OSCARS
Open Science Clusters' Action for Research & Society

Funded Project

Enhancing AI-Readiness of Bioimaging Data with Content-Based Identifiers (BIO-CODES)

Sylvia Le Dévédec, Leiden University, ORCID: 0000-0002-0615-9616

Implemented by

Universiteit Leiden

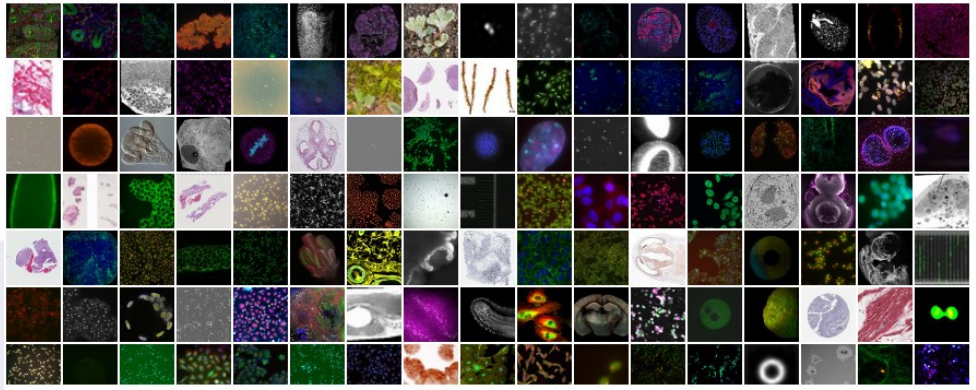
ISCC Foundation

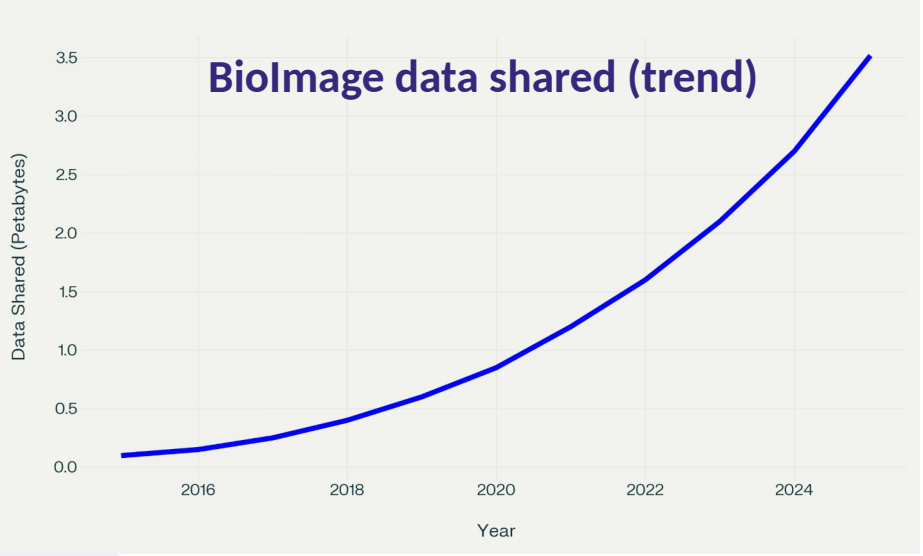
Funded by the European Union

CHALLENGES IN BIOIMAGING

- Rapid **growth of bioimaging data** presents challenges in management, integrity, and reusability.
- **Lack of standardized identifiers** for bioimages leads to difficulties in tracking, verifying, and certifying data.
- Existing **FAIR principles** (*Findability*, *Accessibility*, *Interoperability*, *Reusability*) are not fully implemented in bioimaging workflows.
- **Use of AI in bioimaging** requires transparency in data provenance and authenticity.

Overview IDR website (16-06-2025)





