



**International
Standard**

ISO/IEC 23092-1

**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

**Third edition
2025-01**

This document is a preview generated by EMS



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2025

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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	v
Introduction	vii
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Conventions	4
4.1 Operators and functions.....	4
4.1.1 Arithmetic operators.....	4
4.1.2 Logical operators.....	4
4.1.3 Relational operators.....	4
4.1.4 Bitwise operators.....	4
4.1.5 Assignment operators.....	5
4.1.6 String/Character functions and operator.....	5
4.1.7 Data structure function and operator.....	5
4.1.8 Mathematical functions.....	5
4.1.9 Array operation functions.....	5
4.2 Syntax and semantics.....	6
4.2.1 Method of specifying syntax in tabular form.....	6
4.2.2 Bit ordering.....	6
4.2.3 Specification of syntax functions.....	6
4.2.4 Processes.....	7
5 Structure of coded genomic data	7
5.1 Genomic sequencing data record.....	7
5.2 Genomic annotation data records.....	8
5.3 Data classes.....	9
5.4 Access units.....	10
5.5 Datasets.....	10
5.6 Annotation data tile.....	11
5.7 Annotation tables.....	11
5.8 Annotation access units.....	11
5.9 Selective access.....	12
6 Data format	12
6.1 Format structure.....	12
6.1.1 General.....	12
6.1.2 Box order.....	17
6.2 Syntax for representation.....	18
6.3 Output data unit.....	19
6.4 Data structures common to file format and transport format.....	20
6.4.1 File header.....	20
6.4.2 Dataset group.....	20
6.4.3 Dataset.....	29
6.4.4 Access unit.....	40
6.4.5 Block.....	46
6.4.6 Annotation Table.....	47
6.4.7 Attribute Group.....	57
6.4.8 Annotation access unit.....	59
6.4.9 AAU block.....	63
6.5 Data structures specific to file format.....	64
6.5.1 General.....	64
6.5.2 Indexing.....	64
6.5.3 Descriptor stream.....	74
6.5.4 Offset.....	76
6.6 Data structures specific to transport format.....	77

6.6.1	General	77
6.6.2	Data streams	77
6.6.3	Dataset mapping table list	77
6.6.4	Dataset mapping table	78
6.6.5	Packet	80
6.7	Reference procedures to convert transport format to file format	81
6.7.1	Procedure for genomic sequencing data	81
6.7.2	Procedure for genomic annotation data	83
7	String indexing technologies	87
7.1	Master string index	87
7.1.1	General	87
7.1.2	Syntax	87
7.1.3	Master String Index Header	87
7.1.4	String index	88
7.1.5	Compressed string index	90
7.2	Decoding and querying processes	96
7.2.1	String index payload	96
7.2.2	Helper functions	97
7.2.3	Substring decoding process	98
7.2.4	Suffix array lookup process	99
7.2.5	Inverse suffix array process	99
7.2.6	Character decoding process	100
7.2.7	LF-mapping process	101
7.2.8	Extended LF-mapping process	101
7.2.9	Substring position search process	102
7.2.10	Searching for substring positions with the string index	103
7.2.11	Decoding a subset of the string index	104
7.2.12	Decoding all the strings of a specific annotation data tile	104
7.2.13	Retrieving whole strings with the string index	106
7.2.14	Retrieving data tile index(es) associated with a position and record indexes	107
8	Indexing for numeric range searches	110
8.1	B-Tree indexing	110
8.1.1	General	110
8.1.2	Syntax	110
8.1.3	Semantics	111
Annex A (informative) IETF RFC 3986 specification summary		112
Annex B (informative) Selective access strategies for genomic sequencing data		113
Annex C (informative) Selective access strategies for genomic annotation data		116
Annex D (informative) Depacketization process		132
Annex E (informative) Efficient handling of symmetric annotation data		135
Bibliography		137

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 or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take 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 and IEC had 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 and <https://patents.iec.ch>. ISO and IEC 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. In the IEC, see www.iec.ch/understanding-standards.

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 third edition cancels and replaces the second edition (ISO/IEC 23092-1:2020), which has been technically revised.

