International Standard

ISO 24138

Information and documentation — International Standard Content Code (ISCC)

First edition

PROOF/ÉPREUVE

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/PRF 24138
https://standards.iteh.ai/catalog/standards/iso/4bd4b8ba-b648-409e-8713-22cb3cbbe602/iso-prf-24138

COPYRIGHT PROTECTED DOCUMENT

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 46, *Information and documentation*, Subcommittee SC 9, *Identification and description*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

While ISO/TC 46/SC 9 has established a variety of specific identifier standards, a content-dependent identifier for digital assets in all content formats has not yet been agreed.

Digital content is dynamic, always in motion, and acted upon globally by a variety of entities with different interests and requirements. Digital content continuously re-encodes, resizes, and re-compresses, changing its data as it travels through a complex network of actors and systems.

The International Standard Content Code (ISCC) is an identifier for numerous types of digital assets. An ISCC-CODE is generated from the digital content itself. It is the result of processing the digital content using a variety of algorithms including hash algorithms. The generated ISCC-CODE supports data integrity verification and preserves an estimate of the data, digital content and metadata similarity. However, ISCC has different functionality from content recognition systems.

The ISCC supports the association of higher-level identifiers (like work and product identifiers) with the digitally encoded manifestations of content. The ISCC does not specify a system for managing authoritative metadata. Other content identifier standards can use ISCC to support discoverability of their identifiers and metadata based on digital content.

Organizations, individuals and machines may generate ISCCs for numerous kinds of digital assets and use them for identification and management of those assets.

ISCCs are neither manually nor automatically assigned to digital media assets. Instead, ISCCs are derived from media assets according to the procedures described in this document. Unrelated parties can independently derive the same ISCC from a given media asset.

ISCCs exclusively reference media assets without any implication about ownership. As such, ISCCs are not managed authoritatively by any institution or entity.

The ISCC enables interoperability between different actors and systems using digital assets and supports scenarios that require content deduplication, database synchronization and indexing, integrity verification, timestamping, versioning, data provenance, similarity clustering, anomaly detection, usage tracking, allocation of royalties, fact-checking and general digital asset management use-cases.

This document includes sections targeting a general audience but also descriptions of more technical procedures.

Future editions of this document can be developed as outlined in [Annex C](#).

Information and documentation — International Standard Content Code (ISCC)

1 Scope

This document specifies the syntax and structure of the International Standard Content Code (ISCC), as an identification system for digital assets (including encodings of text, images, audio, video or other content across all media sectors). It also describes ISCC metadata and the use of ISCC in conjunction with other schemes, such as DOI, ISAN, ISBN, ISRC, ISSN and ISWC.

An ISCC applies to a specific digital asset and is a data-descriptor deterministically constructed from multiple hash digests using the algorithms and rules in this document. This document does not provide information on registration of ISCCs.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 10646:2020, *Information technology — Universal coded character set (UCS)*

ISO/IEC 15938, *Information technology — Multimedia content description interface*

ISO/IEC 21778, *Information technology — The JSON data interchange syntax*

IETF RFC 4648, *The Base16, Base32, and Base64 Data Encodings*[1]

IETF RFC 2397, *The "data" URL scheme*[2]

IETF RFC 8785, *JSON Canonicalization Scheme (JCS)*[3]

W3C, C14N 1.1, *Canonical XML Version 1.1*[4]

W3C, JSON-LD 1.1, *A JSON-based Serialization for Linked Data*[5]

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

---

[1] Online available: https://datatracker.ietf.org/doc/html/rfc4648

[2] Online available: https://datatracker.ietf.org/doc/html/rfc2397

[3] Online available: https://datatracker.ietf.org/doc/html/rfc8785

[4] Online available: https://www.w3.org/TR/xml-c14n11

[5] Online available: https://www.w3.org/TR/json-ld/

3.1
bit
atomic unit of information in a computer system

3.2
byte
sequence of 8 *bits* ([3.1](#))

3.3
nibble
half a *byte* ([3.2](#)), which can be represented by a single hexadecimal digit

[SOURCE: ISO 20038:2017, 3.12]

3.4
data
ordered sequence of *bits* ([3.1](#))

3.5
file
stored *data* ([3.4](#)) with a known number of *bits* ([3.1](#)) and a filename

3.6
stream
data ([3.4](#)) in transit with a known or unknown number of *bits* ([3.1](#))

3.7
content
information organized to provide value to a user

3.8
digital content
manifestation of *content* ([3.7](#)) in form of *data* ([3.4](#)) structured according to a set of rules

3.9
metadata
data ([3.4](#)) that defines and describes other data

[SOURCE: ISO 24531:2013, 4.32]

3.10
seed metadata
initial *metadata* ([3.9](#)) used as input to a *hash algorithm* ([3.1](#)) function

3.11
content format
set of rules used to structure *digital content* ([3.8](#))

3.12
media type
two-part *identifier* ([3.15](#)) specifying the nature of the referenced *data* ([3.4](#))

[SOURCE: ISO/IEC 19757-4:2006, 3.9]

