BIO-CODES

Enhancing AI-Readiness of Bioimaging Data

ISCC Introduction

2024-11-29, *Titusz Pan*

About me I Titusz Pan

- Founder and CEO of Craft AG, Freiburg
- Open-Source Software Developer
- Expert at ISO/TC 46/SC 9/WG 18
- Chairman at non-profit ISCC Foundation

Who is the ISCC Foundation

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

Titusz Pan
Chairman

Kira Lemke
Treasurer

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

ISCC Foundation
Activities

- Content Identification Research
- Open source development
- IT infrastructure for ISCC services
- Promoting ISCC adoption
- Governance of the ISCC

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

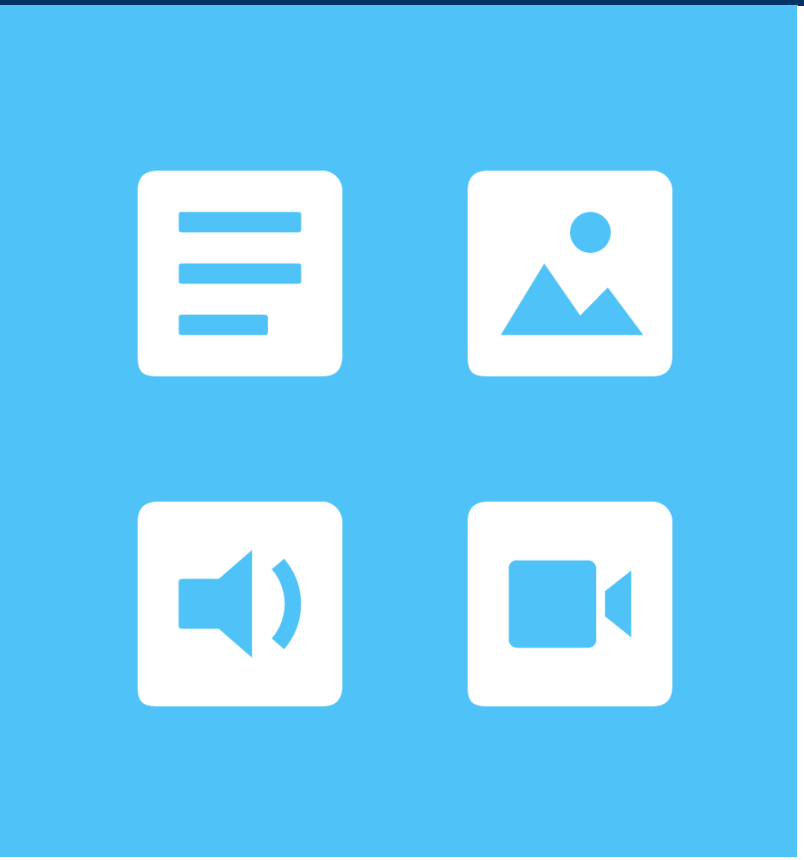
ISO 2108:2017	ISBN	International Standard Book Number
ISO 3297:2022	ISSN	International Standard Serial Number
ISO 3901:2019	ISRC	International Standard Recording Code
ISO 15706-1:2023	ISAN	International Standard Audiovisual Number
ISO 15707:2022	ISWC	International Standard Musical Work Code
ISO 27729:2012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

