Introduction to the FAIR principles and the bioimage data type

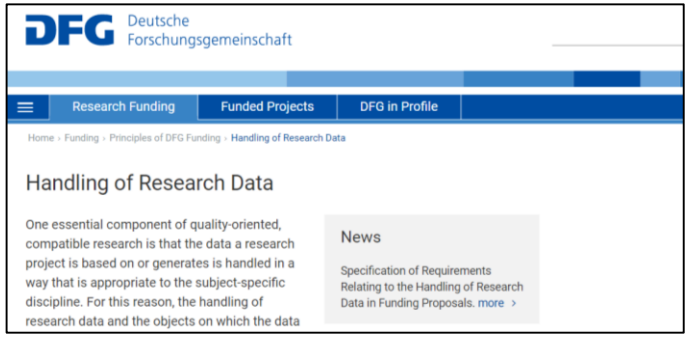
Workshop: **FAIR data handling for microscopy: Structured metadata annotation in OMERO**

April 29th & 30th, 2024, Day 1 – Session 2

Trainers: Tom Boissonnet, Vanessa Fuchs, **Christian Schmidt**

Significance of research data management

- **Intrinsic motivation**
- Impact and efficiency of my research
- **Extrinsic factors**
- Funding agency demands
- Good Scientific Practice
- **Ethical aspect**
- Make science more sustainable
- Promote *Open Science* and data sharing



- https://www.dfg.de/en/research_funding/principles_dfg_funding/research_data/index.html
- https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf
- https://twitter.com/BMBF_Bund/status/1571801906074337280?s=20&t=krDcwOPMuPwjs-VisYBgVg
- https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

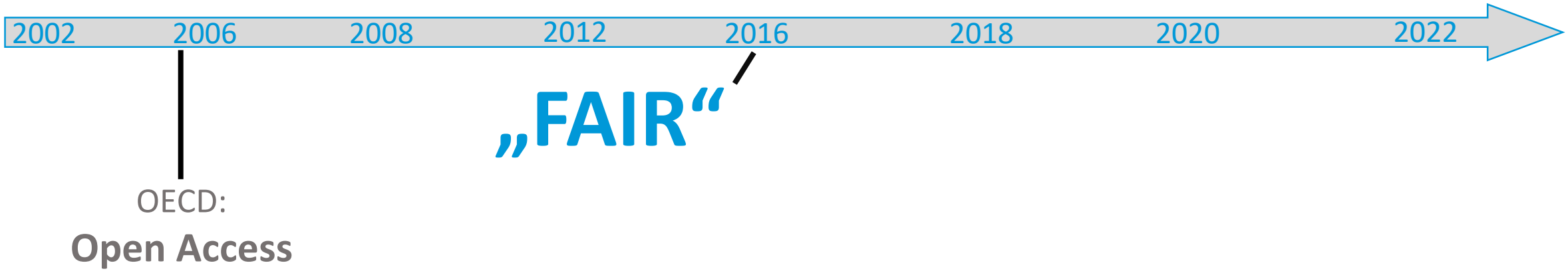
Research Data Ecosystems Emerging in Europe

How to promote sharing and reuse?

- *Guiding principles for data (FAIR)*

Germany

Sharing and reproducibility
https://www.youtube.com/watch?v=j-6N3bLgYyQ&t=10s



Europe (& beyond)

The FAIR principles for data and data stewardship

Findable **Accessible** **Interoperable** **Reusable**

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be 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

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

- Wilkinson et al., 2016, Scientific Data, DOI: 10.1038/sdata.2016.18, CC-BY 4.0 (http://creativecommons.org/licenses/by/4.0)
- Jacobsen et al., 2020, FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, DOI: 10.1162/dint_r_00024

***Guiding* principles**
(*not* a dictation of technical solutions)

Goal is machine-readability
(*not* only human ability to reuse)

Guidelines for data producers *and* publishers

Applies to data and their provenance

FAIR is *not*
- the same as „open“
- a standard

The FAIR principles for data and data stewardship

Findable Accessible Interoperable Reusable

Mark D. Wilkinson, Talk: FAIR in a Series of Vignettes, CoRDI 2023, Karlsruhe, ***https://tinyurl.com/FAIR-CoRDI23***, CC-BY:

The word “data” causes enormous amounts of heartache and confusion, especially when we then add “metadata” to the mix!

