OSCARS
Open Science Clusters' Action for Research & Society

bio-codes.io

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

Dr. Martin Etzrodt, ISCC Foundation, ORCID: 0000-0003-1928-3904

Implemented by

Universiteit Leiden

ISCC Foundation

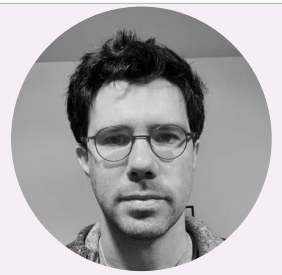
German BioImaging
Gesellschaft für Mikroskopie und Bildanalyse

NL-BIOIMAGING AM

Funded by the European Union

TEAM

Titusz Pan



Martin Etzrodt

Kira Lemke

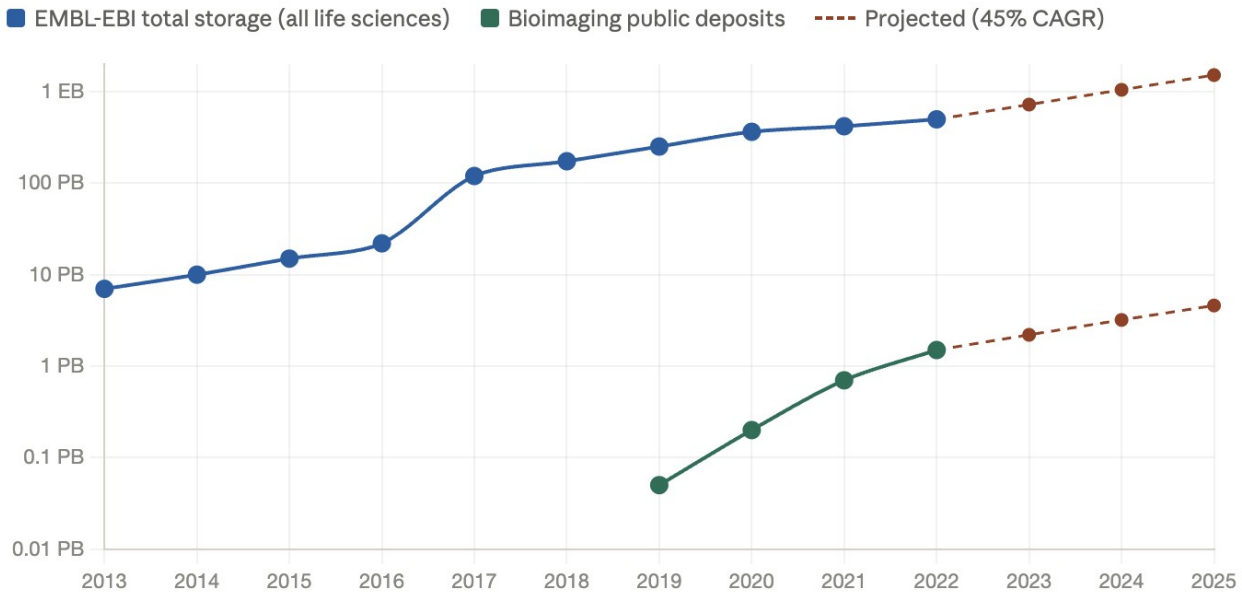
Sebastian Posth

Maarten Paul

Sylvia Le Dévédec

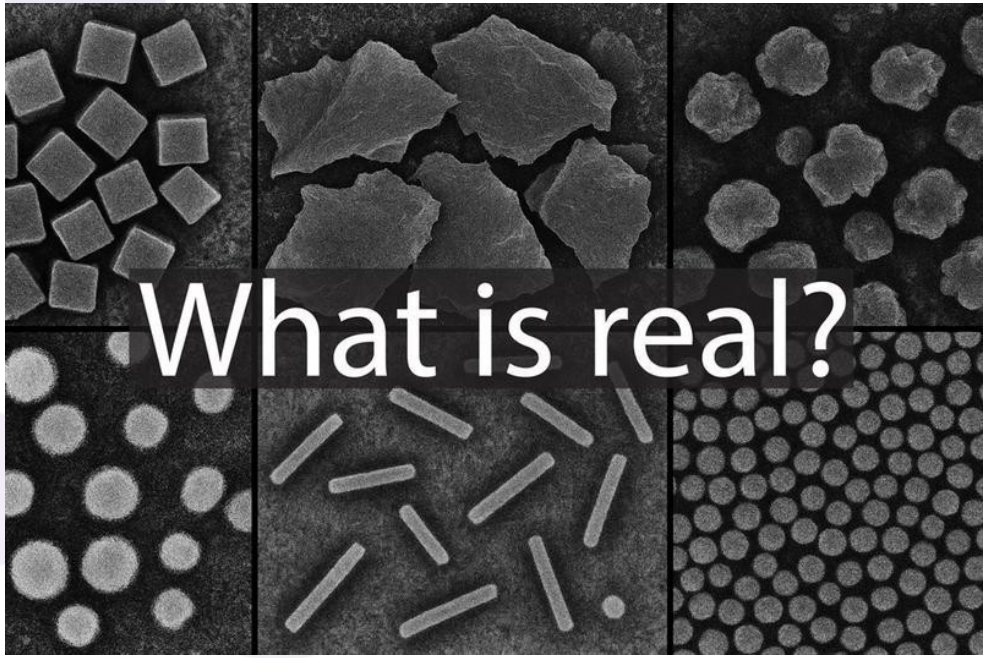
Josh Moore

CHALLENGES IN BIOIMAGING



Sources: Claude.ai with data from Cook et al. 2022 (J Mol Biol); Kersey et al. 2017 (NAR); Ruan et al. 2024 (Nat Methods); Bajcsy et al. 2025 (Nat Methods).

- > Growing volume of data
- > No Audit Trail
- > Lost Provenance
- > ***Inconsistent metadata*** *(hurts AI training)*



https://www.research-in-germany.org/idw-news/en_US/2025/9/2025-09-15_The_Rising_Danger_of_AI-Generated_Images_in_Nanomaterials_Science___Experts_Warn_in_Nature_Nanotechnology.html

SOLUTION: ISCC (International Standard Content Code)

OSCARS

A content fingerprint computed directly from the asset — algorithmic, not assigned. It cannot be removed or decoupled from the data.