3.13
digital asset
file ([3.5](#)) or *stream* ([3.6](#)) encoded in conformance with a specific *content format* ([3.11](#))

3.14
referent
object which is identified

3.15
identifier
sequence of characters that uniquely denotes a *referent* (3.14)

3.16
identifier system
system to enable the provision of *identifiers* (3.15) for a given category of *referents* (3.14)

3.17
content identifier
identifier (3.15) whose *referent* (3.14) is *content* (3.7)

3.18
content-dependent identifier
content identifier (3.18) whose *data* (3.4) depends on the *digital content* (3.8) that it identifies

3.19
content recognition system
system whose primary purpose is to recognise *digital content* (3.8) on a granular level

3.20
algorithm
set of instructions

3.21
hash algorithm
deterministic *algorithm* (3.20) that produces fixed-length *data* (3.4) from an input of arbitrary-length data

3.22
hash digest
result of processing *data* (3.4) with a *hash algorithm* (3.21)

3.23
cryptographic hash function
computationally efficient function mapping binary strings of arbitrary length to binary strings of fixed length, such that it is computationally infeasible to find two distinct values that hash into a common value

3.24
similarity hash
hash digest (3.22) that preserves correlations between inputs to the *hash algorithm* (3.21)

3.25
content defined chunking
CDC
method to split *data* (3.4) into variable length chunks based on internal features such that chunk boundaries are more resistant to *byte* (3.2) shifting

3.26
actor
human or non-human (hardware or software) entity that interacts with a system

3.27
Merkle tree
tree data structure in which every leaf node is labelled with the *hash digest* (3.22) of a data element and every non-leaf node is labelled with the hash digest of the labels of its child nodes

3.28
Merkle root
root node of a *Merkle tree* (3.27)

[SOURCE: ISO 22739:2024, 3.57]

3.29
ISCC processor
application that generates ISCCs for *digital content* ([3.8](#))

3.30
plain text
data ([3.4](#)) with a known text encoding that can be transcoded to Unicode

3.31
whitespace
nondisplaying formatting characters such as spaces, tabs, etc., that are embedded within a block of free text

[SOURCE: ISO/IEC/IEEE 31320-2:2012, 3.1.210]

4 Structure and format of the ISCC

4.1 General structure

a) An ISCC shall be composed of an ISCC-HEADER and an ISCC-BODY (see [Figure 1](#)).

b) The ISCC-HEADER shall describe the MainType, SubType, Version, and Length of its ISCC-BODY.

c) An ISCC-UNIT shall be an ISCC based on one specific algorithm.

d) An ISCC-CODE shall be an ISCC composed from two or more different ISCC-UNITs.

ISCC-CODE

ISCC:KEC2LO3CU6XL6ZVXS3G5LUWUZDEJNBWUBDZPXNR542DDKYVPG372D2A

ISCC-HEADER	ISCC-BODIES (concatenated)
0101000100000101	10100101101110110110000101010011110101110101111110110011010110...
ISCC-IMAGE-V0-MCDI	a5bb62a7aebf66b796cdd5d2d4c8c89686d408f2fbb63de6863562af36ffa1e8

composed / decomposed (as per algorithm)

Meta-Code	Content-Code	Data-Code	Instance-Code
AAA2LO3CU6XL6ZVX	EEAZNTOV2LKMRSEW	GAAYNVAI6L53MPPG	IAAYMNLCV43P7IPI
similarity hash	*similarity hash*	*similarity hash*	*data checksum*

