Introduction to the International Standard Content Code (ISCC)

ISO 24138:2024

ISCC - ISCC Discovery Protocol / NL-BioImaging

2025-09-30, *Martin Etzrodt*

ISCC

Who is the ISCC Foundation

The **ISCC Foundation** is a non-profit, dedicated to developing, standardizing and promoting open-source technology for universal content identification.

Titusz Pan
Chairman

Kira Lemke
Director

Martin Etzrodt
Director

Todd Carpenter
Advisory Board

Roanie Levy
Advisory Board

Giacomo D'Angelo
Advisory Board

Frank Shulleri
Advisory Board

Sebastian Posth
Advisory Board

Lambert Heller
Advisory Board

Philippe Rixhon
Advisory Board

TC 46 - Information and Documentation
SC 09 - Identification and Description

ISCC

ISO 24138:2024

Conceived
2016-06-29

WG 18 Started
2019-10-29

Published
2024-05-15

ISO 2108:2017	ISBN	International Standard Book Number
ISO 3297:2022	ISSN	International Standard Serial Number
ISO 3901:2019	ISRC	International Standard Recording Code
ISO 15706-1:2023	ISAN	International Standard Audiovisual Number
ISO 15707:2022	ISWC	International Standard Musical Work Code
ISO 27729:2012	ISNI	International Standard Name Identifier
ISO 26324:2022	DOI	Digital Object Identifier System
ISO 24138:2024	ISCC	International Standard Content Code

Why is ISO 24138:2024
a significant publication?

- Open-source and interoperable content identification & fingerprinting system
- One universal code for digital text, image, audio and video
- Cross-sector standard (publishing, science, arts, etc.)
- A neutral, transparent and interoperable content identification system

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

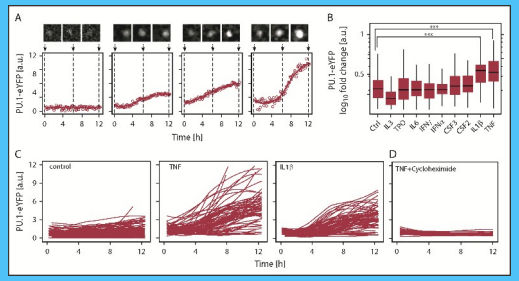
LEIDEN CELL OBSERVATORY

Funded by the European Union

Motivation

Acquisition

Storage/ sharing



Data analysis & interpretation

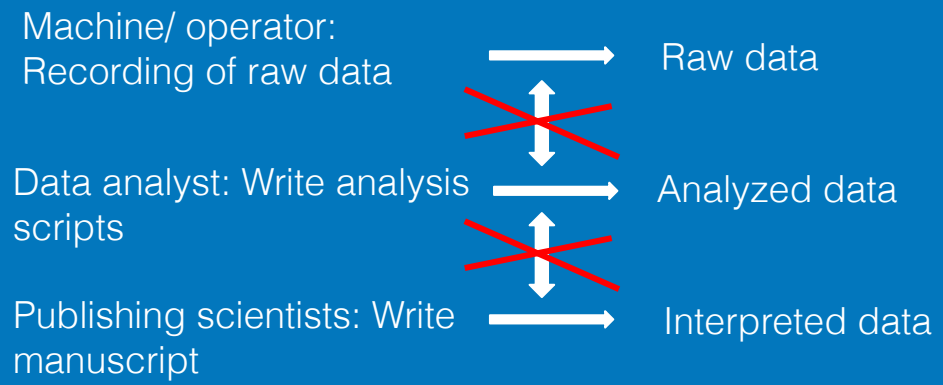
It is often impossible to recapitulate the exact way of how and what raw imaging data went into a publication.

Consequences:

- unnecessary friction (reproducibility)
- wasted resources (reusability)
- fraud, deep fakes…

Motivation

It should be possible to create a full “audit-trail” from image acquisition to final data output.



