

FAIR by design at the bench

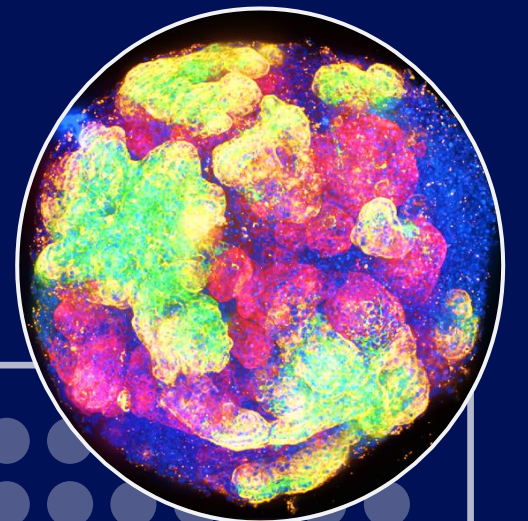
What it takes to produce FAIR (bioimaging) data
from the instrument-up

Sylvia Le Dévédec · Leiden University · LACDR
Leiden Cell Observatory

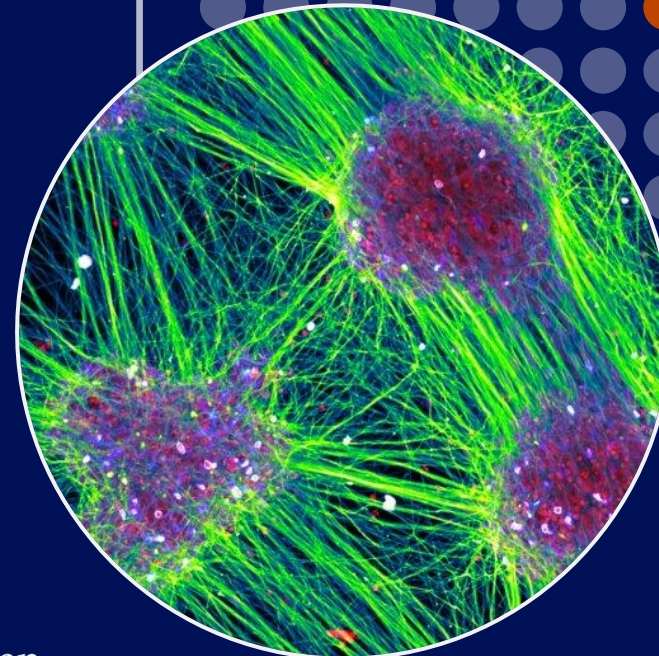
FAIA: AAC; AC: Co-Creation; Claude Sonnet, 5.5 medium



