INTERNATIONAL STANDARD

ISO/IEC 23092-1

Second edition
2020-10

Information technology — Genomic information representation —

Part 1: Transport and storage of genomic information

Technologie de l'information — Représentation des informations génomiques —
Partie 1: Transport et stockage des informations génomiques

Reference number
ISO/IEC 23092-1:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

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 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

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 Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

This second edition cancels and replaces the first edition (ISO/IEC 23092-1:2019), which has been technically revised.

The main changes compared to the previous edition are as follows:

— reference box syntax and semantics have been updated;

— syntax, semantics and decoding process for cluster signatures has been fixed;

— the scope of some parameters has been changed from dataset_header to dataset_parameter_set;

— new dataset_group_ID and dataset_ID fields have been added to the metadata and protection boxes;

— minor fixes in transport format;

— editorial changes.

A list of all parts in the ISO/IEC 23092 series can be found on the ISO website.

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

The advent of high-throughput sequencing (HTS) technologies has the potential to boost the adoption of genomic information in everyday practice, ranging from biological research to personalized genomic medicine in clinics. As a consequence, the volume of generated data has increased dramatically during the last few years, and an even more pronounced growth is expected in the near future.

At the moment, genomic information is mostly exchanged through a variety of data formats, such as FASTA/FASTQ for unaligned sequencing reads and SAM/BAM/CRAM for aligned reads. With respect to such formats, the ISO/IEC 23092 series provides a new solution for the representation and compression of genome sequencing information by:

— Specifying an abstract representation of the sequencing data rather than a specific format with its direct implementation.

— Being designed at a time point when technologies and use cases are more mature. This permits addressing one limitation of the textual SAM format, for which the incremental ad-hoc addition of features followed along the years, resulting in an overall redundant and suboptimal format which was unnecessarily complicated.

— Separating free-field user-defined information with no clear semantics from the genomic data representation. This allows a fully interoperable and automatic exchange of information between different data producers.

— Allowing multiplexing of relevant metadata information with the data since data and metadata are partitioned at different conceptual levels.

— Following a strict and supervised development process which has proven successful in the last 30 years in the domain of digital media for the transport format, the file format, the compressed representation and the application program interfaces.

The ISO/IEC 23092 series provides the enabling technology that will allow the community to create an ecosystem of novel, interoperable, solutions in the field of genomic information processing. In particular it offers:

— Consistent, general and properly designed format definitions and data structures to store sequencing and alignment information. A robust framework which can be used as a foundation to implement different compression algorithms.

— Speed and flexibility in the selective access to coded data, by means of newly-designed data clustering and optimized storage methodologies.

— Low latency in data transmission and consequent fast availability at remote locations, based on transmission protocols inspired by real-time application domains.

— Built-in privacy and protection of sensitive information, thanks to a flexible framework which allows customizable, secured access at all layers of the data hierarchy.

— Reliability of the technology and interoperability among tools and systems, owing to the provision of a procedure to assess conformance to this document on an exhaustive dataset.

— Support to the implementation of a complete ecosystem of compliant devices and applications, through the availability of a normative reference implementation covering the totality of the ISO/IEC 23092 series.

The fundamental structure of the ISO/IEC 23092 series data representation is the *genomic record*. The genomic record is a data structure consisting of either a single sequence read, or a paired sequence read, and its associated sequencing and alignment information; it may contain detailed mapping and alignment data, a single or paired read identifier (read name) and quality values.

Without breaking traditional approaches, the genomic record introduced in the ISO/IEC 23092 series provides a more compact, simpler and manageable data structure grouping all the information related to a single DNA template, from simple sequencing data to sophisticated alignment information.

The genomic record, although it is an appropriate logic data structure for interaction and manipulation of coded information, is not a suitable atomic data structure for compression. To achieve high compression ratios, it is necessary to group genomic records into clusters and to transform the information of the same type into sets of descriptors structured into homogeneous blocks. Furthermore, when dealing with selective data access, the genomic record is a too small unit to allow effective and fast information retrieval.

For these reasons, this document introduces the concept of access unit, which is the fundamental structure for coding and access to information in the compressed domain.