“FAIR Digital Objects” or “FAIR Research Objects” seem to capture what we meant with less confusion

R1.3: (meta)data meet domain-relevant community standards

- the same as „open“
- a standard

- Wilkinson et al., 2016, Scientific Data, DOI: 10.1038/sdata.2016.18, CC-BY 4.0 (http://creativecommons.org/licenses/by/4.0)
- Jacobsen et al., 2020, FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, DOI: 10.1162/dint_r_00024

The FAIR principles for data and data stewardship

Technical infrastructure (generic operations)
Data (domain-specific content)

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be 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

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

FAIR requires **collaboration**:

- infrastructure providers
  (IT, libraries, publishers, etc.)
- stakeholder community members
  (researchers, core facility specialists, research support staff)

- Schultes E, 2020, GO FAIR conversion matrix – slides, https://osf.io/ceyvj/, (CC-BY 4.0: http://creativecommons.org/licenses/by/4.0)

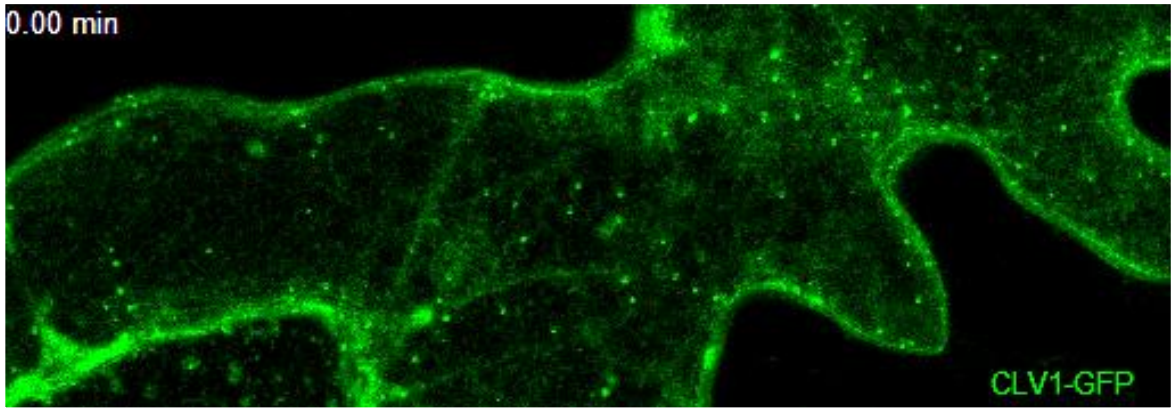
The bioimage data type

Microscopy data is (often):

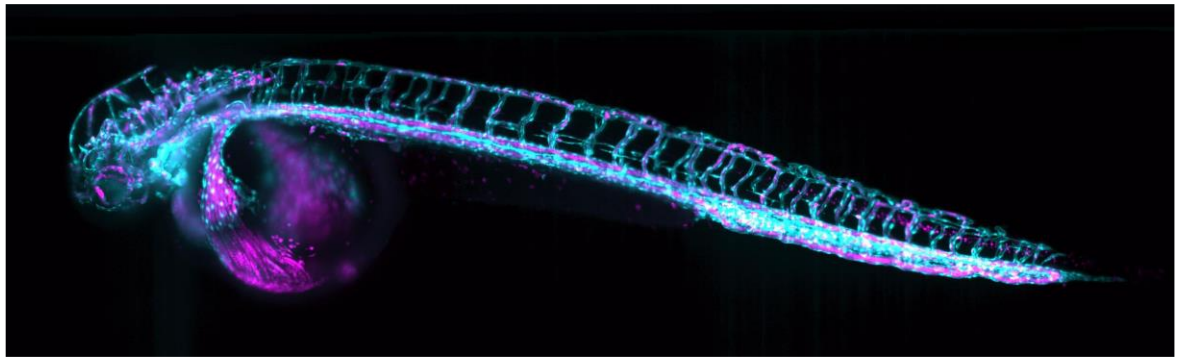
- high-dimensional (X, Y, Z, Channel, Time, …)
- saved in proprietary file formats
- of large file size

  (often in GB-, sometimes in TB-range)
- produced with complex experimental setups
- used for quantitative analysis → derived data

- … i.e. cumbersome to handle, store, and share?