Currently the connection is broken
Content derived identifiers could provide a solution

Biocodes goals

1. Evaluate ISCC Standard for Bioimaging
   - Assess ISCC for bioimaging data
   - Create proof-of-concept implementation
2. Implement Content-Derived Identification
   - Deploy ISCC generator service at Leiden Cell Observatory
   - Develop data integrity verification and provenance tracking
3. Integrate with Existing Infrastructure
   - Add ISCC codes to OMERO workflows
   - Integrate in bioimage repositories (IDR/ Bioimage archive)

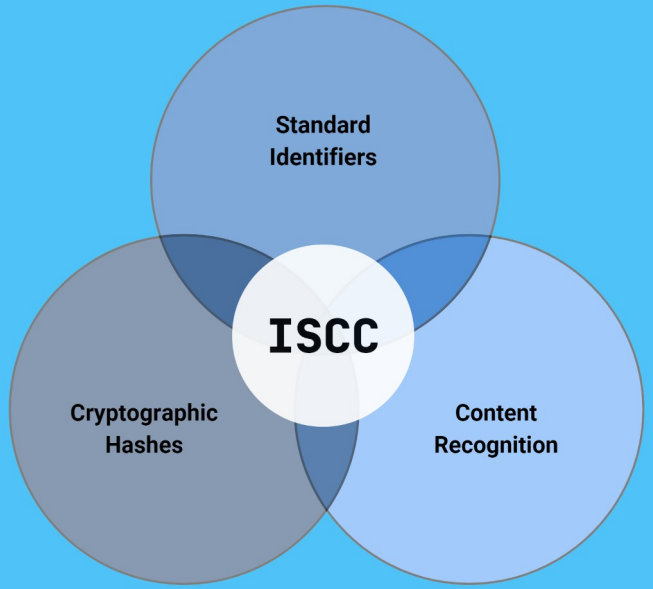
Motivation

Choice of identifier for bio(image) data?

		Content based identification	Timestamp	Robust to changes/ similarity	Can regenerate/ store
1	Standard identifiers (DOI)	✗	✗ ✓	✗	No, external registry/ handle system
2	UUID (v4)	✗	✓	✗	No, every time different
3	Checksums	✓	✗	✗	Yes - deterministic
4	Cryptographic hash (sha256)	✓	✗	✗	Yes - deterministic
5	ISCC	✓	✓	✓	Yes - deterministic

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

CCDFPFc87Mhd T | CTWAGYJ9HZGj l | CDhydSjQXDXV k | CRd5bk4SrBpzt

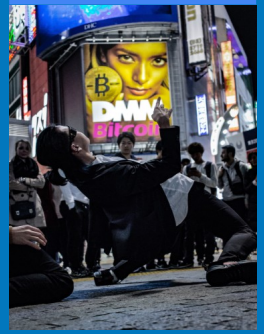
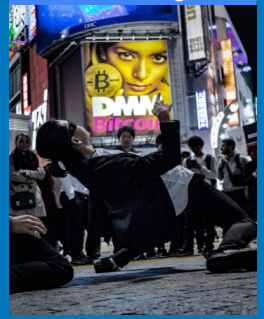
Content-Code (Image)

Similarity hash over normalized generic data. Self-Describing and media-type specific.

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

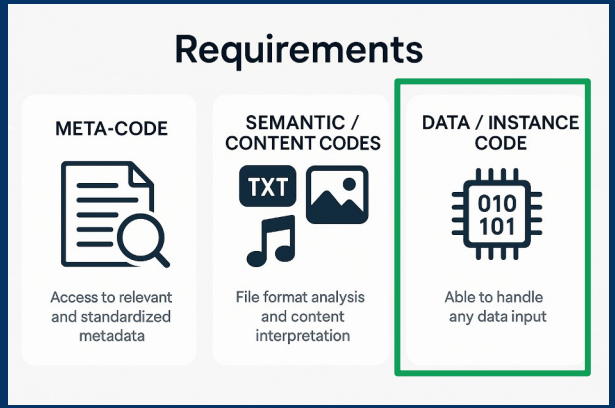
- Two “identical” images
- Yet the data is completely different
- Due to different file formats
- ISCC encodes information structure - not raw data