Universiteit Leiden

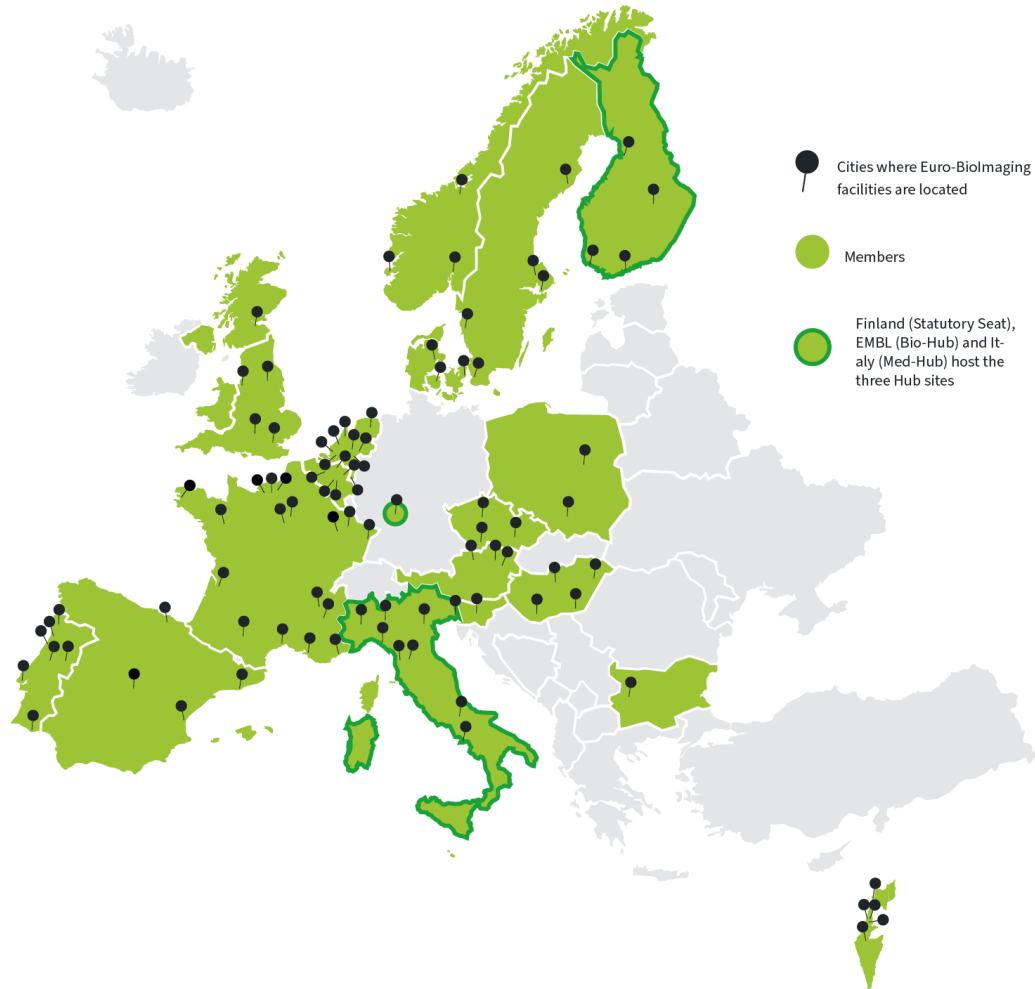


Kidney organoids



Human iPSC derived neurons

Euro-BioImaging in numbers



FAIR Repositories



BioImage Archives



Image Data
Resources

295

imaging facilities in 39 Euro-BioImaging nodes

416 TB

in IDR: 14 M images from 143 studies

~3 PB

open image data in the BioImage Archive and EMPIAR (2023)

2

Large, fast-growing and varied: FAIR has to be automated at the source.

Two routes to FAIR data: prevent and repair (we do both)