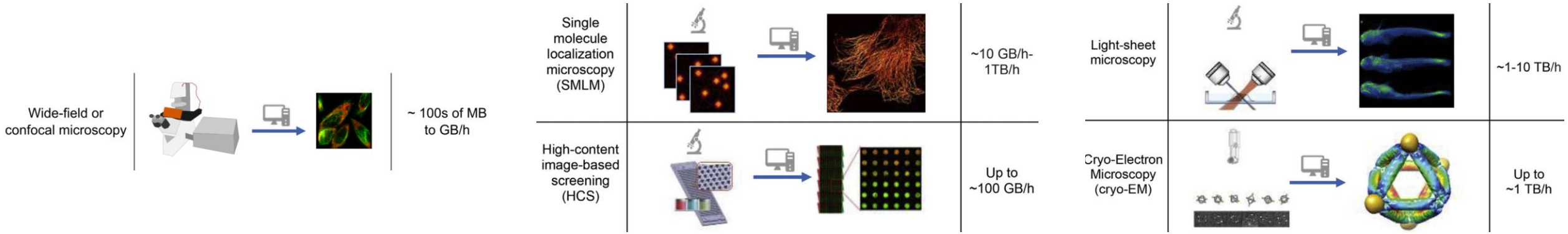
Courtesy of: Y. Stahl, S. Weidtkamp-Peters, HHU Düsseldorf



Courtesy of: Jan Huisken, University of Göttingen

The bioimaging data challenge

Increase in data(set) size per experiment

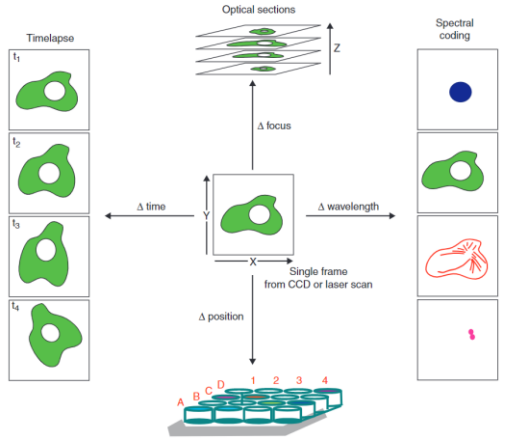


Modified after: Ouyang a& Zimmer, 2017, Curr Op Sys Biol, doi: 10.1016/j.coisb.2017.07.011

Increase in complexity of data structures

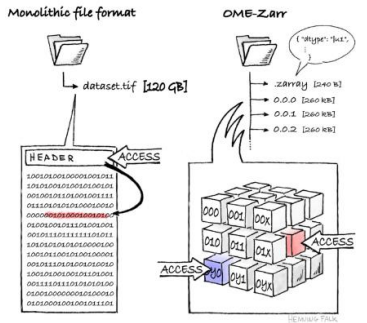
2D image planes

5D image stacks



Goldberg et al., 2005, Genome Biol, doi: 10.1186/gb-2005-6-5-r47

n-dimensional image



Falk, H. zarr-developers/zarr-illustrations-falk-2022 doi: 10.5281/zenodo.7037367

Research Data Ecosystems Emerging in Europe

How to promote sharing and reuse?

- *Guiding principles for data (FAIR)*

In practice?

- *Data ecosystem at the European level (→ EOSC)*
- ***Data ecosystem at the German level (→ NFDI)***

Germany

Nationale Forschungsdateninfrastruktur (NFDI)

2002 2006 2008 2012 2016 2018 2020 2022

„FAIR“

OECD:
Open Access

European Council:
European Open Science Cloud (EOSC)

Europe (& beyond)