The access unit is the smallest data structure that can be decoded by a decoder compliant with ISO/IEC 23092-2. An access unit is composed of one block for each descriptor used to represent the information of its genomic records; therefore, a block payload is the coded representation of all the data of the same type (i.e. a descriptor) in a cluster.

In addition to clusters of genomic records compressed into access units, reads are further classified in six data classes: five classes are defined according to the result of their alignment against one or more reference sequences; the sixth class contains either reads that could not be mapped or raw sequencing data. The classification of sequence reads into classes enables the development of powerful selective data access. In fact, access units inherit a specific data characterization (e.g. perfect matches in Class P, substitutions in Class M, indels in Class I, half-mapped reads in Class HM) from the genomic records composing them, and thus constitute a data structure capable of providing powerful filtering capability for the efficient support of many different use cases.

Access units are the fundamental, finest grain data structure in terms of content protection and in terms of metadata association. In other words, each access unit can be protected individually and independently. Figure 1 shows how access units, blocks and genomic records relate to each other in the ISO/IEC 23092 series data structure.

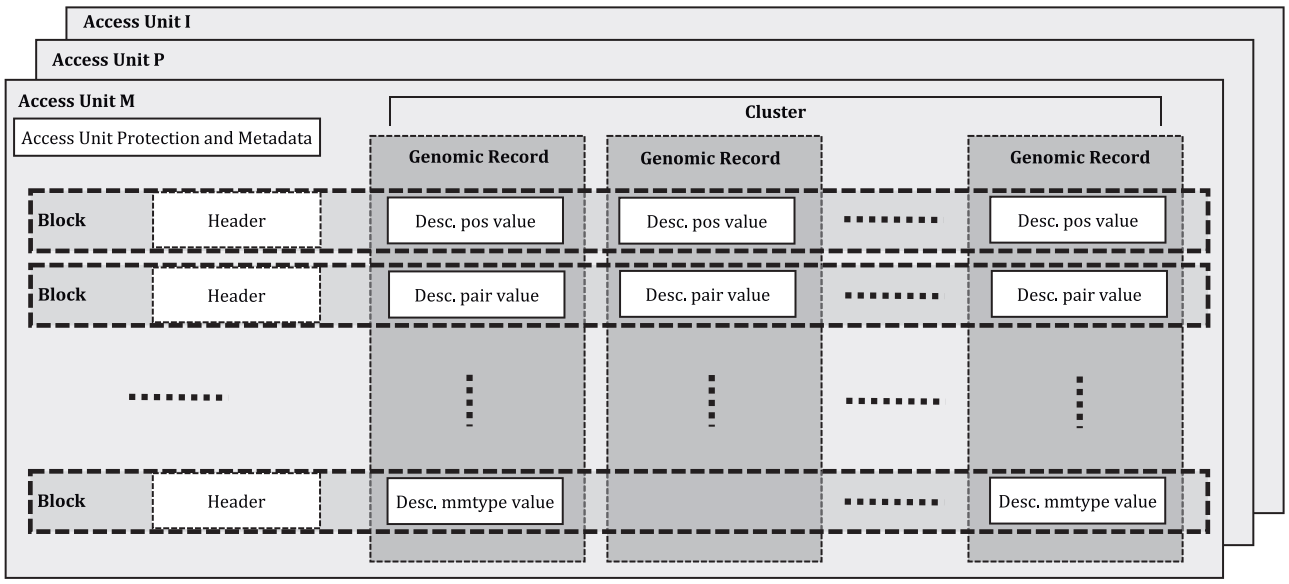

Figure 1 — Access units, blocks and genomic records

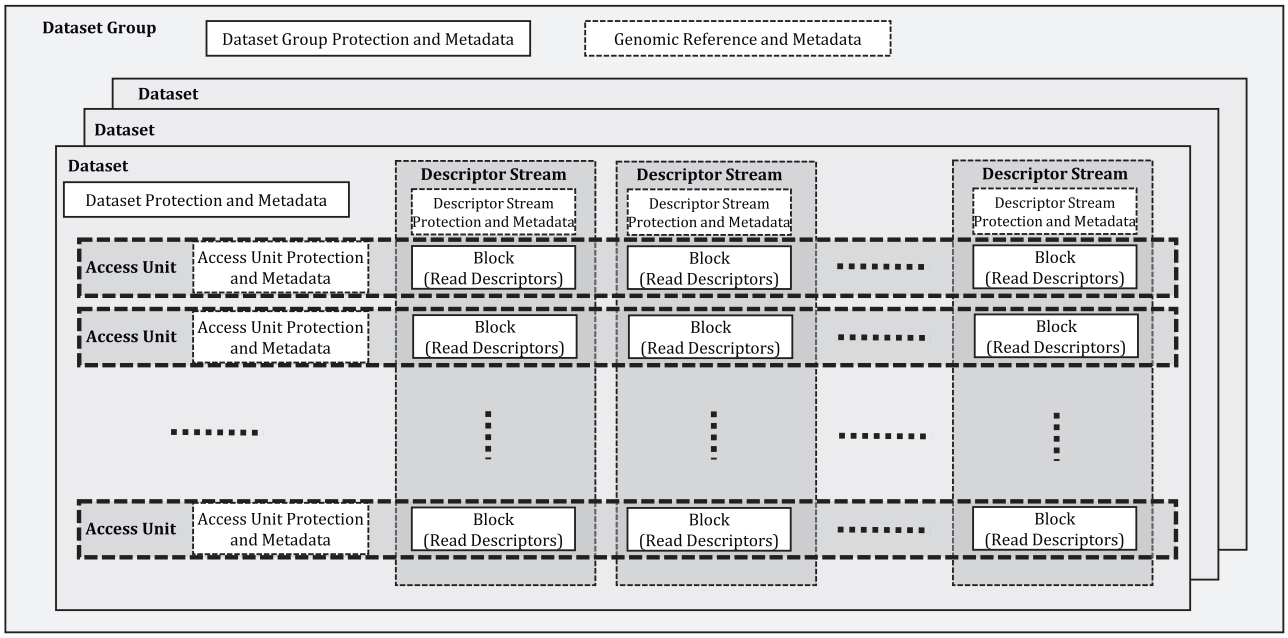

Figure 2 — High-level data structure: datasets and dataset group

A dataset is a coded data structure containing headers and one or more access units. Typical datasets could, for example, contain the complete sequencing of an individual, or a portion of it. Other datasets could contain, for example, a reference genome or a subset of its chromosomes. Datasets are grouped in dataset groups, as shown in Figure 2.

A simplified diagram of the dataset decoding process is shown in Figure 3.