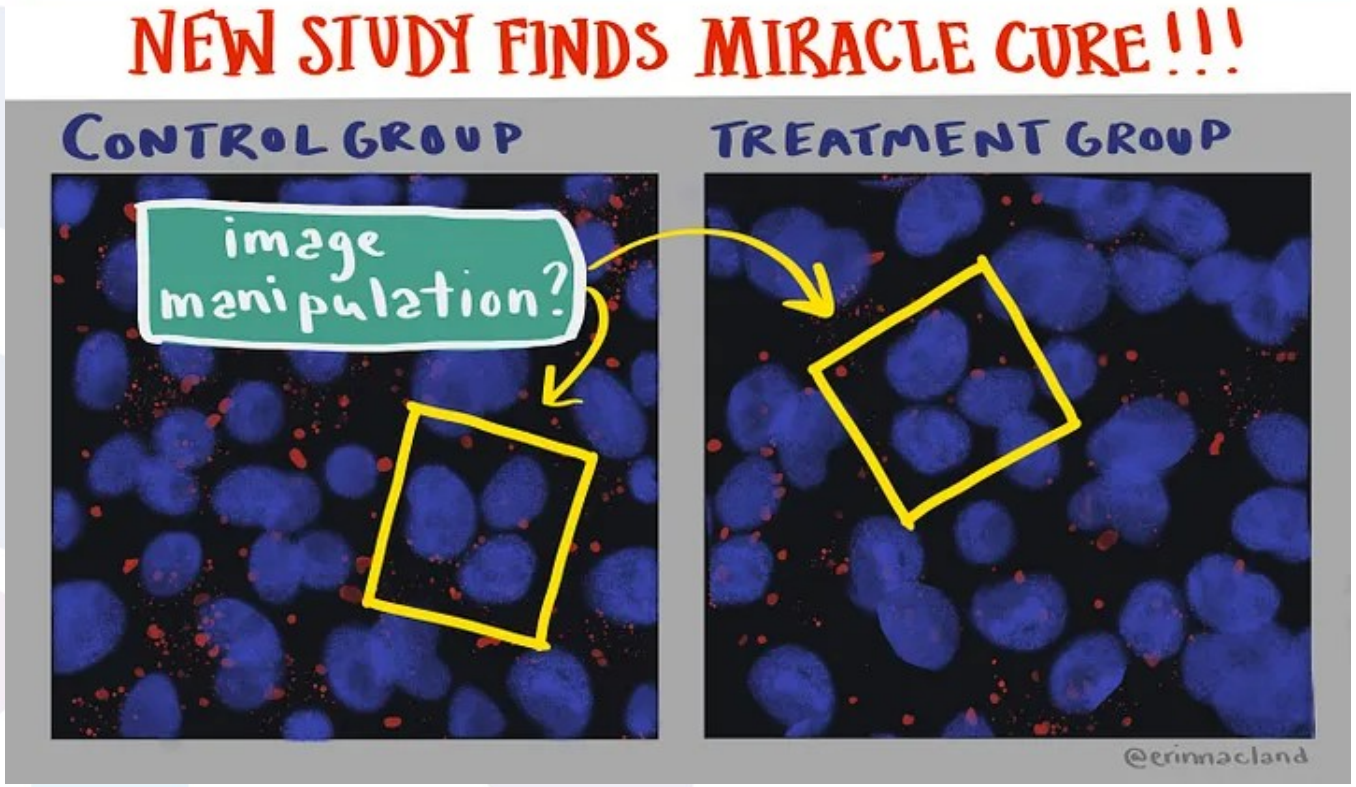
BIOIMAGING DATA INTEGRITY

Table 5
Distribution of retraction reasons of retracted articles from 2003 to 2022.

Retraction Reason	Count (%)
Data and Results Issues	6160 (28.8)
Plagiarism and Duplication	4085 (19.1)
Investigations and Findings	3588 (16.8)
Authorship and Ethical Concerns	1918 (9.0)
Misconduct and Fraud	1288 (6.0)
Image Manipulation and Fabrication	1092 (5.1)
Peer Review and Editorial Issues	985 (4.6)
Withdrawal and Retraction Notices	763 (3.6)
Institutional and Policy Issues	643 (3.0)
Complaints and Objections	418 (2.0)
Miscellaneous	412 (1.9)
Procedural and Legal Issues	70 (0.3)
Total	**21422 (100)**

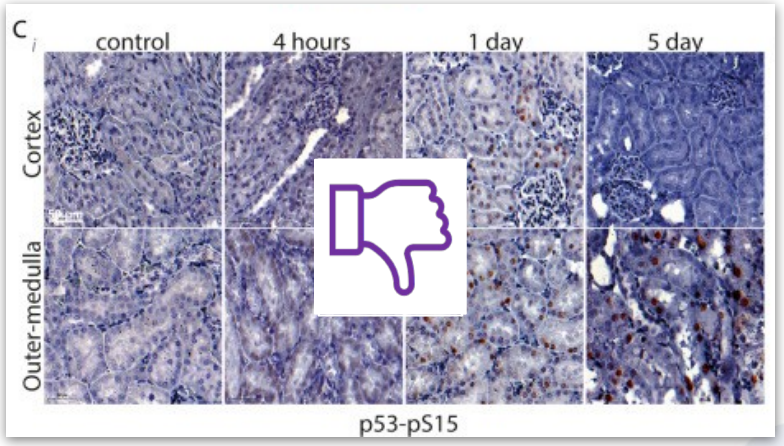
Of the 8466 articles identified in the Web of Science, 7375 were successfully linked with records from the Retraction Watch Database. As each article may have multiple retraction reasons, a total count of 21,422 for the various reasons was documented.

From Malcolm Koo and Shih-Chun Lin, Helyon 2022

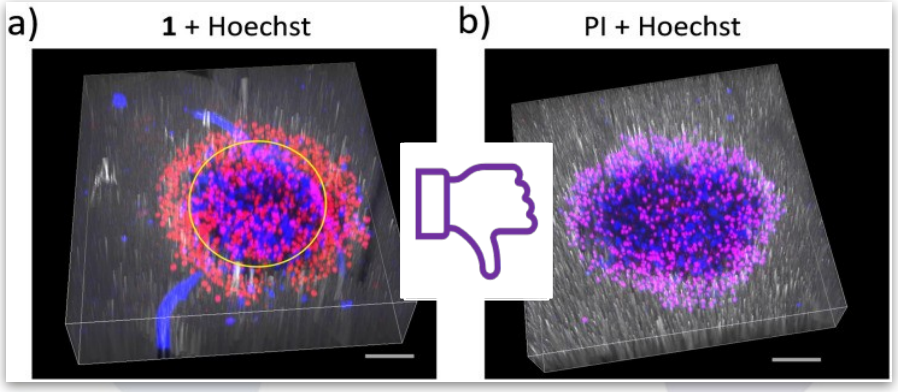


https://medium.com/@erinnacland/a-science-browser-extension-everyone-should-use-b15ad69ad6de