NFDI Nationale Forschungsdateninfrastruktur

Nationale Forschungsdaten Infrastruktur

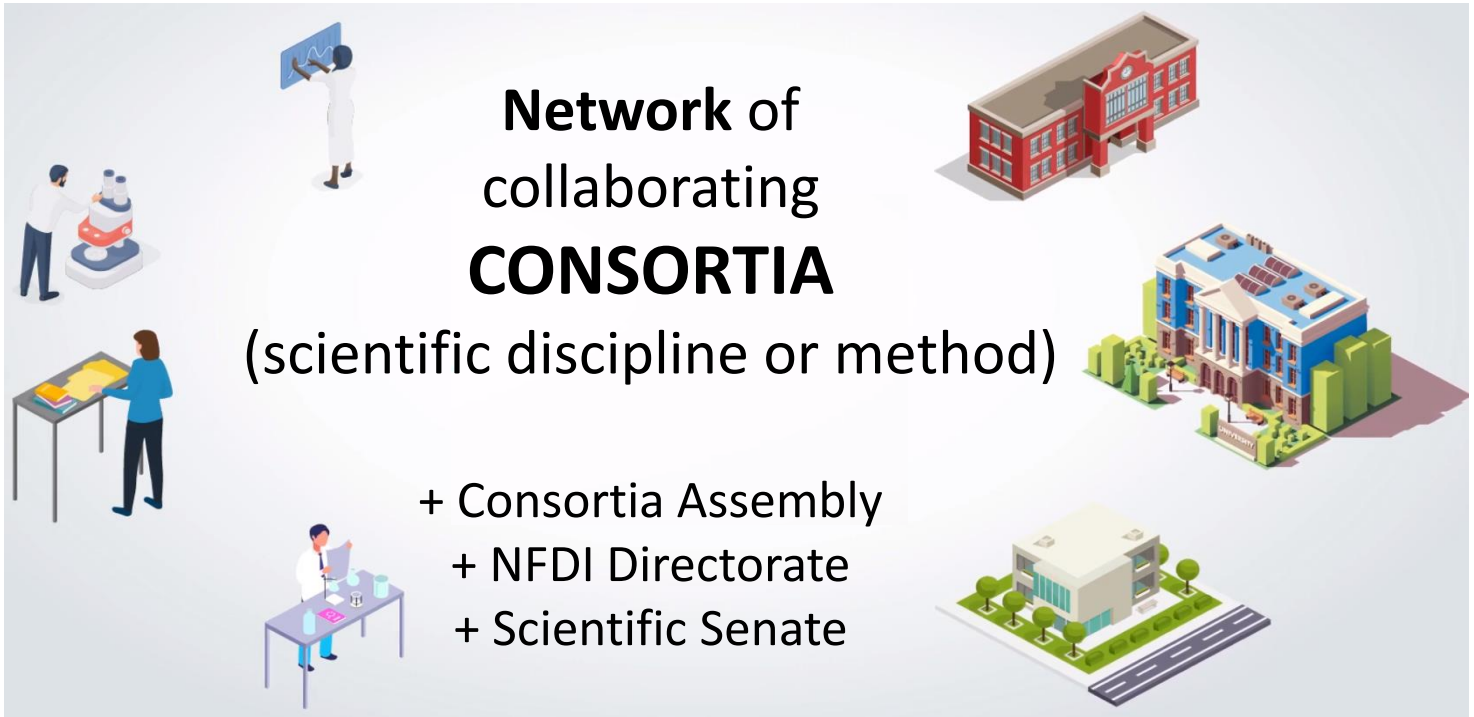
"The aim (…) is to systematically manage scientific and research data, provide long-term data storage, backup and accessibility, and network the data both nationally and internationally."

- **Science-driven (bottom-up)**
- "**Invest into people, not into metal**"

- 3 calls for application (2019 – 2021)