Open standard

Open source

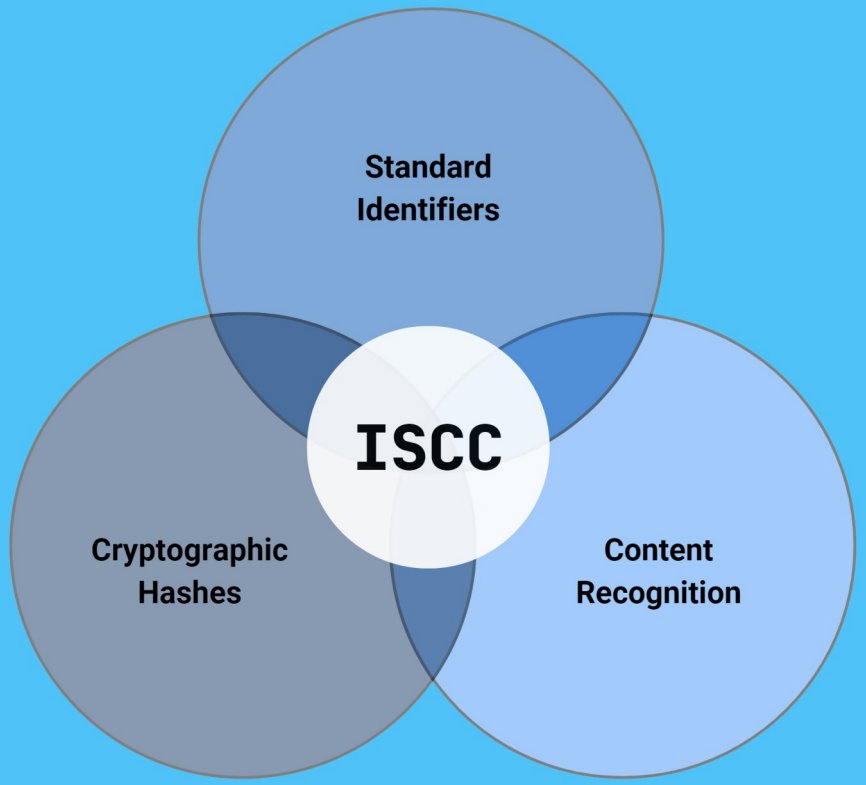
No registry needed

ISO 24138:2024

Any researcher can independently compute an ISCC from available bioimaging data - without registration, permission, or central infrastructure.

ISCC

At the Intersection of Digital Content Identification



The ISCC combines properties from multiple content identification paradigms:

- **Standard Identifiers:**

  Identifies abstract works
- **Content Recognition:**

  Clusters "*similar*" content
- **Cryptographic Hashes:**

  Verifies media asset

The DNA of your digital content

Estimate similarity using ISCC-CODEs

ISCC: KED572P4AOF5K6QXQA4T6OJD5UGX7UBPFW2TVQNTHBCKFRFCANCZARQ4K6NSFZQSH4GQ

Meta-Code	Semantic-Code	Content-Code	Data-Code	Instance-Code
AAA572P4AOF5K6QX	CEAYAOJ7HER62DL7	EEA5ALZNWU5MDMZY	GAAUJIWEUIBULECG	IAARYV43ELTBEPYN

Abstract & Persistent → Concrete & Volatile

Metadata Similarity	Semantic Similarity	Syntactic Similarity	Data Similarity	Data Integrity

Components are self-describing and can be used standalone or in combination and at different length

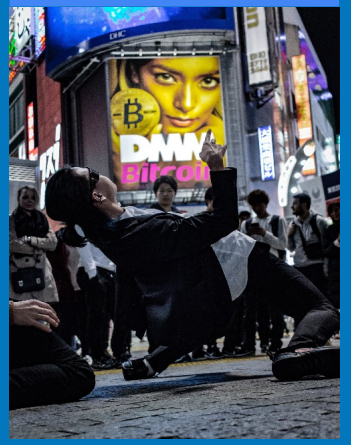
Content-Code (Image)

CCDFPFc87Mhd T	CTWAGYJ9HZGj l	CDhydSjQXDXV k	CRd5bk4SrBpzt

If we want to identify “Content” we cannot rely on “Data”:

- Two “identical” images
- Yet the data is completely different
- Due to different file formats
- ISCC Content-Code identifies information structure - not raw data

Perceptual Identification

JPG Image	JPG Data	JPG SHA1	JPG Content-ID
	49 74 27 73 20 6e 6f 74 20 61 62 6f 75 74 20 62 61 6e 6b 69 6e 67 20 74 68 65 20 75 6e 62 61 6e 6b 65 64 2e	7b 24 1f 77 f0 f2 96 df 73 b5 e0 38 97 6a 5e 3b d0 12 bd 23	CYHa5UMqq1iQS
=	≠	≠	=
PNG Image	PNG Data	PNG SHA1	PNG Content-ID
	54 68 65 20 43 75 72 72 65 6e 63 79 20 75 73 65 64 20 6f 6e 20 43 6f 62 6c 6f 20 69 73 20 43 68 61 72 6d 2e	7e bd c5 c5 c0 30 d5 4c 30 c0 31 df 4c 9e ff d5 b2 ad e8 2d	CYHa5UMqq1iQS

BIOCODES Goal

OSCARS

1. Data acquisition

Bioimaging pipeline

OMERO

2. Data analysis

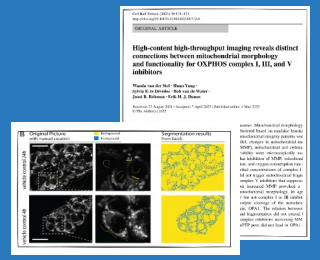
OMERO.figure

Galaxy

3. Publication and deposit at repository

BioImage Archive

IDR



ISCC Codes

Data Fingerprinting & Similarity

Raw data

- Data Similarity
- Data Integrity

Processed data & scripts

- Syntactic Similarity
- Data Similarity
- Data Integrity

Published assets

- Metadata Similarity
- Semantic Similarity
- Syntactic Similarity
- Data Similarity
- Data Integrity