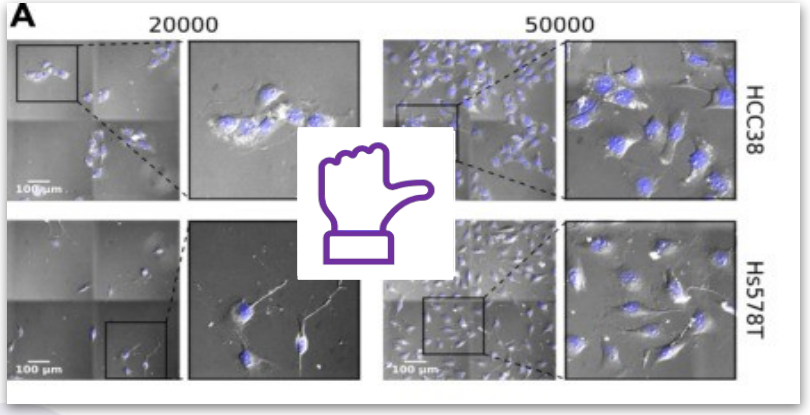
DATA PROVENANCE IN SCIENTIFIC PUBLICATION LACKING



Lukas S Wijaya et al., Cell Biol Toxicol . 2025

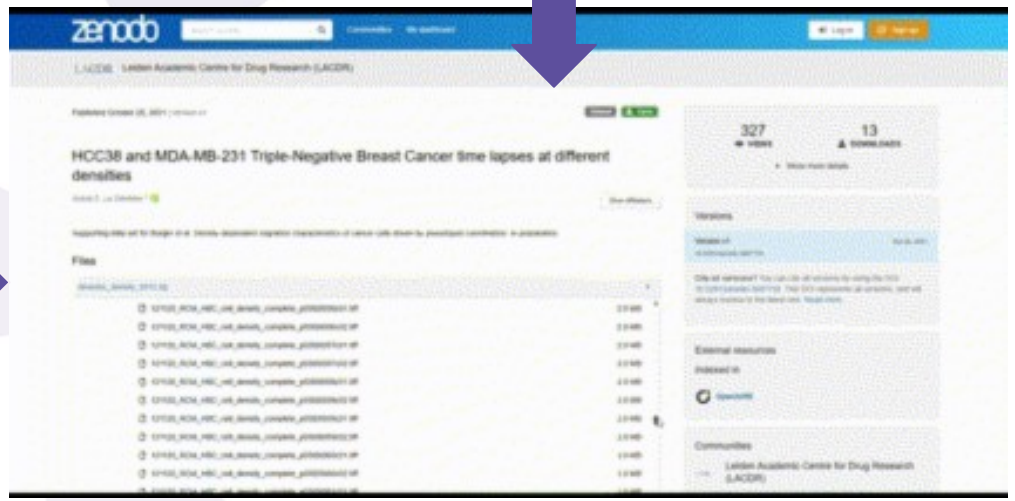


Ramu Vadde et al, Chem Commun (Camb). 2024

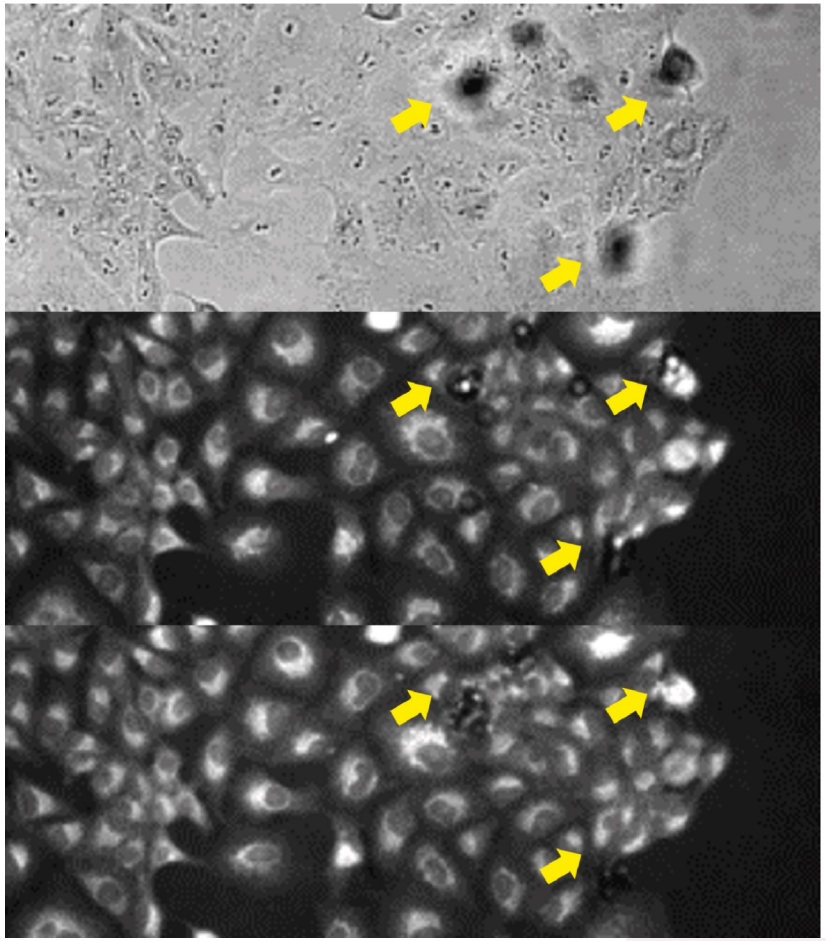


Gerhard Burger, Front Cell Dev Biol . 2022

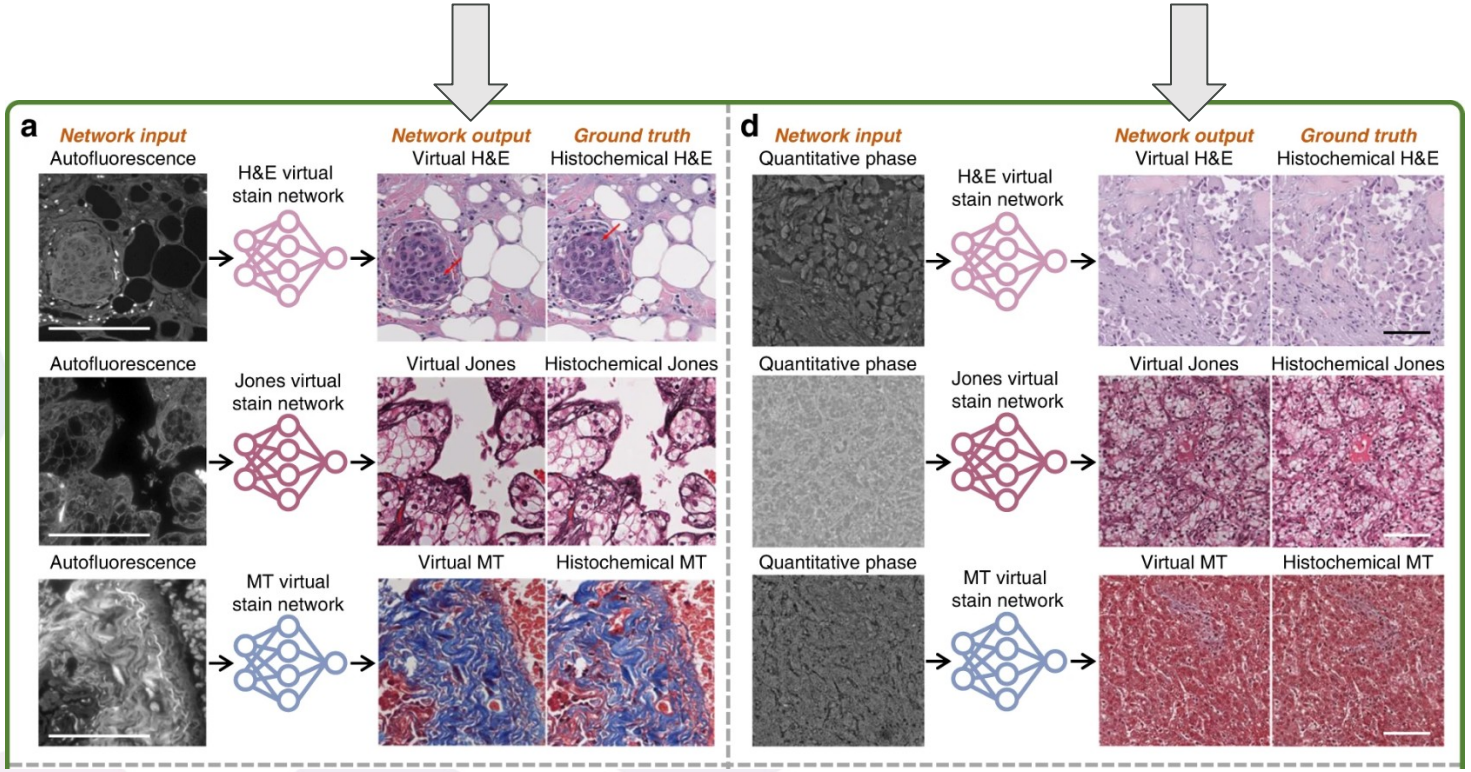