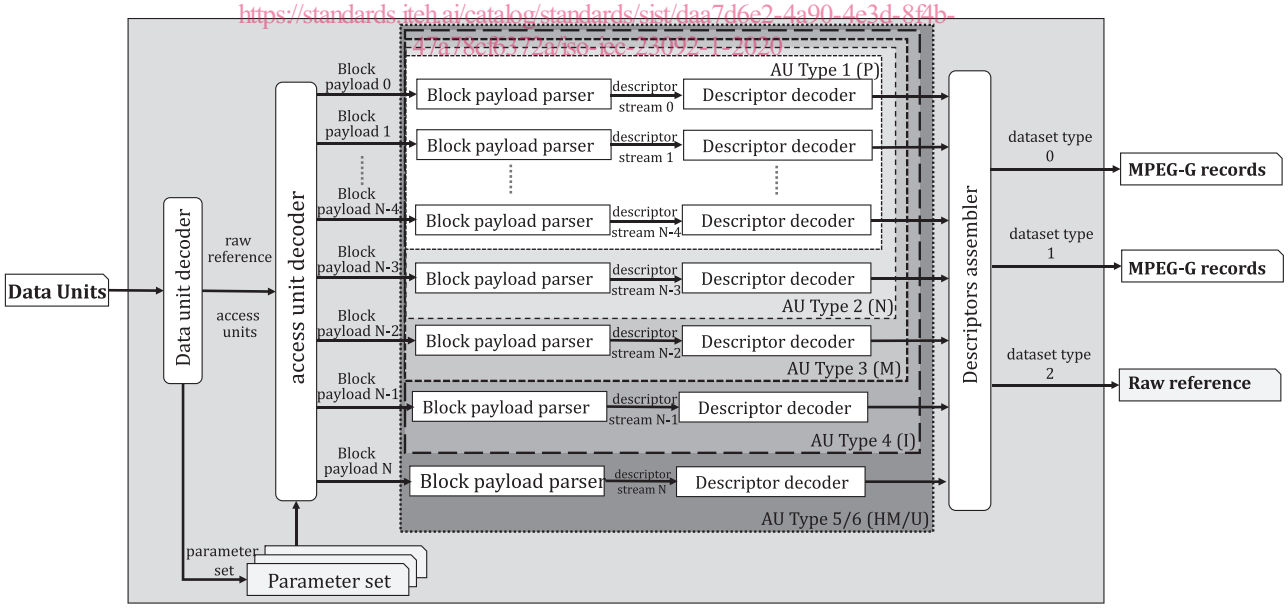

Figure 3 — Decoding process

This document defines the syntax and semantics of the data formats for both transport and storage of genomic information. According to this document, the compressed sequencing data can be multiplexed into a bitstream suitable for packetization for real-time transport over typical network protocols. In storage use cases, coded data can be encapsulated into a file format with the possibility to organize blocks per descriptor stream or per access units, to further optimize the selective access performance

to the type of data access required by the different application scenarios. This document further provides a reference process to convert a transport stream into a file format and vice versa.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO and IEC take no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured ISO and IEC that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO and IEC. Information may be obtained from the patent database available at www.iso.org/patents.

Information technology — Genomic information representation —

Part 1:
Transport and storage of genomic information

1 Scope

This document specifies data formats for both transport and storage of genomic information, including the conversion process.

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, *Information technology — Universal Coded Character Set (UCS)*

ISO/IEC 23092-2, *Information technology — Genomic information representation — Part 2: Coding of genomic information*

ISO/IEC 23092-3, *Information technology — Genomic information representation — Part 3: Metadata and application programming interfaces (APIs)*

IETF RFC 3986, *Uniform Resource Identifier (URI): Generic Syntax*

IETF RFC 7320, *URI Design and Ownership*

3 Terms and definitions

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

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

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

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

3.1
access unit
logical data structure containing a coded representation of genomic information to facilitate bit stream access and manipulation

3.2
access unit covered region
genomic range comprised between the access unit start position and the access unit end position, inclusive

3.3
access unit start position
position of the left-most mapped base among the first alignments of all genomic records contained in the access unit, irrespective of the strand

3.4
access unit end position
position of the right-most mapped base among the first alignments of all genomic records contained in the access unit, irrespective of the strand

3.5
access unit range
genomic range comprised between the access unit start position and the right-most genomic record position among all genomic records contained in the access unit

3.6
access unit covered region
genomic range comprised between the access unit start position and the access unit end position inclusive

3.7
alignment
information describing the similarity between a sequence (typically a sequencing read) and a reference sequence (for instance, a reference genome)

3.8
box
object-oriented building unit defined by a unique type identifier and length

3.9
cluster
aggregation of genomic records

3.10
cluster signature
signature
sequence of nucleotides that is common to most or all genomic records belonging to a cluster

3.11
container box
box (3.8) whose sole purpose is to contain and group a set of related boxes

3.12
data stream
set of *packets* (3.20) transporting the same data type

3.13
extended access unit start position
position of the left-most mapped base among all alignments of all genomic records contained in the access unit, irrespective of the strand

3.14
extended access unit end position
position of the right-most mapped base among all alignments of all genomic records contained in the access unit, irrespective of the strand

3.15
file format
set of data structures for the storage of coded information

3.16
genomic position
position
integer number representing the zero-based position of a nucleotide within a reference sequence

3.17
genomic region
region
genomic interval between a start nucleotide position and an end nucleotide position, inclusive

3.18
genomic range
range
interval of positions on a reference sequence defined by a start position *s* and an end position *e* such that $s \leq e$; the start and the end positions of a genomic range are always included in the range

3.19
mapped base
base of the aligned read that either matches the corresponding base on the reference sequence or can be turned into the corresponding base on the reference sequence via a substitution

3.20
packet
transmission unit transporting segments of any of the data structures defined in this document

3.21
reference genome
representative example of the sequences for a species' genetic material

Note 1 to entry: Genetic material meaning the sequences of the DNA molecules present in a typical cell of that species.