- ISCC audit trail allows cryptographic data verification
- Improved data integrity, transparency, and reusability
- Enhanced AI-readiness of bioimaging datasets for trusted AI applications

BIOCODES: ISCC-SUM & ISCC-BIO

OSCARS

High-Performance Implementations for BIO-CODES

DATA / INSTANCE CODE

Able to handle any data input

- Optimized ISCC for Scientific TB-scale datasets ~1GB/s
- Implements Data & Instance code
- TREEWALK : deterministic directory hashing for large dataset collections
- ISCC-BIO : Implements the IMAGEWALK spec for deterministic Z–C–T plane traversal,
- format-agnostic, reproducible content hashing of microscopy volumes
- READY: Rust library, Python bindings, CLI tool

Implement ISCC in OMERO/ Napari / Lab Management software

OSCARS

OMERO (Available)

- Server plugin for automatic ISCC generation on image import
- Automatic ISCC generation on import
- Facility-level deduplication
- FAIR-compliant metadata annotation

OMERO.figure

Napari & Digital Lab Notebook integration (Planned)

- Plugin integration for interactive image analysis and ISCC annotation within the Napari viewer
- Interactive ISCC annotation
- In-viewer provenance display
- DLN integration for images, data and text

USE ISCC in GALAXY

OSCARS

Galaxy Tool

https://usegalaxy.eu/tools/list?search=ISCC

Preprint @BioHackrXiv with Galaxy team

https://osf.io/preprints/biohackrxiv/tsxby_v1

Galaxy Europe

Generate ISCC-CODE with ISCC-SUM

Generates an ISCC-CODE (International Standard Content Code) for datasets using the ISCC-SUM algorithm.

Show tool help

Imaging | Copy Link | Favorite Tool | Open

Verify ISCC-CODE with ISCC-SUM

Verifies that a file (dataset) matches an expected ISCC-CODE (International Standard Content Code) for exact content verification. This tool uses ISCC-SUM, which generates an ISCC-CODE containing Data-Code and Instance-Code units for bit-level file comparison.

Show tool help

Imaging | Copy Link | Favorite Tool | Open

Find datasets with similar ISCC-CODEs with ISCC-SUM

Finds similar datasets by comparing their ISCC-CODE Data-Code components using Hamming distance.

Show tool help

Imaging | Copy Link | Favorite Tool | Open

Generate ISCC-SUM codes from (imaging) data

Verify that a dataset matches an expected ISCC code

Compare codes using Hamming-distance to find similar datasets

Resources

OSCARS

Project homepage:

https://bio-codes.io/

Galaxy Tool

https://usegalaxy.eu/tools/list?search=ISCC

Image SC (Community forum TBA)

ISCC Foundation

Universiteit Leiden

NL-BIOIMAGING AM

German BioImaging

Gesellschaft für Mikroskopie und Bildanalyse

Use cases in bioimaging

01

Duplicate & near-duplicate detection

Find redundant datasets across IDR, EMPIAR, and BioImage Archive — including reprocessed, reformatted, or recropped versions that byte-level checksums miss entirely.

02

Data integrity & transfer verification

High-speed (~4 GB/s) Instance-Code fingerprinting for verifying file integrity during ingestion of multi-hundred-GB Zarr, HDF5, and OME-TIFF files at scale.

03

Reproducibility & dataset citation

Bind ISCC codes at submission time so downstream researchers verify exact data identity (Instance-Code) or a known derivative (Data-Code) — a core FAIR reproducibility requirement.

04

Version tracking across processing stages

Link raw → denoised → segmented → pseudo-colored outputs via Data-Code similarity, making the full provenance chain discoverable without manual annotation.

05

AI training data provenance & synthetic image flagging

ISCC fingerprints enable training corpus audits for synthetic contamination and repository-level flagging of AI-generated micrographs.