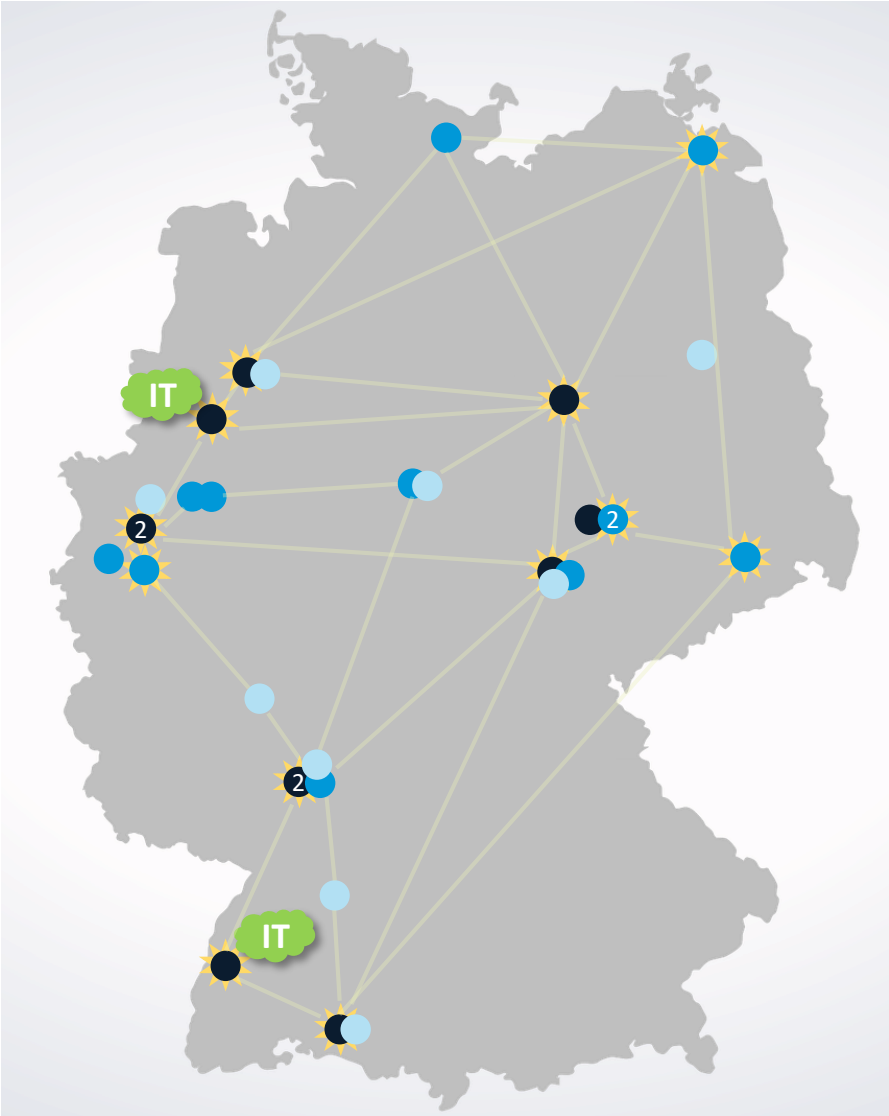
→ 26 consortia + 1 cross-consortia project

- NFDI4BIOIMAGE started in 03/2023



Adapted from: https://www.dfg.de/foerderung/programme/nfdi/informationsmaterialien/index.html, copyright: Deutsche Forschungsgemeinschaft

The consortium NFDI4BIOIMAGE



- 11 co-applicant institutions (6 Task Areas)
  Lead institution: Heinrich Heine University Düsseldorf
  Speaker: Prof. Dr. Stefanie Weidtkamp-Peters

IT infrastructure *(storage & playground: @Uni Münster & @Uni Freiburg)*
But not: a central data archive for all bioimaging data

- 12 participating institutions
- **data stewards (DaSts)** & **research software engineers**
- community use cases

Community-oriented services (implemented or planned)

- Help Desk for use community support
- Training portfolio
- Bioimage RDM playground infrastructure
- Supporting reproducible image analysis