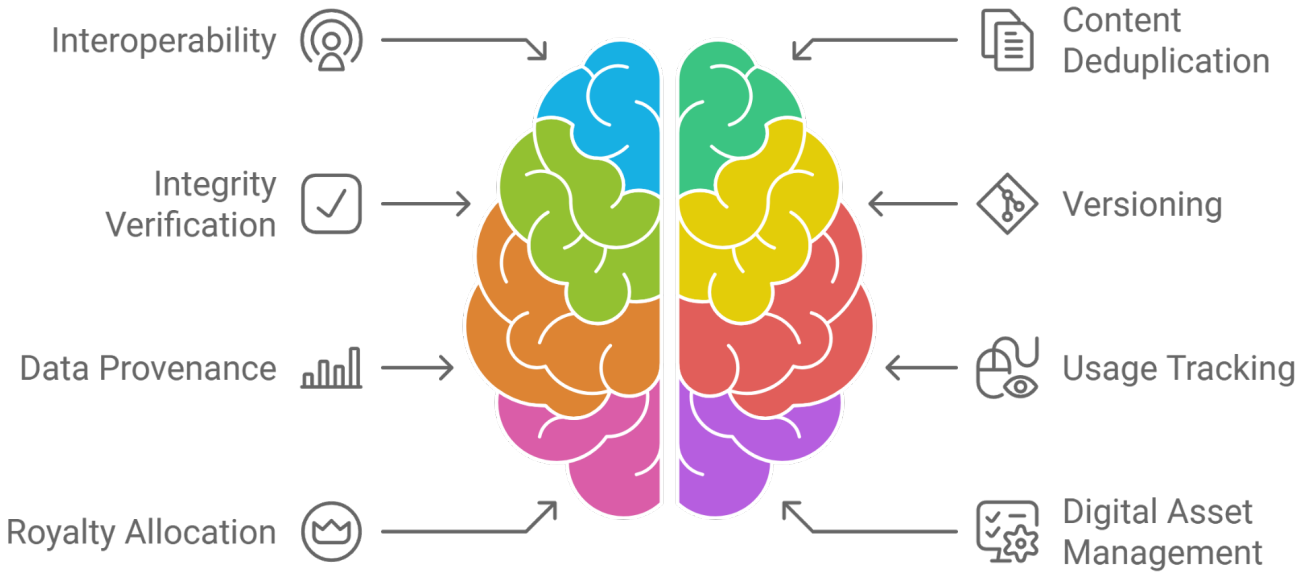
Why is ISO 24138:2024
a significant publication?