Instrument



Facility



Analysis



Repository

**Ex-post
FAIRification**
Repair afterwards

Files and notes, no
structure

Stored as is

Steps documented
later

Metadata
reconstructed by
curation or algorithms

FAIR by design
Prevent the gaps

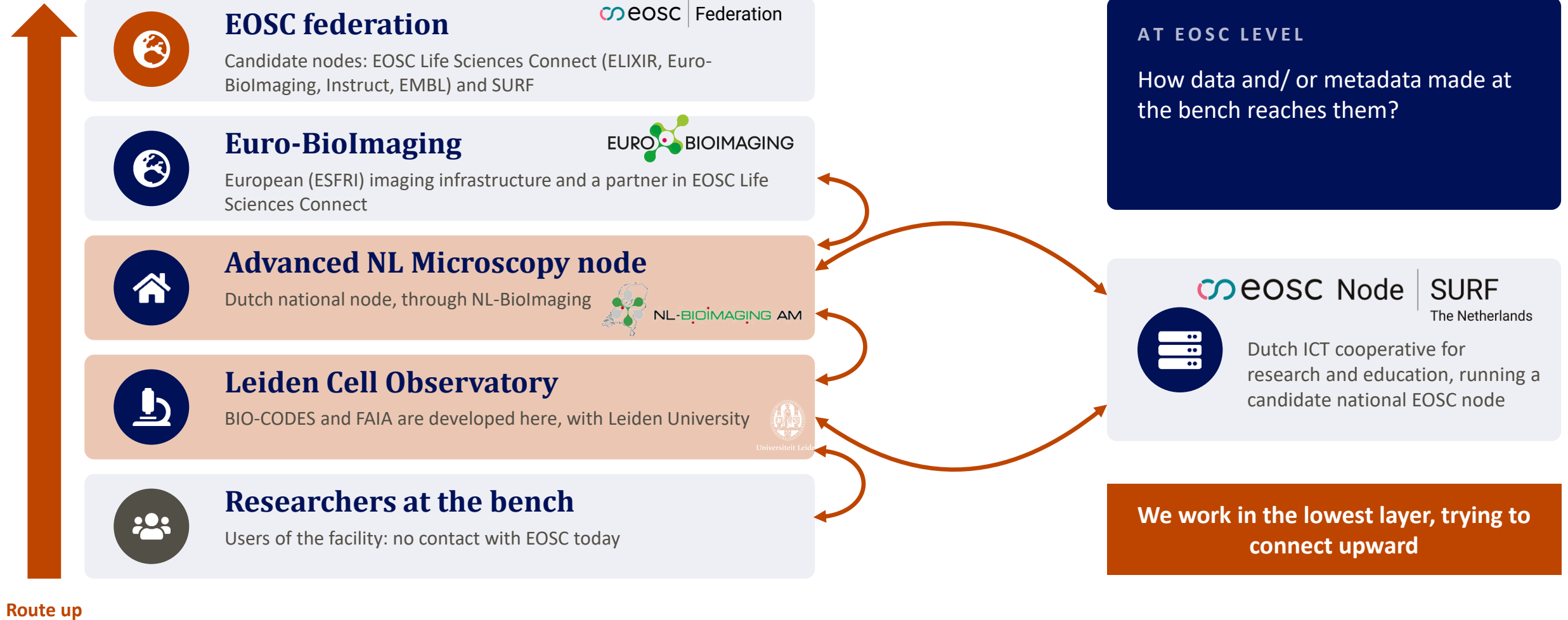
Experiment described
before it runs

Metadata and content-
based ID attached on
ingest

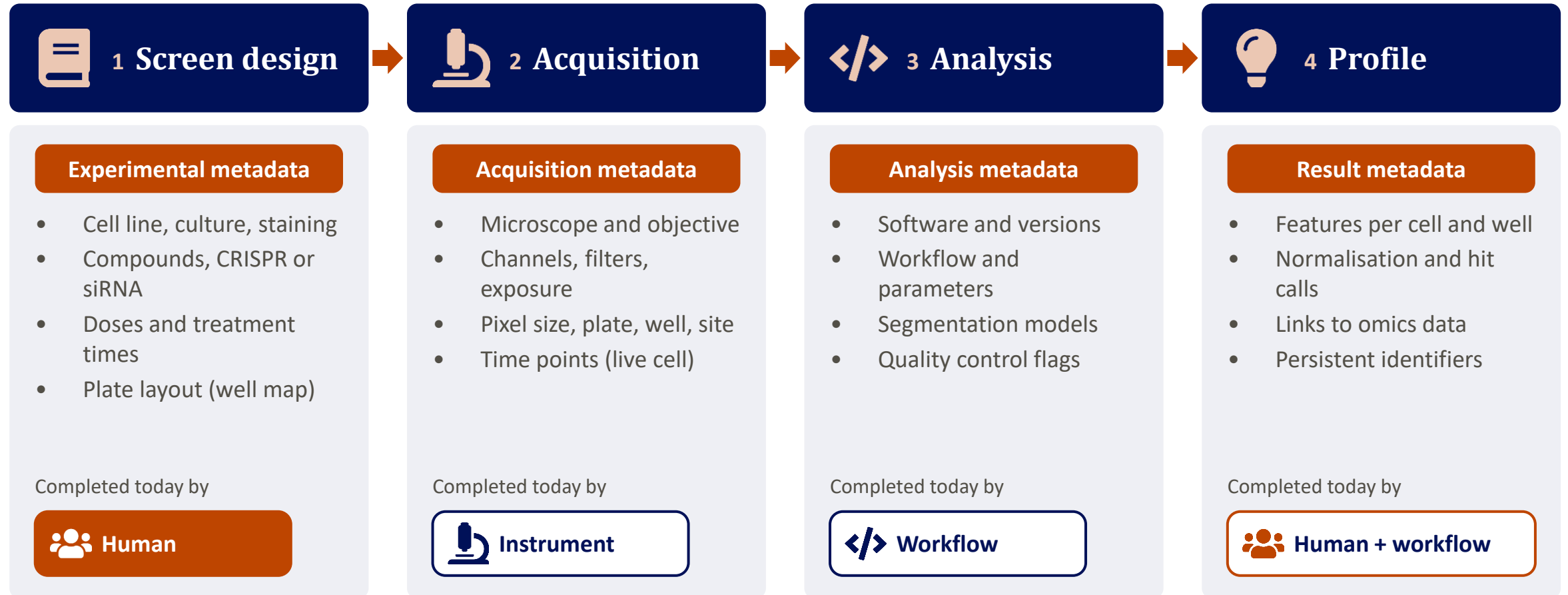
Provenance recorded
by the workflow

Deposited with
identifiers

From the bench towards EOSC

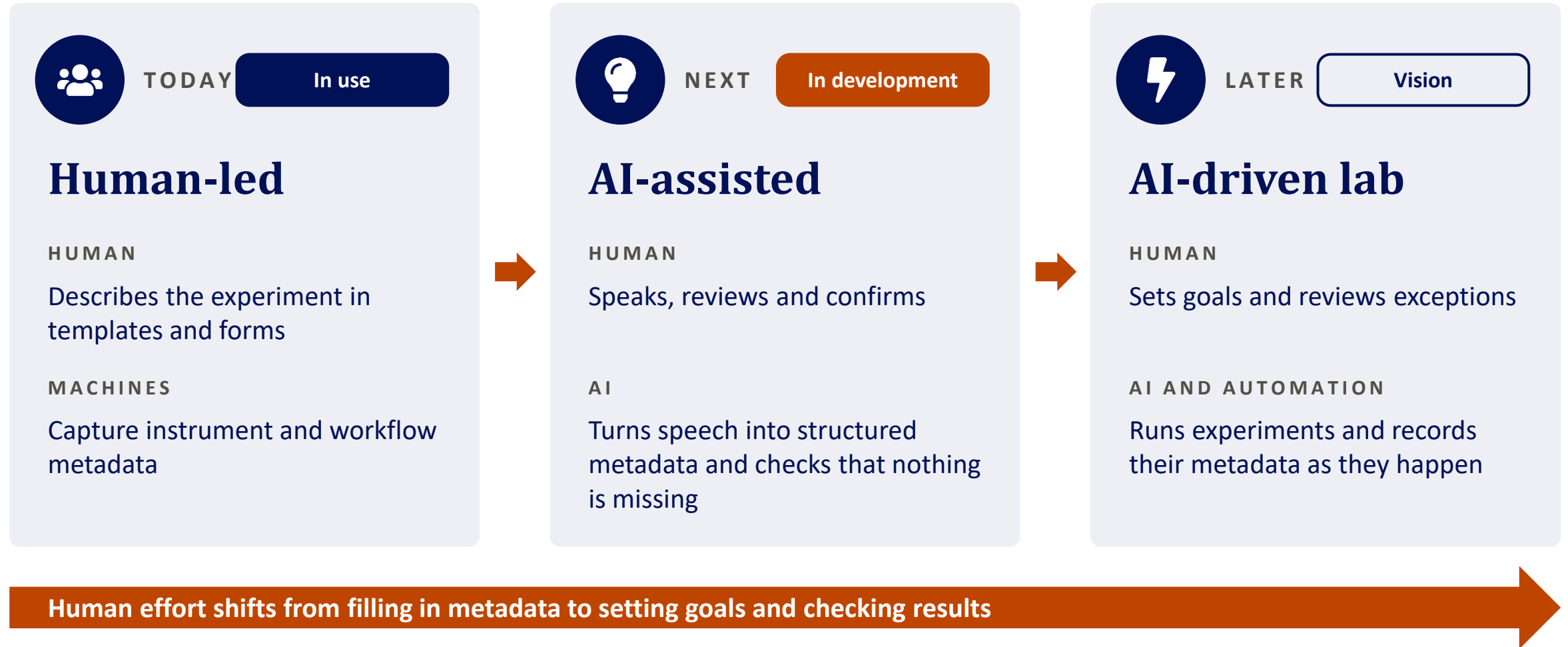


From microscope to phenotypic profile



Humans are still heavily needed mainly where the biology is described and interpreted

From human-led to AI-driven metadata



Layer 1: the pipeline, instrument-up



FAIR by design across all five stages

Each stage captures structured metadata and hands it to the next, so nothing has to be reconstructed later.

The same image has no shared identity across raw data, repository and publication.



No shared identity

No audit trail

Lost provenance

BIO-CODES: content-based identifiers

ISCC:KED572P4AOF5K6QXQA4T6OJD5...

International Standard Content Code (ISO 24138),