Data deposited in Zenodo **but no clue which images** were used to generate the figure in the scientific publication



NEW CHALLENGE: GENAI and SYNTHETIC IMAGES



Xiaodan Xing et al, Computers in Biology and Medicine. 2024



Bijie Bai et al Light: Science & Applications. 2023

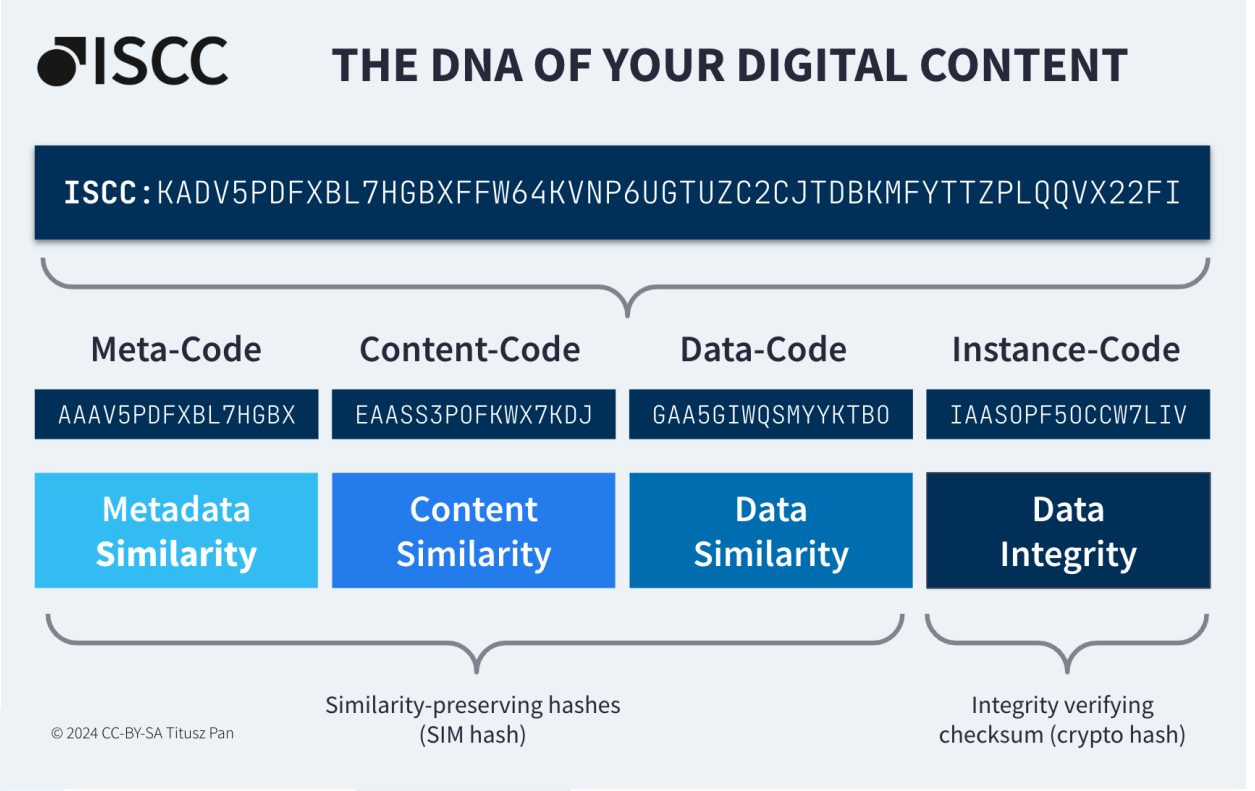
What is **real** (=ground truth) and what is **synthetic**?