- 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 global system

ISCC

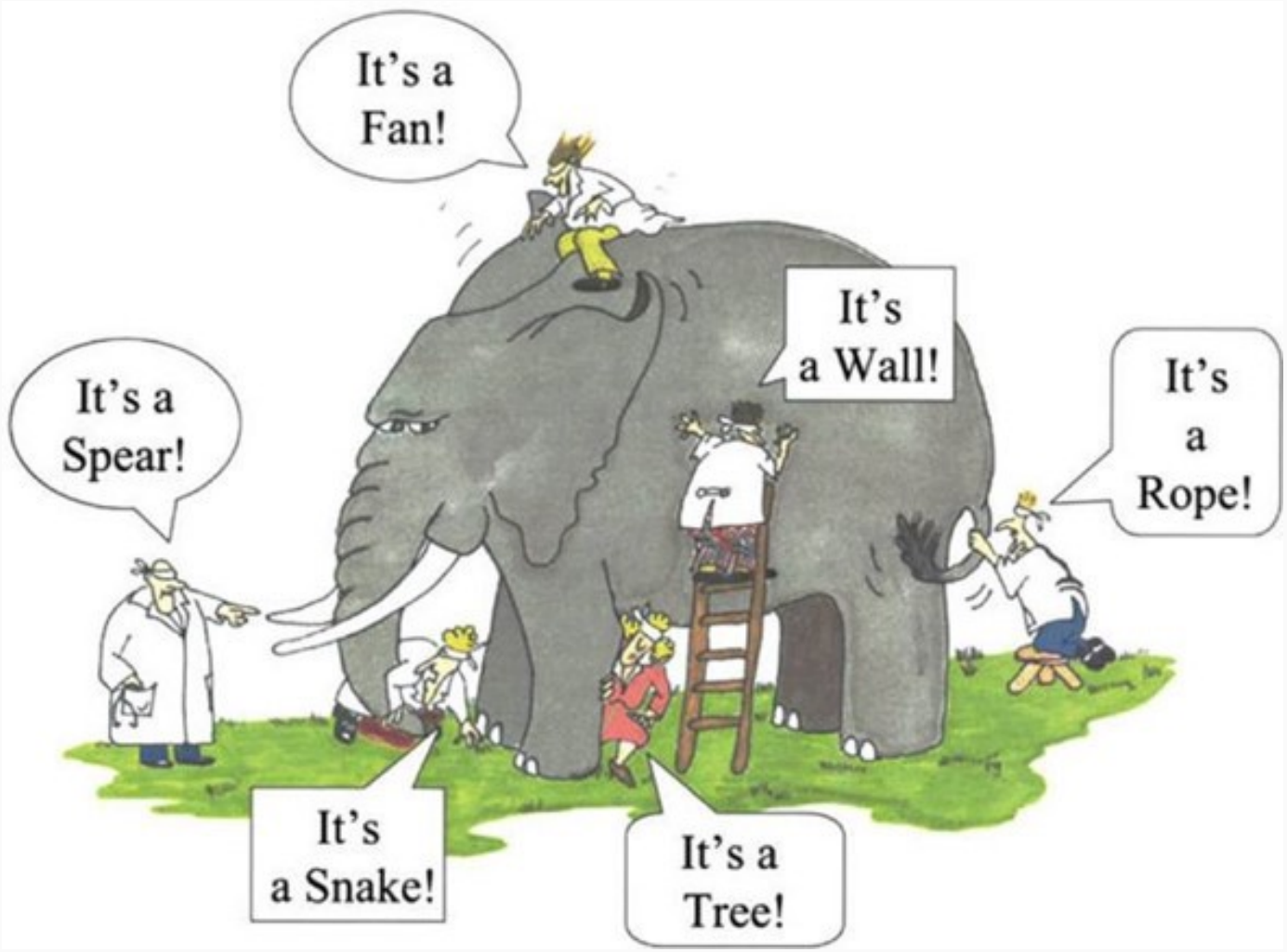
Key Benefits of Implementing ISCC





What is **Not** ISCC

- A persistent identifier
- A content recognition system
- A cryptographic hash
- A content registry



Layers of “Content” Identification

In our model for digital content identification we distinguish 6 layers that exist naturally on a scale from abstract to concrete.

Existing content identifiers usually operate on one or two of these layers.

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

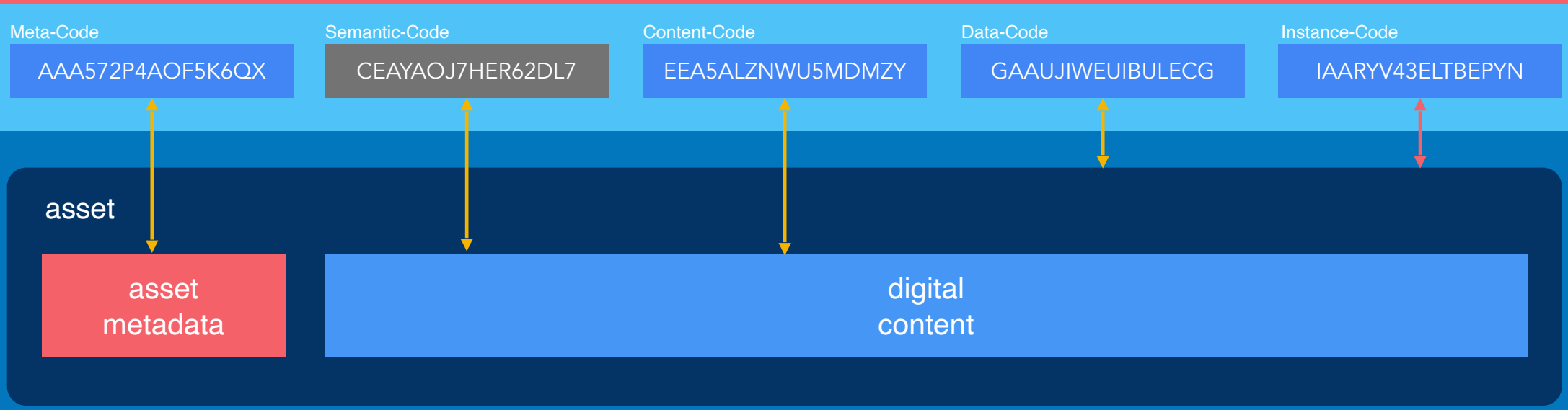
Metadata Similarity	Semantic Similarity	Syntactic Similarity	Data Similarity	Data Integrity

