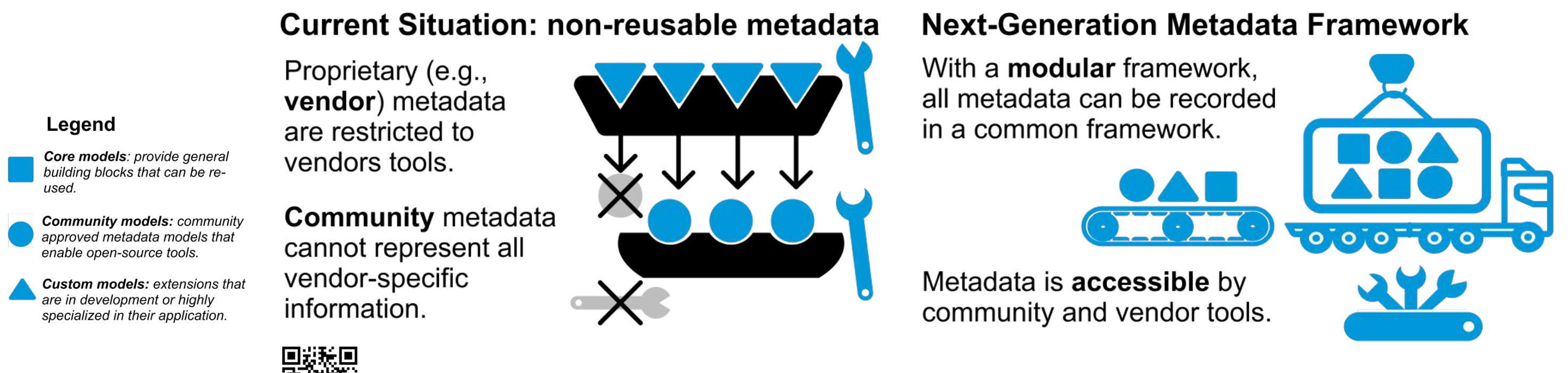
This approach focuses on understanding the **relationships** and **meaning** of the data elements to facilitate efficient data sharing, analysis, and interpretation in the field of bioimaging.
For successful semantic data integration, certain requirements for metadata are essential :
- Unique identifiers (PID)
- Relationships and Linkages (RDF)
- Semantic annotations (Ontologies)
- Metadata standards (schemas, models)

Such metadata enables **mapping** of different **schemas through ontologies** and realizes **entity mapping** based on a **knowledge graph.**

If this metadata is missing, this requires time-consuming subsequent labeling to create the context required for the AI.

*"How many humans does it take to make tech seem human? **Millions**"*

Engagement with **hardware vendors** is needed to enable **complete and re-usable metadata** directly from acquisition systems. ❶❷❹



Legend
■ ***Core models****: provide general building blocks that can be re-used.*
● ***Community models:*** *community approved metadata models that enable open-source tools.*
▲ ***Custom models:*** *extensions that are in development or highly specialized in their application.*

Current Situation: non-reusable metadata

Proprietary (e.g., **vendor**) metadata are restricted to vendors tools.

Community metadata cannot represent all vendor-specific information.

Next-Generation Metadata Framework

With a **modular** framework, all metadata can be recorded in a common framework.

Metadata is **accessible** by community and vendor tools.

Figure is co-authored with Caterina Strambio De Castillia, CC-BY

NFDI4 BIOIMAGE

DFG Deutsche Forschungsgemeinschaft
In cooperation with nfdi