POSSIBLE SOLUTIONS?

- **Advanced Data Management**
  Digitalization, automation, biobanking, large volumes, velocity, bioimaging data, quality, reproducibility
- **FAIR Principles**
  Findable, Accessible, Interoperable, Reusable, data sharing, AI-driven biomedical research
- **Data Integrity & Provenance**
  Data integrity, conventional data, AI-generated synthetic data, transparency, trustworthiness
- **Standardized Content Identification**
  Content-based identifiers, unique identification, deduplication, synchronization, provenance tracking

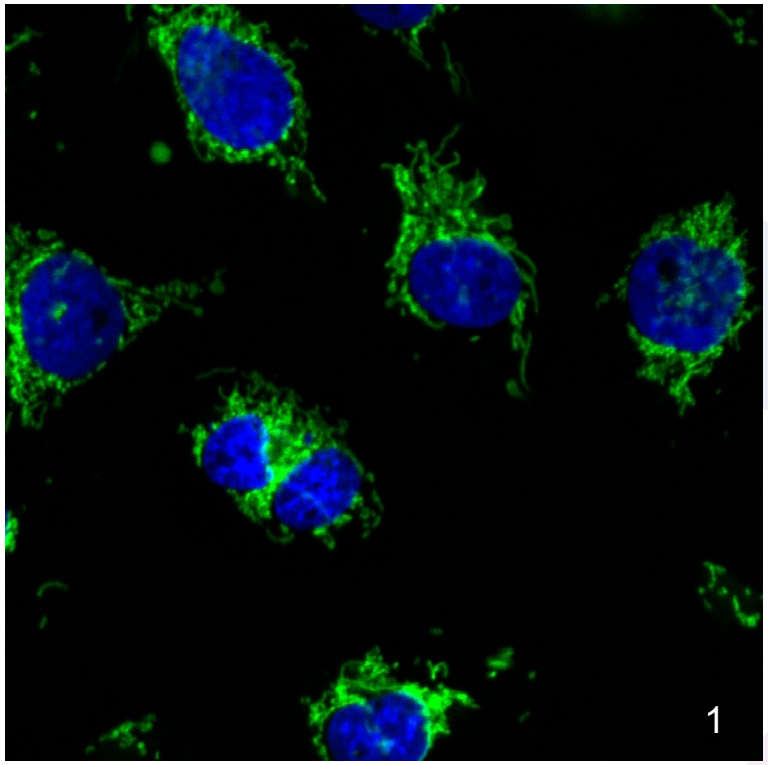
WHAT IS ISCC (INTERNATIONAL STANDARD CONTENT CODE)?

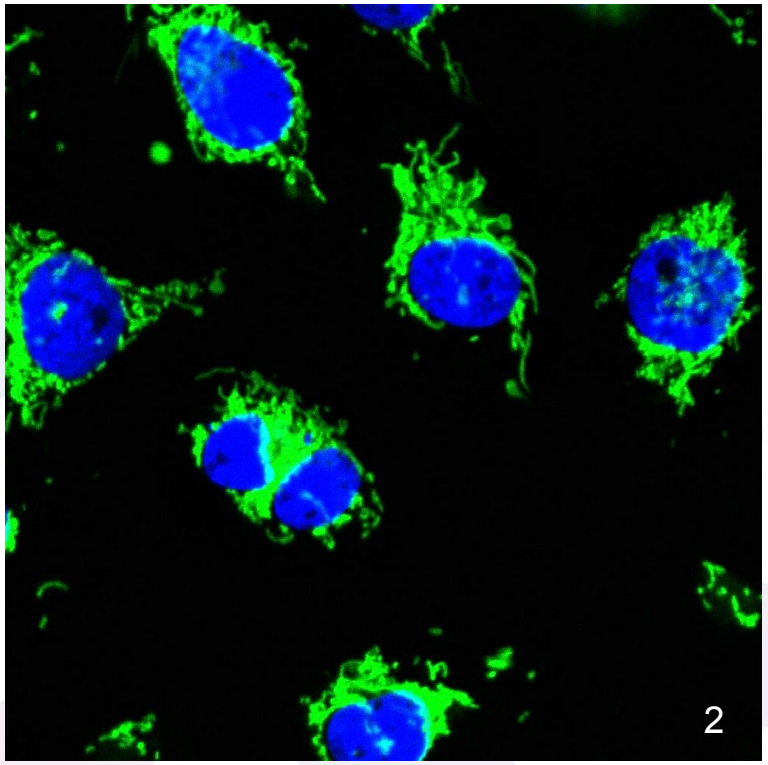
- **ISO standard** (ISO 24138) for content-derived identification of digital media (text, images, video, audio, mixed).
- **A composite identifier** that exhibits similarity-preserving properties
- **Generated directly from the media asset** – no central database or manual registration needed.
- ISCC combines **cryptographic and similarity hashes**, enabling:
  - Content integrity verification
  - Similarity detection
  - Transparent tracking of content provenance and versions