The main changes are as follows:

- Updates to the overall hierarchy of data structures and box order in [subclause 6.1](#)
- Extensions for the transport and storage of genomic annotation data, in addition to genomic sequencing data, in support of ISO/IEC 23092-6:2023 specifications while maintaining backward compatibility, which include:
 - An overview of genomic annotation data records in [subclause 5.2](#), with detailed formats specified in Part 6
 - Basic annotation table information in dataset header (as specified in [subclause 6.4.3.2](#)) and annotation encoding parameters in dataset parameter set (as specified in [subclause 6.4.3.7](#))
 - Additional data structures such as annotation table (atcn, as specified in [subclause 6.4.6](#)), attribute group (agcn, as specified in [subclause 6.4.7](#)), annotation access unit (aauc, as specified in [subclause 6.4.8](#)), AAU block (as specified in [subclause 6.4.9](#)), attribute data byte offset (adbo, as specified in [subclause 6.5.2.3](#)) and annotation table index (atix, as specified in [subclause 6.5.2.4](#))
 - The reference procedure for conversion from transport format to file format for genomic annotation data in [subclause 6.7.2](#)

ISO/IEC 23092-1:2025(en)

- Data structure for B-Tree indexing (as specified in [subclause 8.1](#)) and selective access strategies for genomic annotation data (as specified in [Annex C](#))
- Extensions in support of ISO/IEC 23092-3:2022 which include:
 - New container boxes for metrics metadata: DT_metrics (dtmt, as specified in [subclause 6.4.3.4](#)) and AU_metrics (aamt, as specified in [subclause 6.4.4.5](#)), containing statistical information (with detailed formats specified in Part 3), which allows for fast and direct extraction of statistics associated with the dataset and access unit content
 - New container boxes for clinical data linkage (CDL) metadata: DG_CDL (dgcd, as specified in [subclause 6.4.2.7](#)), DT_CDL (dtcd, as specified in [subclause 6.4.3.5](#)) and AT_CDL (atcd, as specified in [subclause 6.4.6.4](#)), for establishing linkages to external data sources, which enables access to the clinical data of individual samples
- The inclusion of FM-Index-based entropy coding algorithm (as specified in [Clause 7](#)), which provides string search capabilities in the compressed domain

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

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 and www.iec.ch/national-committees.

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.