Layer 3 - 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

ISCC-SUM

Key Features for BIO-CODES

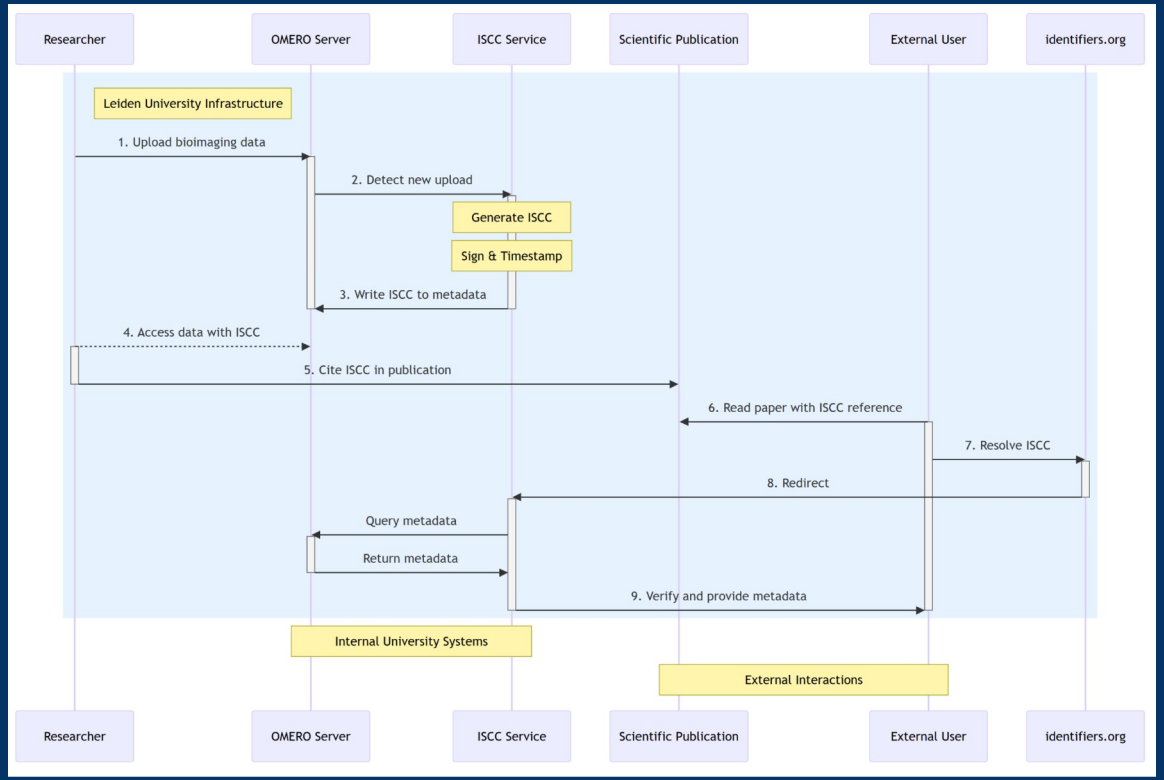


- Handles large proprietary imaging formats efficiently
- Handles container formats
- Cross-platform (Linux, Mac, Windows):
  Works on all lab computers and servers
- Simple command-line interface
  (like md5sum, sha1sum, sha256)
- Python integration for use in OMERO/analysis pipelines