Source: iscc.io

EXAMPLE





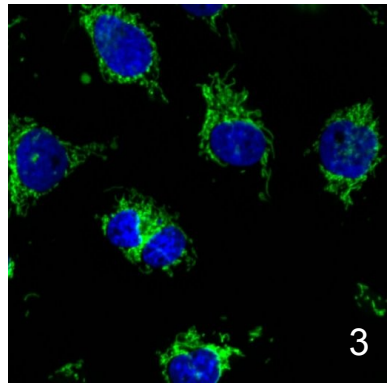
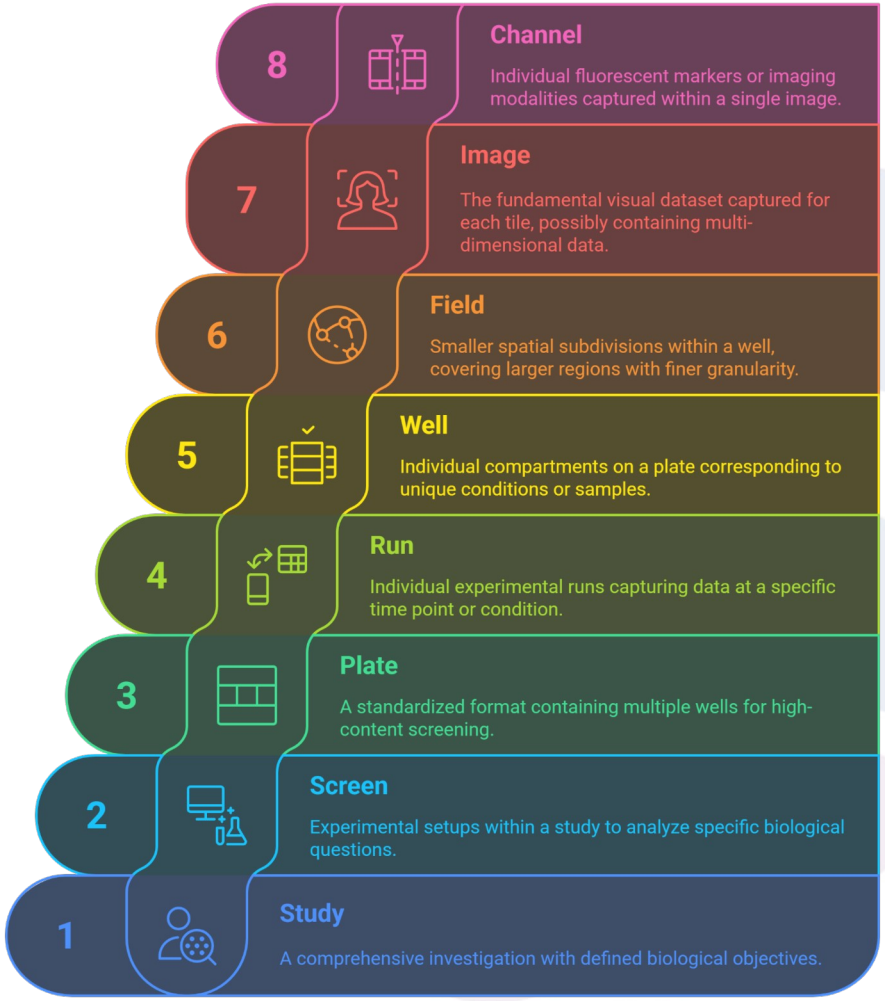


Image 1 (original; 1024x1024) ->
ISCC:KEC5FEDFHVYTLFAWZPRUARXFXGMYX5MSO2XYTVQWXBUAH74TU4YG3ZY
Image 2 (increased contrast; 1024x1024) -> ISCC:KECV74PVO3A7XPC7ZPRUARXFXGMYWN2WQDAR56XLTGAT4ELZMQTATEA
Image 3 (512x512) ->
ISCC:KEC5PML5X5233XYWZPRUARXFXGMYXQY3ZTHZJU3HI62GLIIMH3OCYMY