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

ISCC & Content Binding

a binding of bindings

ISCC:KED572P4AOF5K6QXQA4T6OJD5UGX7UBPFW2TVQNTHBCKFRFCANCZARQ4K6NSFZQSH4GQ



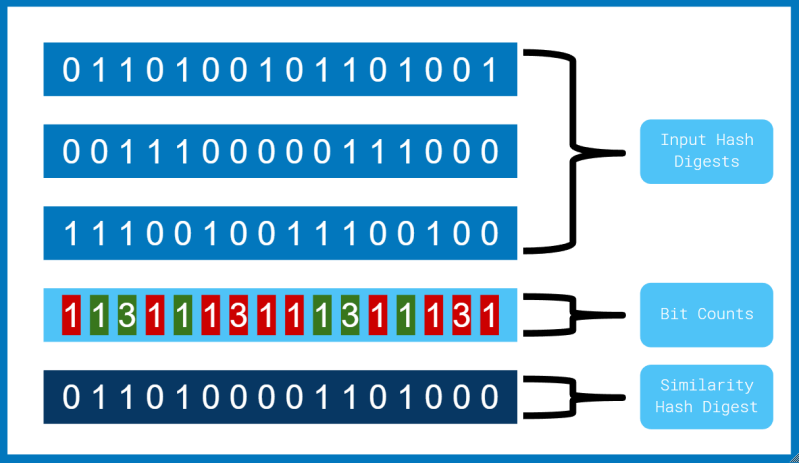
Yellow arrows are soft bindings, red arrows are hard bindings.
Additionally ISO 24138:2024 defines a Meta-Hash algorithm which is a secure hard binding to be part of ISCC metadata.

CCDFPFc87MhdT | CTWAGYJ9HZGj1 | CDhydSjQXDXVk | CRd5bk4SrBpzt

Meta-Code

A similarity preserving hash over metadata.

ISCC - Similarity Hash Diagram



Layer 1 - Abstract Identification

The Meta-Code is seeded from Metadata

name: Title for content or work or series (max 128 bytes)

description: Description of the identified content (max 4096 bytes)

meta: Subject, industry, or use-case specific metadata, encoded as RFC-2397 Data-URL (max 16384 bytes).

- Identifies at any desired level of abstraction (series, work ...)
- Top level of grouping a content collection or hierarchy
- Independent of digital manifestations
- Supports *progressive disambiguation*
- Should be embedded into digital assets for reproducibility
- Requires minimal metadata

Seed Metadata is metadata that is used to establish a Meta-Code and stays frozen (immutable) throughout its existence. Floating Metadata is any mutable metadata that is managed in context with an **ISCC**.

CCDFPFc87MhdT | CTWAGYJ9HZGj1 | CDhydSjQXDXVk | CRd5bk4SrBpzt

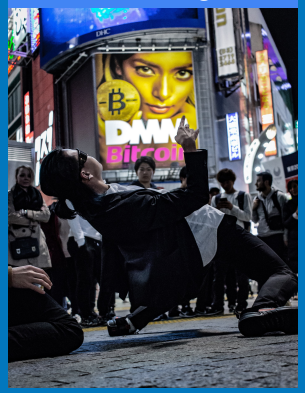
Content-Code (Image)

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
- Content-ID 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

CCDFPFc87MhdT | CTWAGYJ9HZGj1 | CDhydSjQXDXVk | CRd5bk4SrBpzt

Data-Code

Similarity over raw encoded data.