derived from the data itself



Generate

Automatically, at the point of
acquisition



Store

In OMERO, now next to the
MIHCSME metadata but later
linked.



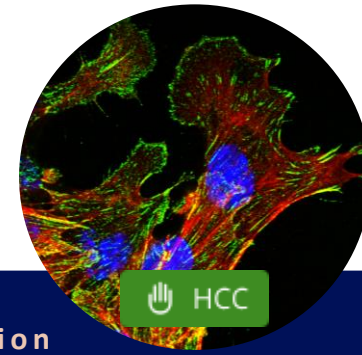
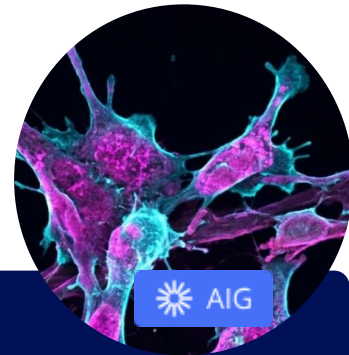
Follow

To the repository and into the
publication

Status: design and proof of concept. The OMERO integration is a beta version.



Layer 3: FAIA, declaring AI involvement



BIO-CODES

Which exact data is this?

Content-based identity (ISCC)



FAIR AI Attribution

Did AI create or touch it, and how?

Linked to the ISCC, signable



Dataset or image



ISCC fingerprint



Signed declaration



BIO-CODES and FAIA
registries

Status: Applying FAIA to bio-imaging data is the next step in collaboration with Leiden University.

Beyond bioimaging: tools for FAIR Digital Objects

 Identifier	 Metadata	 Content	 Provenance
ISCC ISO 24138, computed from any digital asset Generic	MIHCSME An ISA-based template pattern: swap in each field's own descriptors Per field	OME-Zarr Chunked n-dimensional image arrays Image-based data	Galaxy and FAIA Cross-disciplinary workflows, and AI declarations for text, images, audio, video and datasets Generic

The domain part is a template. The automation around it is generic.

Image analysis in Galaxy already serves bioimaging and Earth observation.

What it takes for the facility: cost and effort



People

Data stewardship for the platform



Training

Researchers and facility staff



IT infrastructure

OMERO, storage, compute and backup



Governance

Templates, SOPs and policy kept current



Researcher effort

Describing the experiment before it runs

What could EOSC help with?

ASK 1



ISCC adoption

In EOSC services and in Euro-BiolImaging, so data keeps its identity.

ASK 2



FAIA-aware registries

Registries that store and resolve AI-attribution declarations.

WHERE IT PLUGS IN



Instrument



OMERO



Galaxy workflows



EOSC federation

Take-home messages



FAIR by design is automation moved upstream, and it complements ex-post FAIRification.



Three layers: standards at the instrument, identity at acquisition, attribution for AI.



The cost is people, training and infrastructure. EOSC can help for instance with ISCC adoption and FAIA-aware registries.