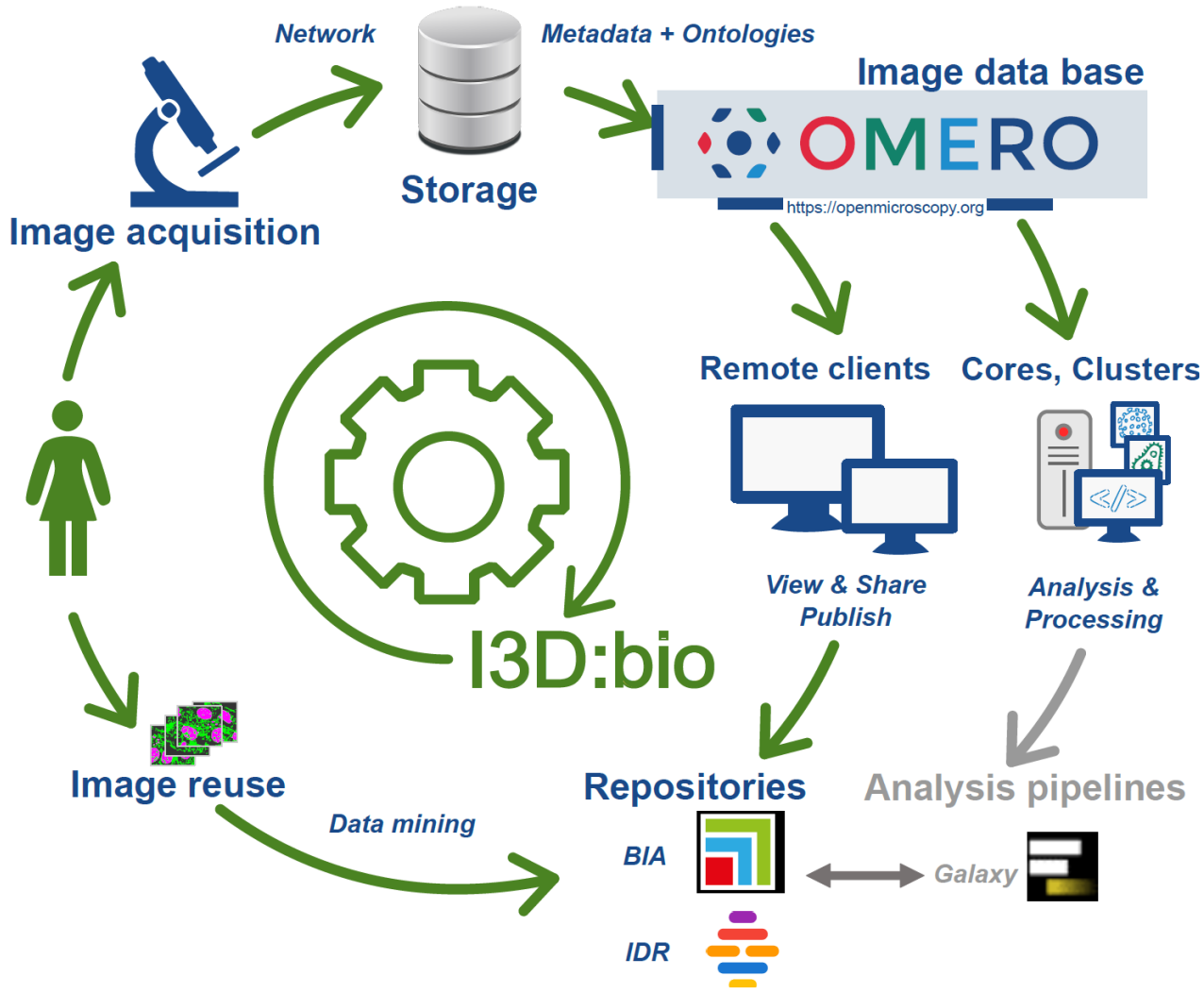
I3D:bio – OMERO-centered RDM project

Information Infrastructure

for BioImage Data (I3D:bio)

OMERO for bioimaging RDM

- **DFG-funded:** 1st phase: 2022-2024
- **Partners:** HHU Düsseldorf, Uni Osnabrück, Uni Freiburg, DKFZ Heidelberg
- **Goal:** Implement new OMERO instances at universities and research institutions in Germany, and provide training for OMERO users



Findability

F1: (Meta) data are assigned globally unique and persistent identifiers

F2: Data are described with rich metadata (described by R1 below)

F3: Metadata clearly and explicitly include the identifier of the data they describe

F4: (Meta)data are registered or indexed in a searchable resource

Mark D. Wilkinson, Talk: FAIR in a Series of Vignettes, CoRDI 2023, Karlsruhe, *https://tinyurl.com/FAIR-CoRDI23*, CC-BY:

Metadata for the purpose of discovery…
in the 'F' section of FAIR!
It should not be defining itself w.r.t. metadata for the purpose of reuse, which is the 'R' in FAIR…

Links resolve to https://www.go-fair.org/fair-principles/

Accessibility

A1: (Meta)data are retrievable by their identifier using a standardised communication protocol

A1.1: The protocol is open, free and universally implementable

A1.2: The protocol allows for an authentication and authorisation procedure where necessary

A2: Metadata should be accessible even when the data is no longer available

Links resolve to https://www.go-fair.org/fair-principles/

Interoperability

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2: (Meta)data use vocabularies that follow the FAIR principles

I3: (Meta)data include qualified references to other (meta)data

Links resolve to https://www.go-fair.org/fair-principles/

Reusability

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1: (Meta)data are released with a clear and accessible data usage license

R1.2: (Meta)data are associated with detailed provenance

R1.3: (Meta)data meet domain-relevant community standards

Mark D. Wilkinson, Talk: FAIR in a Series of Vignettes, CoRDI 2023, Karlsruhe, *https://tinyurl.com/FAIR-CoRDI23*, CC-BY:

CC0

In hindsight, this was probably a bad choice of words…

Maybe “data access/usage policy or license” would have been better?