Helps verify data integrity and track lineage through pipelines & formats

BioCodes process flow

Signing, Timestamping, and Discovery



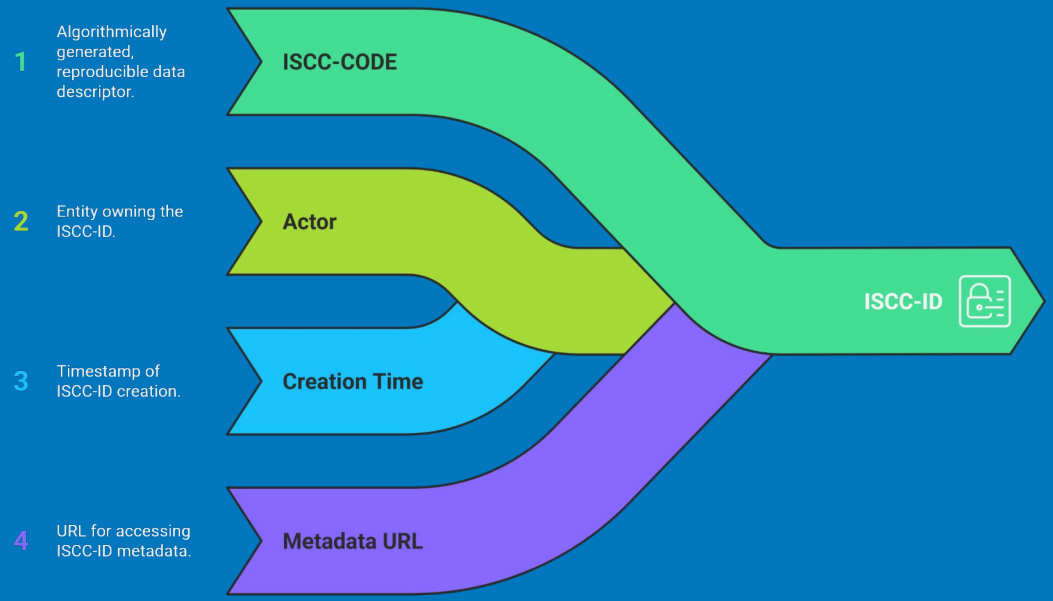
ISCC

What does ISO 24138:2024 NOT standardize (yet)?

- **Who** created an ISCC
- **When** was an ISCC created
- **Where** do I find metadata

ISCC-ID - A new ISCC Type

for binding data, actor, time and metadata



ISCC Discovery Protocol

Signing, Timestamping, Discovery

Minimal IsccNote

```
{
  "iscc_code": "ISCC:KACWN77F73NA44D6EUG3S3QW6Z…",
  "datahash": "1e205ca7815a69c1d53017638eb52803…",
  "nonce": "000faa3f18c7b9407a48536a9b00c4cb",
  "timestamp": "2025-01-15T12:00:00.000Z",
  "signature": {
    "version": "ISCC-SIG v1.0",
    "pubkey": "z6MkmeDbeC5BecFmVnTHAaVUrGLd…",
    "proof": "z5j9nrpPw3oYSAN4XsUtkwrueTD6V…"
  }
}
```

Cryptographic Public Key

Owner's key ensures secure and verifiable ownership.

Timestamp

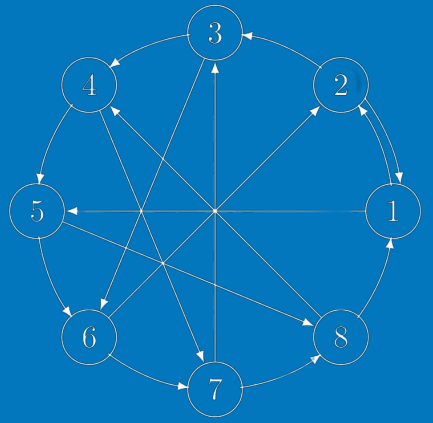
Timestamp provides proof of existence and creation time.

URL Location

URL links to metadata and associated services.

Digital Content

Content represented by ISCC-CODE for identification.



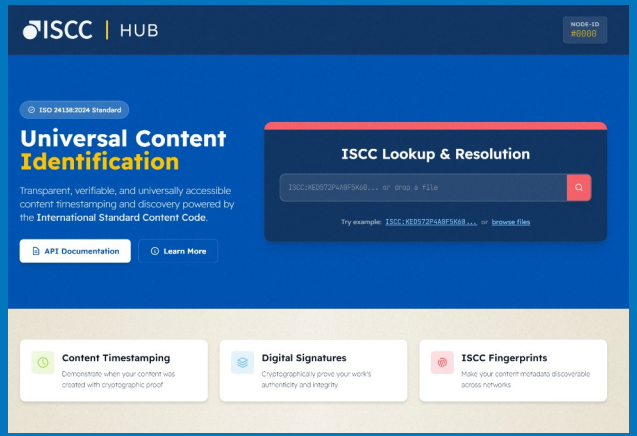
Scalable content based witness and discovery network

ISCC

Current Development

ISCC Discovery Protocol (IDP)

ISCC-HUB - An open-source web service implementation of the IDP



A neutral & transparent infra for:

Fingerprinting

help others find your metadata based on content matching

Digital Signing

cryptographically prove authenticity of data

Timestamping

demonstrate when your data was first created

The ISCC Discovery Protocol creates a federated infrastructure of neutral core services for efficient and interoperable **discovery of content-related metadata**, services and actors.

ISCC

Biocodes

Results (thus far)

- ISCC-SUM - High-Performance ISCC Implementation for bio (imaging) data
- OMERO & Napari integrations in progress
- A prototype ISCC Hub for the discovery protocol is operational