ISCC-UNIT   CONTENT-IMAGE-V0-64-96cdd5d2d4c8c896

common, extendable and self-describing component data structure

0000000000000001	10100101101110110110000101010011110101110101111110110011010110111
ISCC-HEADER	**ISCC-BODY**

common code structure

MainType/SubType specific (similarity)-hash

header data structure

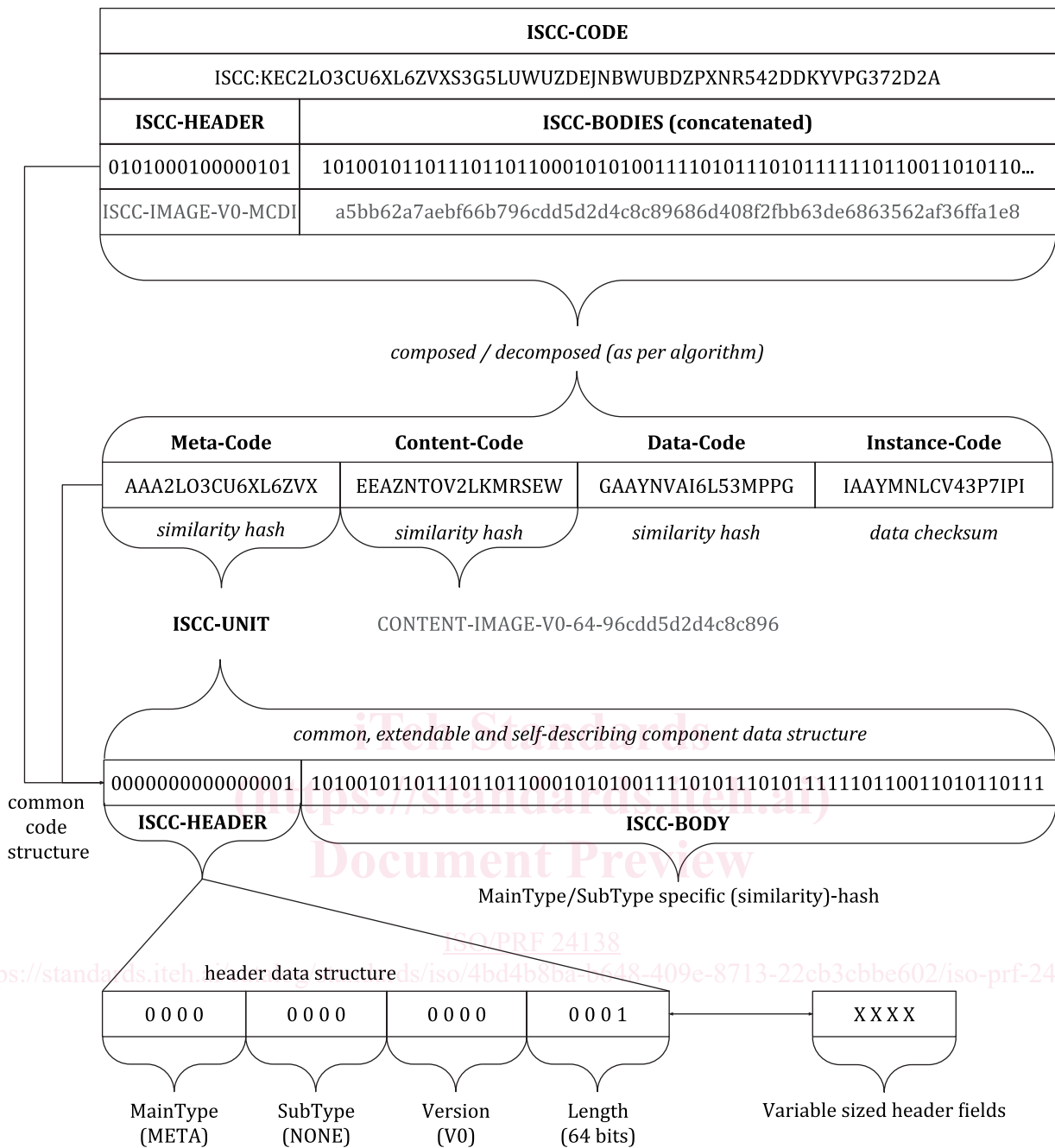| 0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 1	X X X X
MainType (META)	SubType (NONE)	Version (V0)	Length (64 bits)	Variable sized header fields

Figure 1 — General structure of an ISCC

The concatenation of the ISCC-UNITs is based on the underlying data and is not visible in the string representation of the ISCC-CODE itself. See 6.6 on how an ISCC-CODE is composed from individual ISCC-UNITs.

4.2   ISCC-HEADER

4.2.1   General

4.2.1.1   The ISCC-HEADER is a variable sized bitstream composed of an ordered sequence of the 4 header-fields MainType, SubType, Version, Length.

4.2.1.2   Each header-field is a bitstream with a length between 4 and 16 bits and encodes an integer value between 0 and 4679 (see Table 1) with the following encoding scheme.

a)   The total bit-length of a header-field shall be determined by its prefix-bits.

b)   The prefix-bits shall be followed by data-bits.

c)   The data-bits shall be interpreted as unsigned integer values plus the maximum value of the preceding range.

d)   If the total length of all header fields in number of bits is not divisible by 8, the header shall be padded with 4 zero bits (0000) on the right side.

Table 1 — Variable length ISCC-HEADER field encoding

Prefix bits	Number of nibbles	Number of data bits	Integer range
0	1	3	0-7
10	2	6	8-71
110	3	9	72-583
1110	4	12	584-4679

4.2.1.3   The interpretation of the integer value of a header-field shall be context dependent.

a)   For the MainType and SubType fields, it shall be an identifier for the designated type (see 4.2.2 and 4.2.3).

b)   For the Version field it shall be the literal version number (see 4.2.4).

c)   For the Length field of ISCC-UNITs, it shall be a number used as a multiplier to calculate the bit length of the ISCC-BODY (see 4.2.5, Table 6).

d)   For the Length field of ISCC-CODEs, it shall be a bit-pattern encoding the combination of ISCC-UNITs and the bit-length of the ISCC-BODY (see 4.2.5, Table 7).

EXAMPLE   Header Field Examples

0 = 0000

1 = 0001

…

7 = 0111

8 = 1000 0000

9 = 1000 0001

…

4.2.2   MainTypes

The MainType header-field shall signify the type of an ISCC (see Table 2).

Backward incompatible updates to an algorithm associated with a MainType shall be indicated by incrementing the version field of the ISCC-HEADER of the respective MainType.

NOTE   This document specifies initial algorithms (version 0) for all reserved MainTypes with the exception of the SEMANTIC type which is not currently defined.