- Identifies encoded content
- Clusters file versions
- Spectrum of tolerance
- Shift resistant chunking (CDC)
- **Similarity hash (tree) over variable sized chunk hashes**

Layer 4 - Encoded Manifestation

Fixed Size Data Chunking

AAAAIBBBBICCCCIDDDDI

EAAAIABBBIBCCCICDDDID

Content Defined Chunking - Shift Resistant - Variable Size Chunks

AAAAIBBBBICCCCIDDDDI

EAAAAIBBBBICCCCIDDDDI

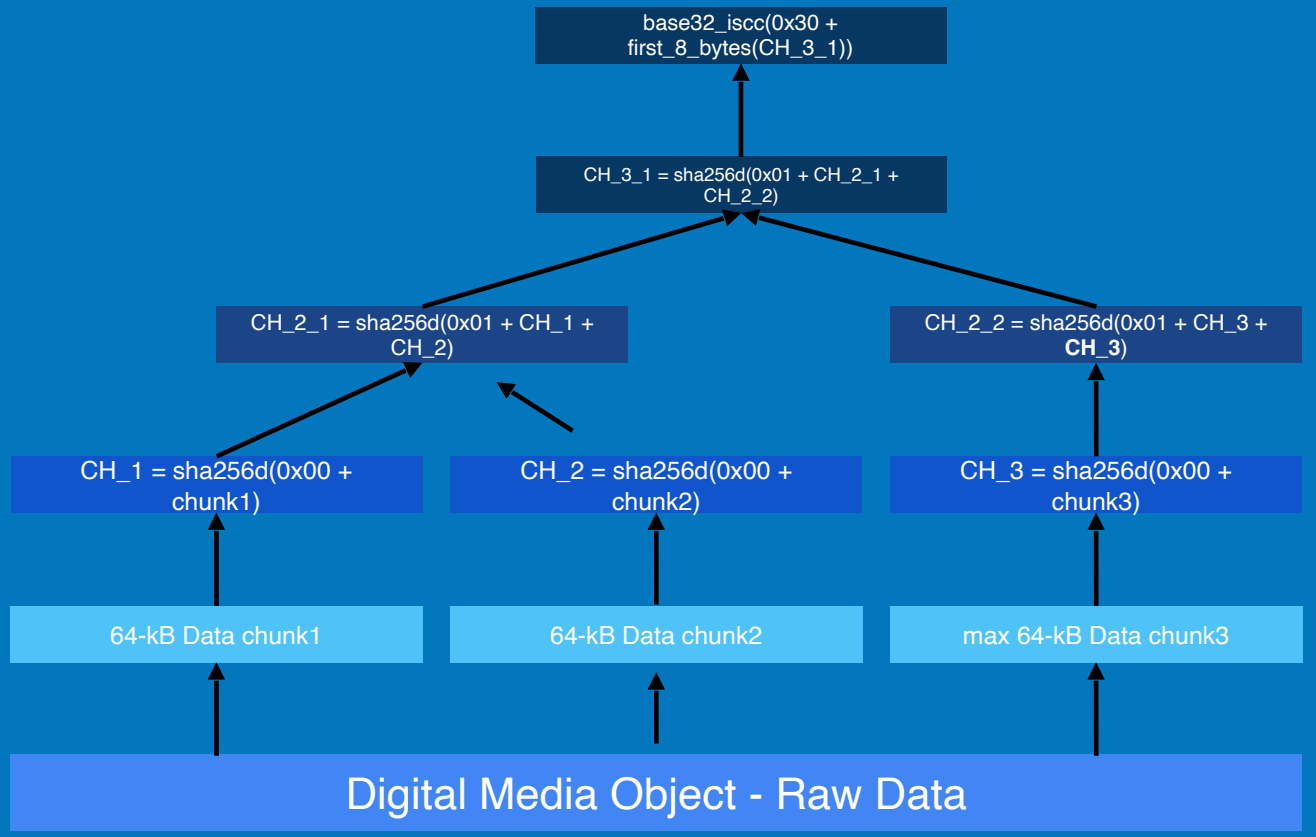
CCDFPFc87MhdT | CTWAGYJ9HZGj1 | CDhydSjQXDXVk | CRd5bk4SrBpzt