***Note**: The Image-Code unit of the ISCC is matching across all 3 image representations

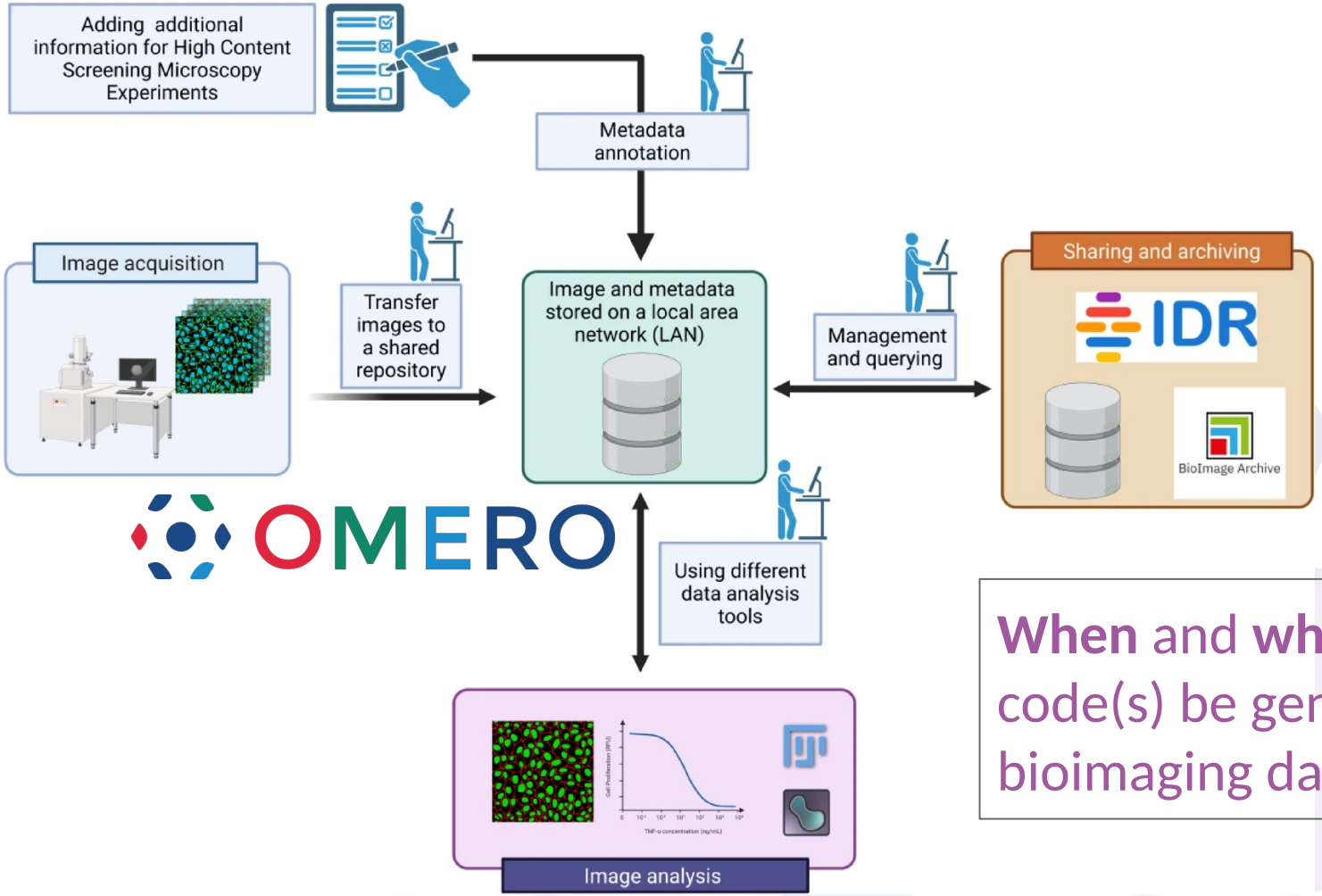
DEFINING IMAGE DATA: A FIRST CHALLENGE



Source: iscc.io

- Defining the level of granularity for generating the ISCC is a challenge.
- What is an ´image´ in bioimaging?