Links resolve to https://www.go-fair.org/fair-principles/

Group exercise:
What does FAIR mean for bioimaging?

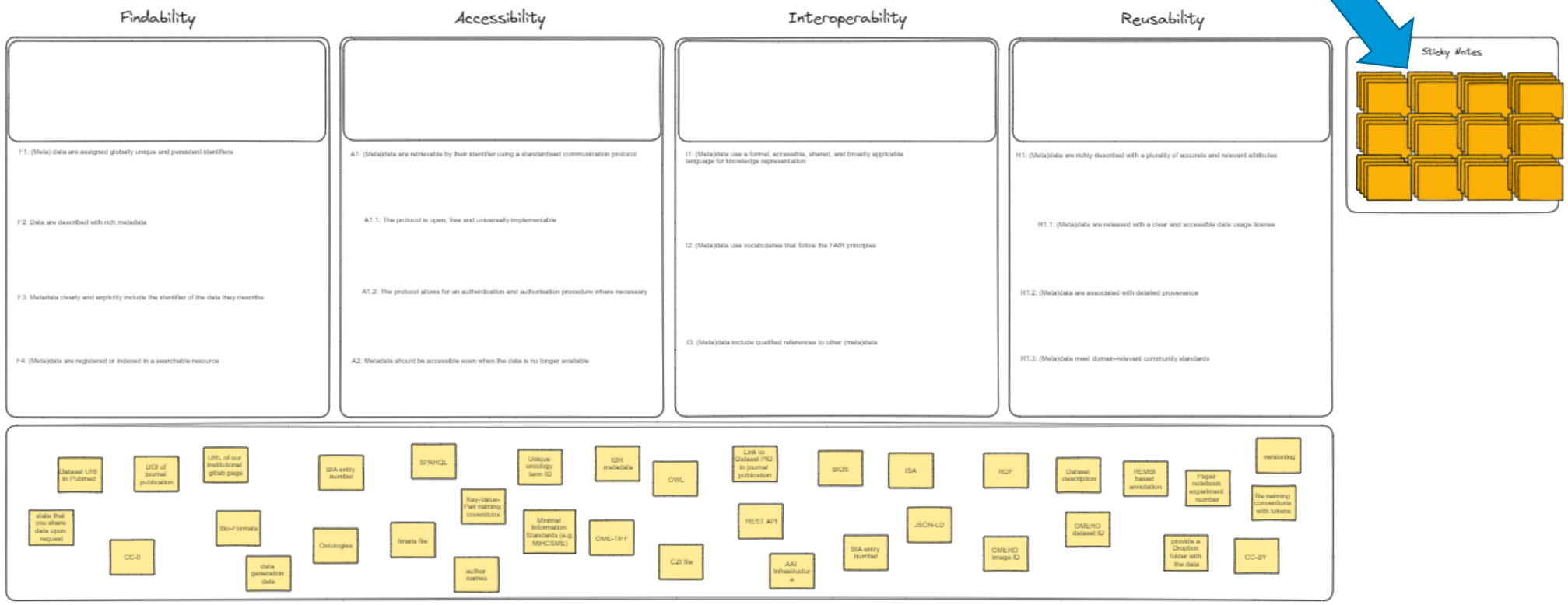
→ Excalidraw

Use the sticky notes to assign concepts, tools, practices, or resources to the FAIR principles

Place sticky notes here:

Or here if you think they are specific for certain sub-category of FAIR

Optional:
You can use these prepared notes if you know the terms and if find them to apply



Acknowledgments

In cooperation with

Information **I**nfrastructure for Bio**I**mage **D**ata (I3D:bio)

https://www.i3dbio.de/

German Cancer Research Center (DKFZ), Heidelberg
Department Enabling Technology

Dr. Christian Schmidt, Project Coordinator

office@nfdi4bioimage.de

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrasstructure – NFDI 46/1 – 501864659

In cooperation with

The NFDI4BIOIMAGE consortium comprises legally independent partners and does not act autonomously towards third parties. The authors represent the contributions from their respective affiliated institutions and work together for the project.