3.22
reference sequence
nucleic acid sequence with biological relevance

Note 1 to entry: Each reference sequence is indexed by a one-dimensional integer coordinate system whereby each integer within range identifies a single nucleotide. Coordinate values can only be equal to or larger than zero. The coordinate system in the context of this standard is zero-based (i.e. the first nucleotide has coordinate 0 and it is said to be at position 0) and linearly increasing within the string from left to right.

3.23
genomic segment
segment
contiguous sequence of nucleotides, typically output of the sequencing process and sequenced from one strand of a template

3.24
sequence read
read
readout, by a specific technology more or less prone to errors, of a continuous part of a nucleic acid molecule extracted from an organic sample

3.25
syntax field
element of data represented in the data format

3.26
template
genomic sequence that is produced by a sequencing machine as a single unit

Note 1 to entry: A template can be made of one or more segments, being called single-end sequencing read when it only has one segment and paired-end sequencing read when it has two segments.

3.27
transport format
set of data structures for the transport of coded information

3.28
variable
parameter either inferred from syntax fields or locally defined in a process description

4 Mathematical operators

NOTE The mathematical operators used in this document are similar to those used in the C programming language. However, integer division with truncation and rounding are specifically defined. The bitwise operators are defined assuming two's-complement representation of integers. Numbering and counting loops generally begin from 0.

4.1 Arithmetic operators

+ addition

– subtraction (as a binary operator) or negation (as a unary operator)

++ increment

* multiplication

/ integer division with truncation of the result toward 0 (for example, 7/4 and –7/–4 are truncated to 1 and –7/4 and 7/–4 are truncated to –1)

4.2 Logical operators

|| logical OR

&& logical AND

! logical NOT

4.3 Relational operators

> greater than

≥ greater than or equal to

< less than

≤ less than or equal to

== equal to

!= not equal to

4.4 Bitwise operators

& AND

| OR

>> shift right with sign extension

<< shift left with 0 fill

4.5 Assignment

= assignment operator

4.6 Unary operators

sizeof(N) size in bytes of N, where N is either a data structure or a data type

5 Structure of coded genomic data

5.1 Genomic records

The genomic record, in this document, is a data structure consisting of either a single sequence read, or paired sequence reads, and its associated sequencing and alignment information. The genomic record may contain detailed mapping and alignment data, a single or paired read identifier (read name) and quality values.

When alignment information is present, the genomic record position is defined as the position of the left-most mapped base of the genomic record on the reference genome. Genomic record positions are 0-based in the ISO/IEC 23092 series. In case of multiple alignments, the position of the first alignment in the record is considered; in such a case, the first alignment shall be the one with the leftmost position among all the alignments with the best score.

In case of unmapped reads (i.e. no alignment information present) the notion of position does not apply to the genomic record.

In case of aligned content, bases that are present in the reads of the genomic record and not present in the reference sequence (*insertions*) and bases preserved by the alignment process but not mapped on the reference sequence (*soft clips*) do not have mapping positions.

Table 1 enumerates all the types of data that a genomic record can contain. ISO/IEC 23092-2 defines technology that allows coding all and only those types of data into a set of descriptors; data, and consequently descriptors, which are mandatory or optional, are also specified in ISO/IEC 23092-2, as well as how they are used to represent multiple alignments.

Table 1 — Genomic records

Data	Semantics
Record identifier	name of the record (e.g. read names)
Sequence reads	sequencing readout, as one or more strings of bases
Quality values	quality scores of the sequence reads
Strandedness	information about the strandedness of each read of the Record
Length	length of the sequence reads
Position	position on the reference genome of the left-most mapped genomic record base
Pairing	position or distance of the mate reads (e.g. in a pair)
Flags	technical, additional alignment information (duplicates, proper pairs, failures)
Mismatches	information about position and type of each mismatch in mapped records
Clips	information about clipped bases (soft and hard clips) in mapped records
Mapping scores	mapping scores for an alignment
Multiple alignments	information about the number of alignments and the alternative alignment information about each segment of the record
Group	read group the genomic record belongs to

ISO/IEC 23092-2 defines an output record format for all types of data in Table 1. These records shall be generated by decoders compliant to ISO/IEC 23092-2 as output of the decoding process.