Instance-Code

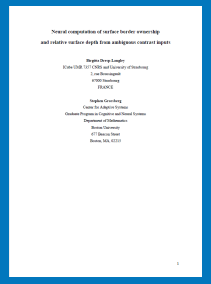
Cryptographic hash. The root of a hash tree over raw data.

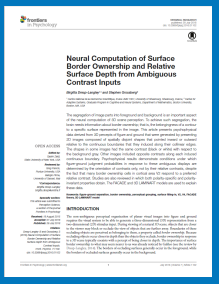
- Precise data identification
- Proof of data containment
- Separate Tophash (256 bit)
- Data integrity (via tophash)

Layer 5 - Exact Representation



Example one **DOI** multiple matching **ISCC**





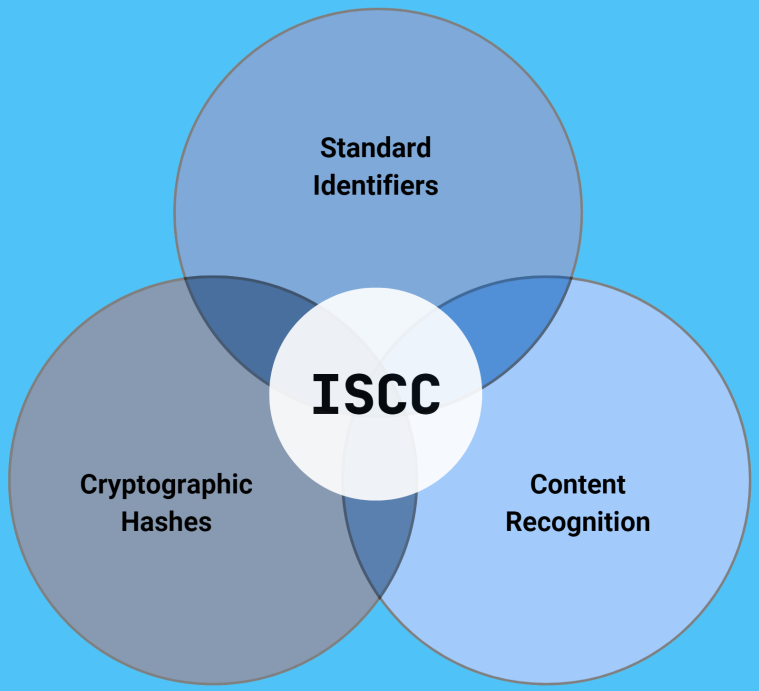
Paper: Neural Computation of Surface Border Ownership and Relative Surface Depth from Ambiguous Contrast Inputs

Host	DOI	ISCC
hal.archives-ouvertes.fr	10.3389/fpsyg.2016.01102	CCDyud5ZWAkDR-CTTq25WFQTWaU-CDbUZg6v3qzzM-CRrxfuPk2nP3Q
arxiv.org	10.3389/fpsyg.2016.01102	CCDyud5ZWAkDR-CTTRs5cQY1D11-CDPqUxrqN7YRx-CRcUmq2SmgN18
hal.archives-ouvertes.fr	10.3389/fpsyg.2016.01102	CCDyud5ZWAkDR-CTfNotD3KMMd1-CD481J7LDBQPH-CR8rZ9QzTzJRL
frontiersin.org	10.3389/fpsyg.2016.01102	CCDyud5ZWAkDR-CTfNotD3KMMd1-CDMXxzVp63Mpt-CRZ5iRuFkENb7

Estimated Similarity of Meta-Code: 100.00 %
Estimated Similarity of Content-Code Text: 84.38 %
Estimated Similarity of Data-Code: 53.12 %