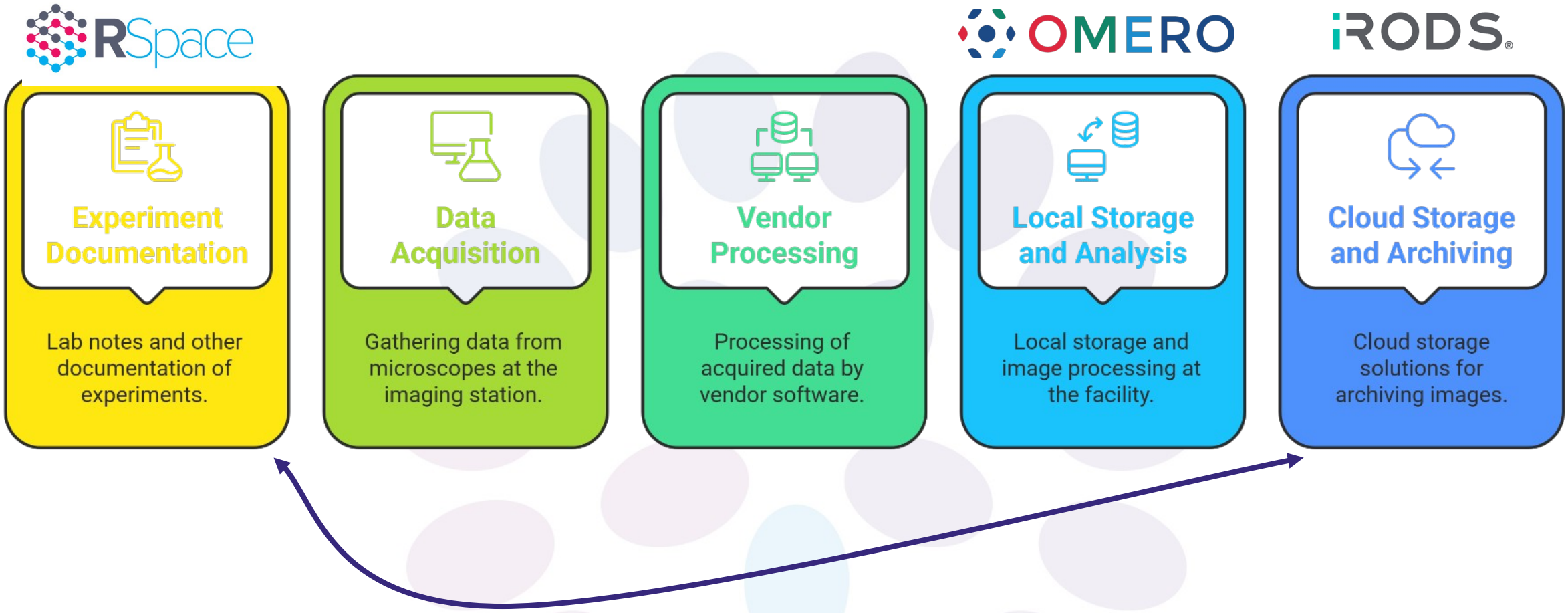
BIOIMAGING DATA WORKFLOW



Adapted from [Riccardo Massei](#) at al, Scientific Reports, 2025

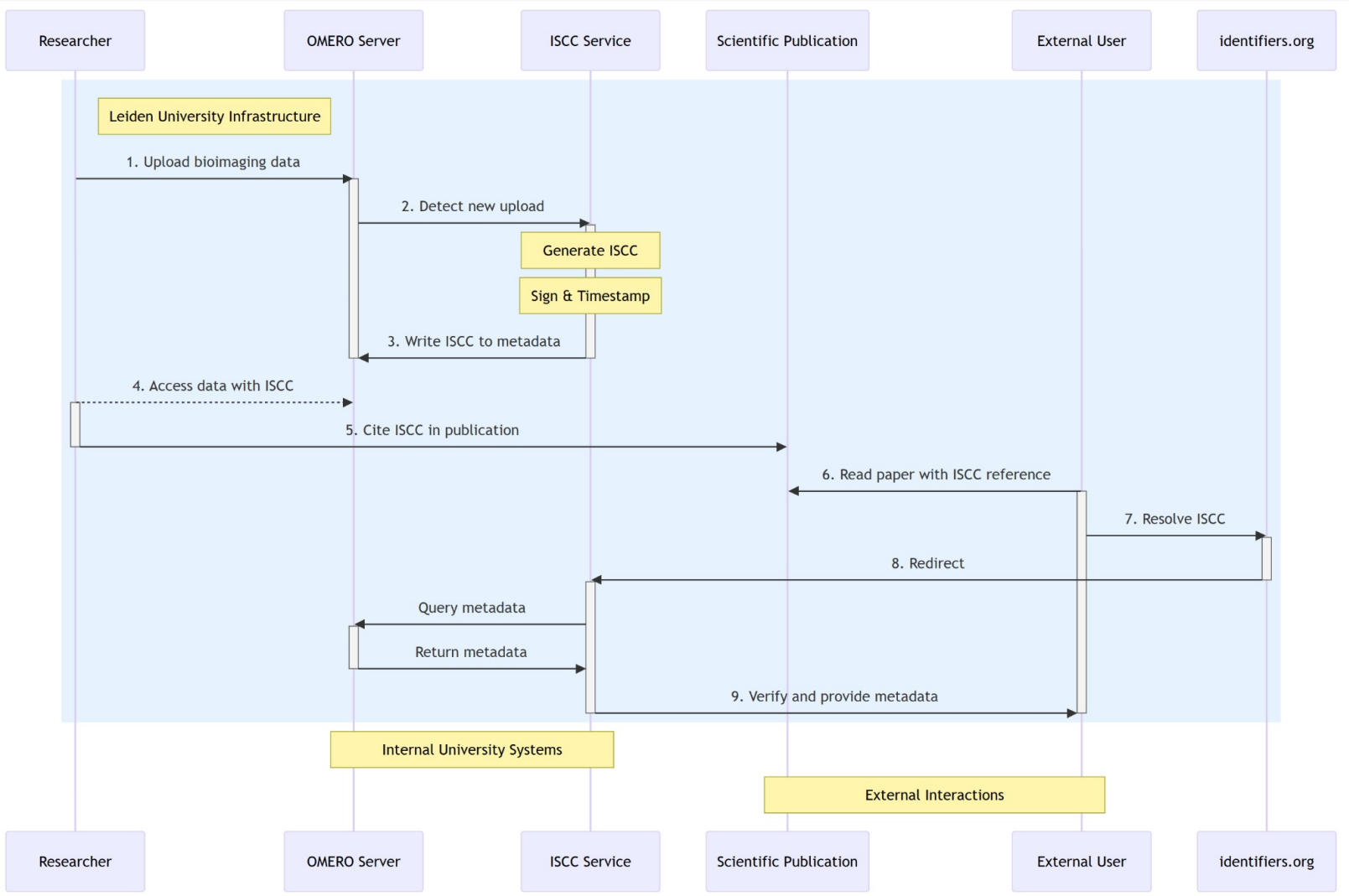
When and **where** should the ISCC code(s) be generated for the bioimaging data?

A SECOND CHALLENGE: MULTIPLE INTEGRATION POINTS



Source: iscc.io

WHERE TO IMPLEMENT THE GENERATION OF ISCC?



Automating ISCC code generation at the point of image acquisition
->
COLLABORATION STARTED WITH NIKON EUROPE BV

Source: iscc.io

EXPECTED PROJECT DELIVERABLES

- **Implementing the International Standard Content Code (ISCC) for bioimaging data (e.g. OME-Zarr)** and **figure renderings**.
- Developing a **proof-of-concept integration** of ISCC with existing bioimaging workflows.
- **Automating ISCC code generation** at the point of image acquisition for better data traceability (start collaboration with Nikon Europe BV).
- Exploring integration with existing bioimaging metadata standards like RO-Crate and platforms like OMERO.

DREAMED OUTCOMES

- Standardized **content-based identifiers for bioimaging data**.
- Improved data **integrity, transparency, and reusability**.
- **ISCC adoption** in Euro-BioImaging and EOSC infrastructures.
- Enhanced **AI-readiness** of bioimaging datasets for trusted AI applications.
- Integration of ISCC within **vendor platforms** and research infrastructures.

- **Adoption Barrier**: Technical complexity may cause delay in adoption by bioimaging facilities and vendors → **Mitigation**: Engage stakeholders early, provide material for simple onboarding.

- **Technical Integration**: Development of ISCC for bioimaging data incl. proprietary image formats is challenging → **Mitigation**: Considering dedicated ISCC code unit for only OME-Zarr (and most common bioimaging formats).

THE TEAM

Titusz Pan

Martin Etzrodt

Sebastian Posth

ISCC Foundation:
ISCC technology implementation

Sylvia Le Dévédec

Maarten Paul

Lennard Voortman

Joost Willemse

Leiden University/ LUMC
Imaging core facilities: data producers

Katy Wolstencroft

ACDC/AMC
FAIR data

Josh Moore

German
BioImaging/OME
OME zarr developper

NL-BIOIMAGING AM

Enhancing AI-Readiness of Bioimaging Data with Content-Based Identifiers (BIO-CODES)