5.2 Data classes

Six data classes are specified to classify genomic records according to the result of the mapping of the encoded sequence reads against one or more reference sequences.

In the case of more than one read in a template, if both reads are mapped, the genomic record belongs to the class of the read with the highest class_ID. In case of multiple alignments, the genomic record belongs to the class of the first alignment in the record.

The data classes and their descriptions are specified in Table 2.

Table 2 — Data classes

Class ID	Class name	Record content
1	CLASS_P	Only reads perfectly matching to the reference sequence.
2	CLASS_N	Reads perfectly matching to the reference sequence or containing mismatches which are unknown bases only.
3	CLASS_M	Reads perfectly matching to the reference sequence or containing substitutions or unknown bases, but no insertions, no deletions, no splices and no clipped bases.
4	CLASS_I	Reads perfectly matching to the reference sequence or containing substitutions, unknown bases, insertions, deletions, splices or clipped bases.
5	CLASS_HM	Paired-end reads with only one mapped read.
6	CLASS_U	Unmapped reads only.

Genomic records of each data class are coded by means of several descriptors; conversely, a descriptor is a coding element needed to represent part of the information. Descriptors for each data class are specified in ISO/IEC 23092-2.

Descriptors are coded in blocks. Blocks are defined in subclause 6.5.5. A sequence of block payloads of a single descriptor composes a descriptor stream. All block payloads in a descriptor stream contain compressed descriptors of a single type representing reads of the same data class.

5.3 Access units

Access units (AUs) are data structures containing a coded representation of genomic information and optionally related metadata to facilitate the bitstream access and manipulation. An access unit contains either genomic records belonging to the same data class or a fragment of a reference sequence.

The access unit is the smallest data organization that can be decoded by a decoder compliant with ISO/IEC 23092-2.

Access units are orthogonal to descriptor streams: an access unit is composed of all and only those blocks of the descriptor streams that are necessary to decode the information contained in a cluster of records of a given data class.

An access unit can be of several types according to the class of the coded data.

Table 3 — Access unit type

Access unit type		Class of data
Name	Value	
P_TYPE_AU	1	CLASS_P
N_TYPE_AU	2	CLASS_N
M_TYPE_AU	3	CLASS_M
I_TYPE_AU	4	CLASS_I
HM_TYPE_AU	5	CLASS_HM
U_TYPE_AU	6	CLASS_U

Depending on the type of coded information, an access unit can be decoded either independently of any other access unit or using information contained in other access units.

5.4 Datasets

A dataset is a data structure containing headers and access units. The set of access units composing the dataset constitutes the dataset payload.

One or more datasets are assembled into a dataset group.

5.5 Selective access

In the case of selective access to a genomic region comprised between a *start* genomic position and an *end* genomic position the decoder shall return: a) all the access units whose covered region overlaps the region defined by *start* and *end* with at least one base, and the parameter sets that are needed to decode them; b) at least the reference portion that is necessary to decode the access units identified in a).

In the case of selective access to signed content identified by a U_cluster_signature signature the decoder shall return all the access units whose signature corresponds to U_cluster_signature, and the parameter sets that are needed to decode them. Examples of selective access strategies are described in Annex B.

6 Data format

6.1 Format structure

6.1.1 General

Table 4 presents the overall data structures and hierarchical encapsulation levels.

Boxes that may occur at the top-level are shown in the left-most column; indentation is used to show possible containment. Not all boxes need be used in all files; the mandatory boxes are marked with an asterisk (*) in the *Mandatory* column: such column refers to the relevant scope (File and/or Transport). Optional boxes are represented with dashed borders in Figure 4 and Figure 5. Mandatory boxes are represented with solid borders. When no entry is present in the *Scope* column, scope is both *File* and *Transport*. See the specification of each individual box for the assumptions when the optional boxes are not present. If the box key is represented in *italic* format in Table 4, the corresponding box is represented either with no Key and no Length, but only Value in the gen_info format, as specified in subclause 6.3, for all boxes but offset, or as specified in subclause 6.6.5.1 for the offset box.