ISCC

At the Intersection of Digital Content Identification



The ISCC combines properties from multiple content identification paradigms:

- **Standard Identifier:**

  Unique media asset indentification
- **Content Recognition:**

  Clusters “*similar*” content
- **Cryptographic Hashes:**

  Verifies media asset integrity

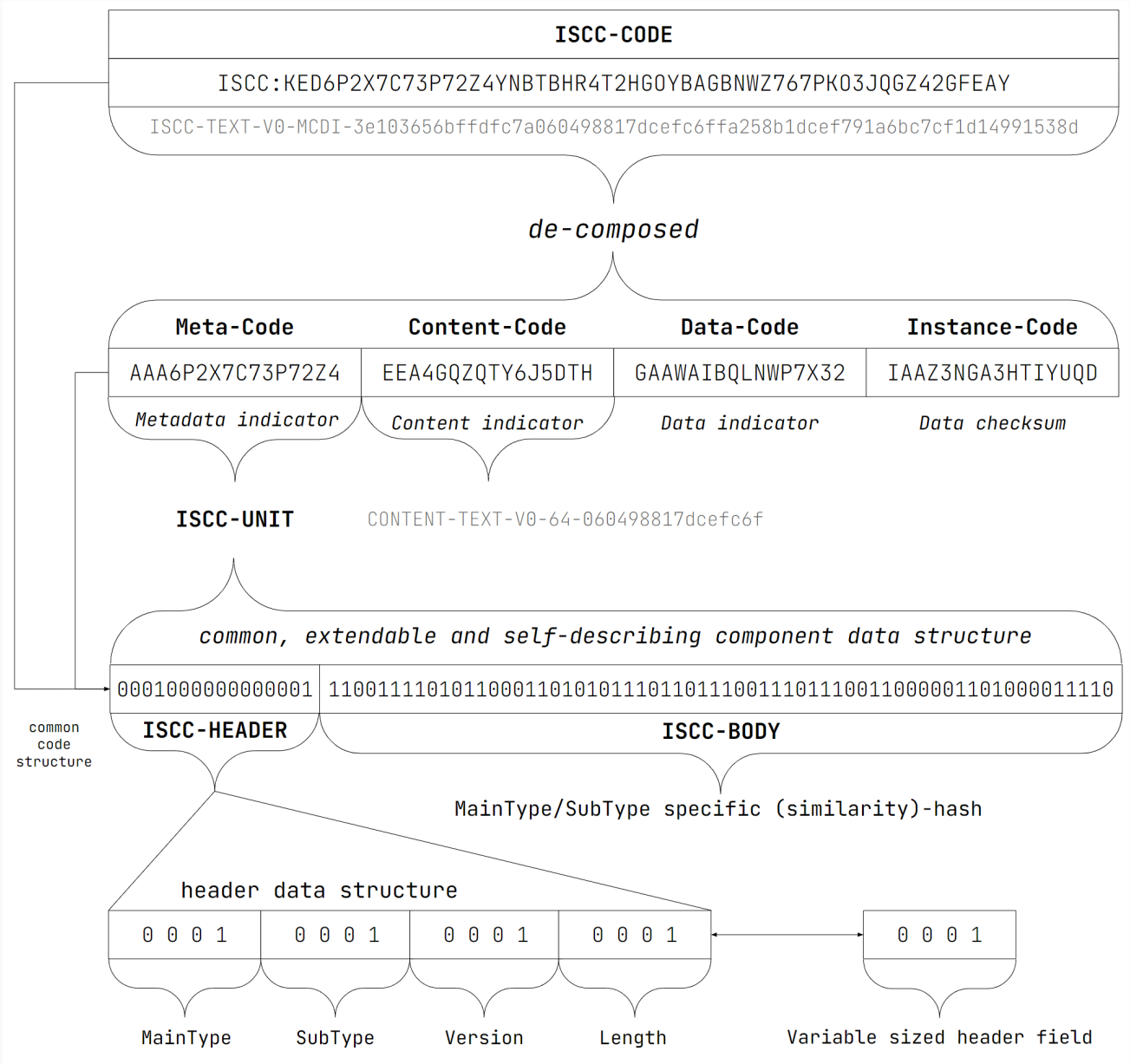
Content Binding

Connecting Content with Metadata

01	Standard Identifiers	• Registration based (bound by database entry) • Abstract “Binding” via descriptive metadata • Assignment requires third-party
02	Watermark (Soft Binding)	• Hides an identifier within the data (proprietary) • Can be removed/changed if algorithm is known • Supports audiovisual content only (no text)
03	Fingerprint (Soft Binding)	• Calculated from the digital asset • Works for all modalities and cannot be removed • Not statistically unique
04	Cryptohash (Hard Binding)	• Calculated from the digital asset • Works for all data • Strong uniqueness guarentee

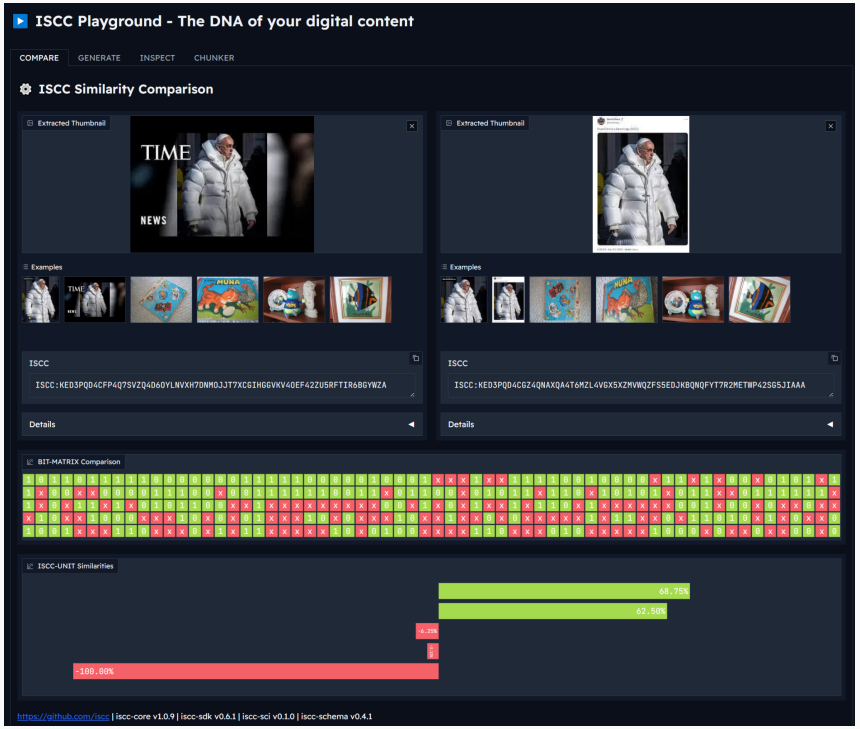
ISCC - Codec

The ISCC is based on a compact, machine-readable, self-describing, modular and extensible codec.



Demo Time :)
Playground

- An interactive playground
- Explore and visualize **ISCC**
- Helps developers and users
- Showcase future updates
- Published on Huggingface



https://huggingface.co/spaces/iscc/iscc-playground

ISCC

Give your DOI superpowers

Bind digital content to a DOI

Timestamps

demonstrate when your content was first created

Signatures

cryptographically prove your work's authenticity

Fingerprints

help others find your book metadata and licensing infos

ISCCs connect your paper to a DOI in a granular, verifiable and location independent way. Individual paragraphs and images can be resolved to a DOI and associated metadata.

Digital Reality

There is too much/
granular
content to manually
assign and track
identifiers.

Good News

All your

already has an ISCC.
It just needs to be
computed.

ISCC

Thank you
for your attention

Contact Information:

Titusz Pan

tp@iscc.io