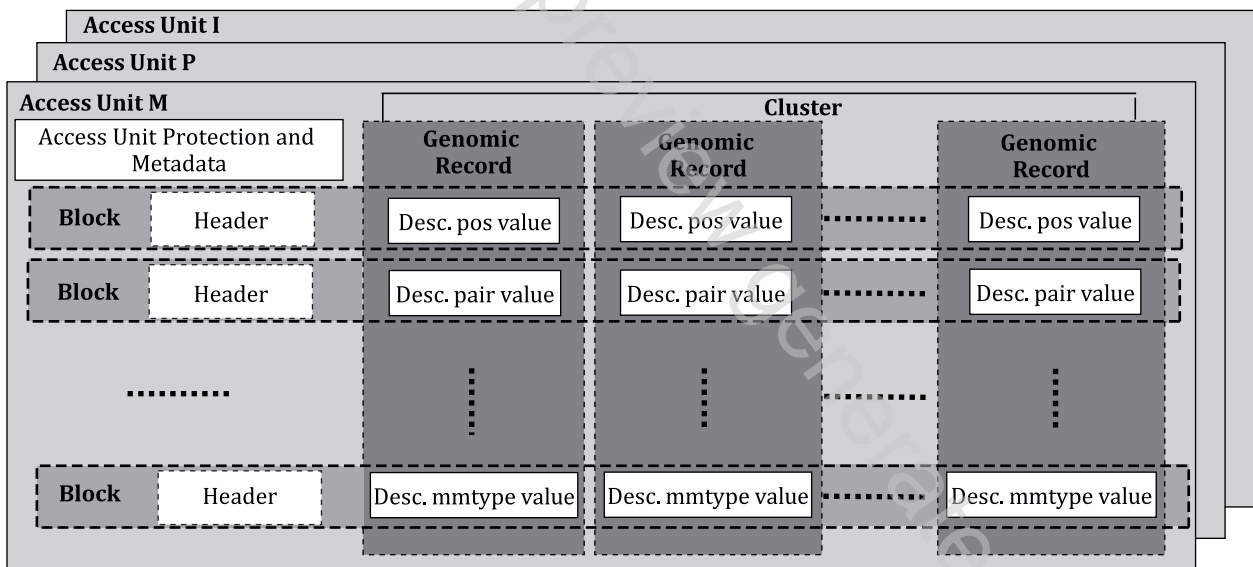


Figure 1 — Access units, blocks and genomic records

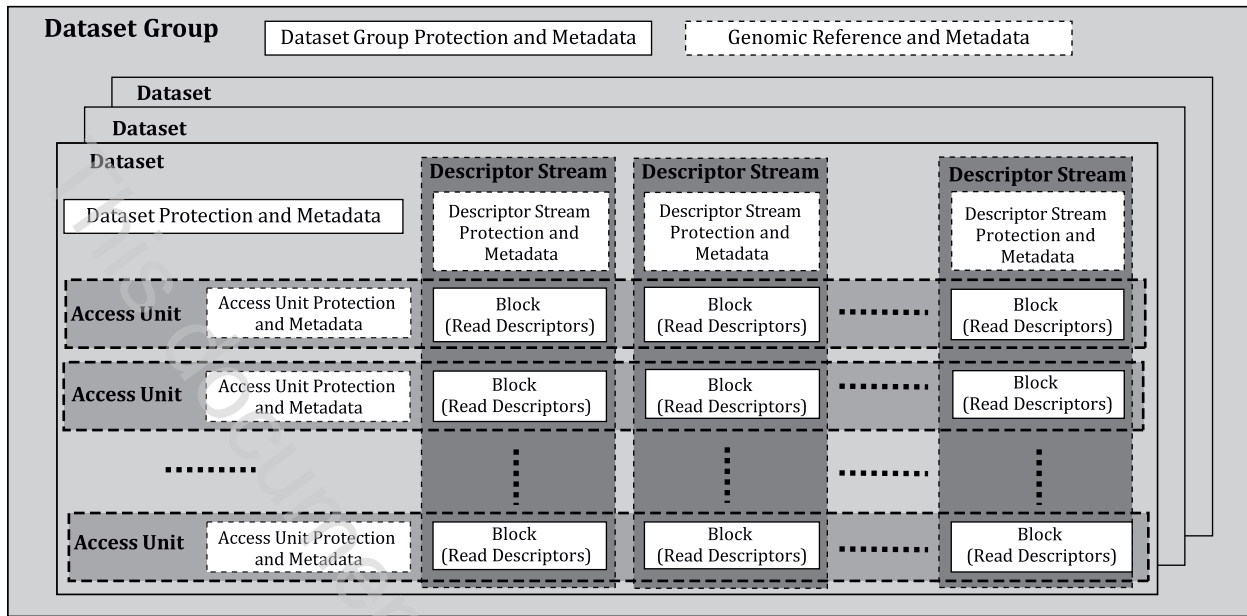


Figure 2 — High-level data structure: datasets and dataset group

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](#).

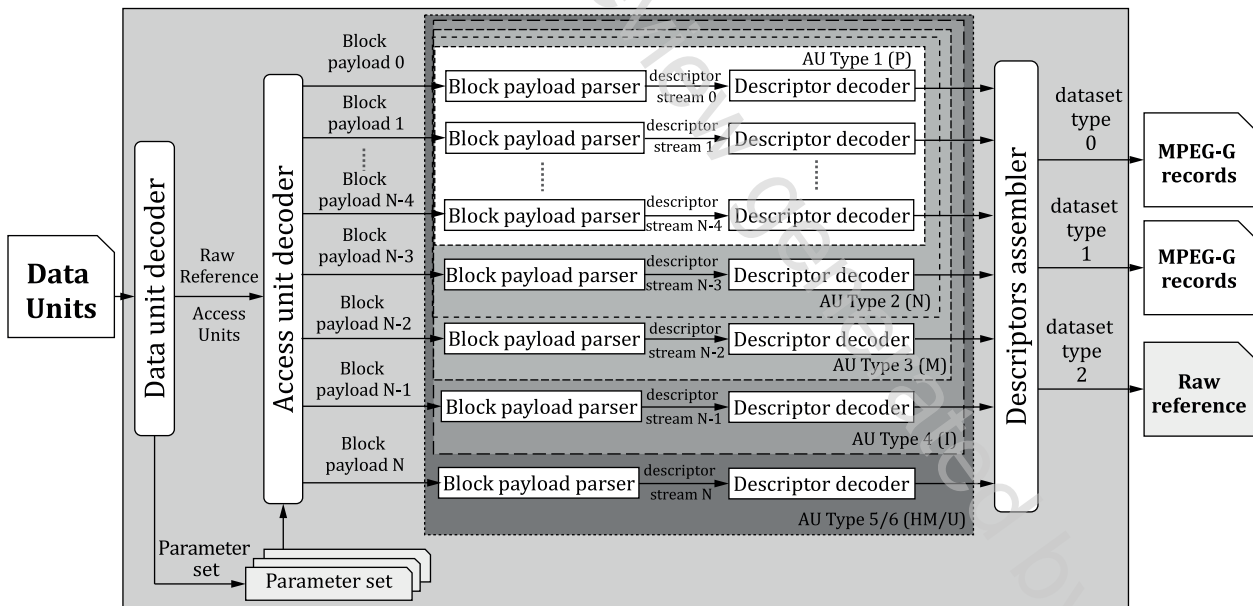


Figure 3 — Decoding process

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.

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)*

ISO/IEC 23092-6, *Information technology — Genomic information representation — Part 6: Coding of genomic annotations*

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 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/>

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