DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 23092-5

ISO/IEC JTC **1**/SC **29** Secretariat: **JISC**

Voting begins on:
2020-01-10

Voting terminates on:
2020-04-03

Information technology — Genomic information representation —

Part 5:
Conformance

ICS: 35.040.99

This document is circulated as received from the committee secretariat.

Reference number
ISO/IEC DIS 23092-5:2020(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
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*.

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 the addressing of one limitation of the textual SAM format, for which incremental ad-hoc addition of features followed along the years, resulting in an overall redundant and suboptimal format which at the same time results not general and unnecessarily complicated.

— Normatively separating free-field user-defined information with no clear semantics from the normative 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.

This document 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 normative 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 series. 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.

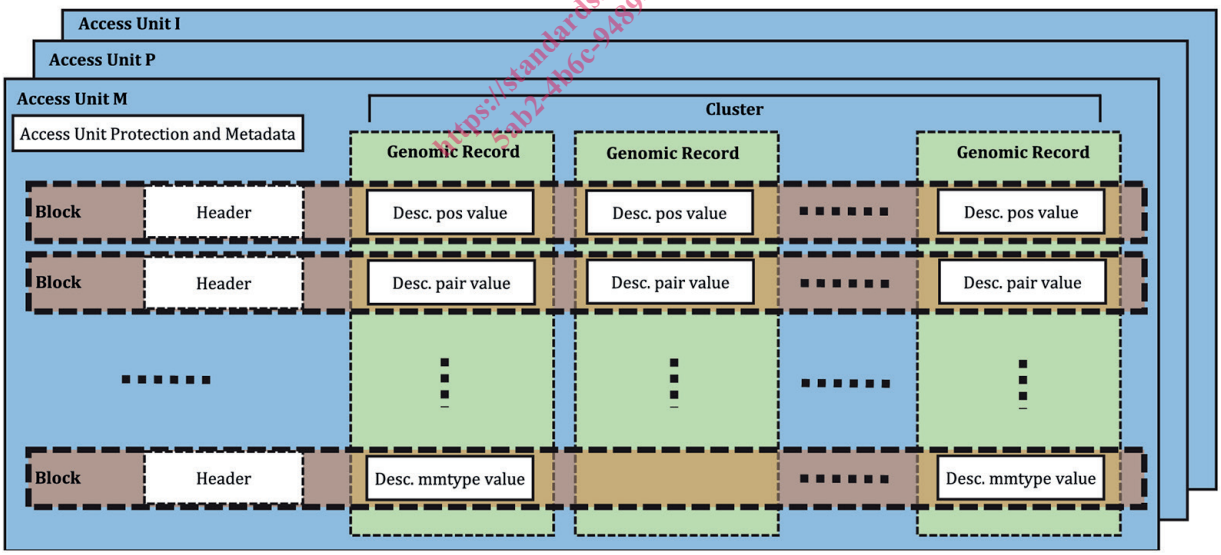

Figure 1 — Access units, blocks and genomic records

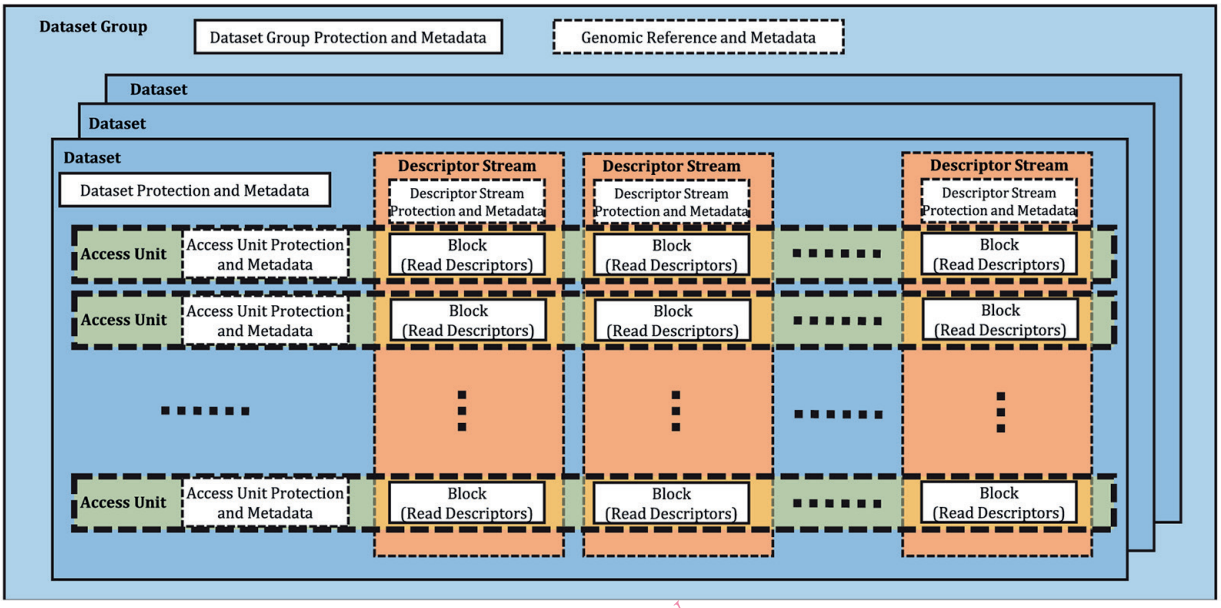

Figure 2 — High-level data structure: datasets and dataset group

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.

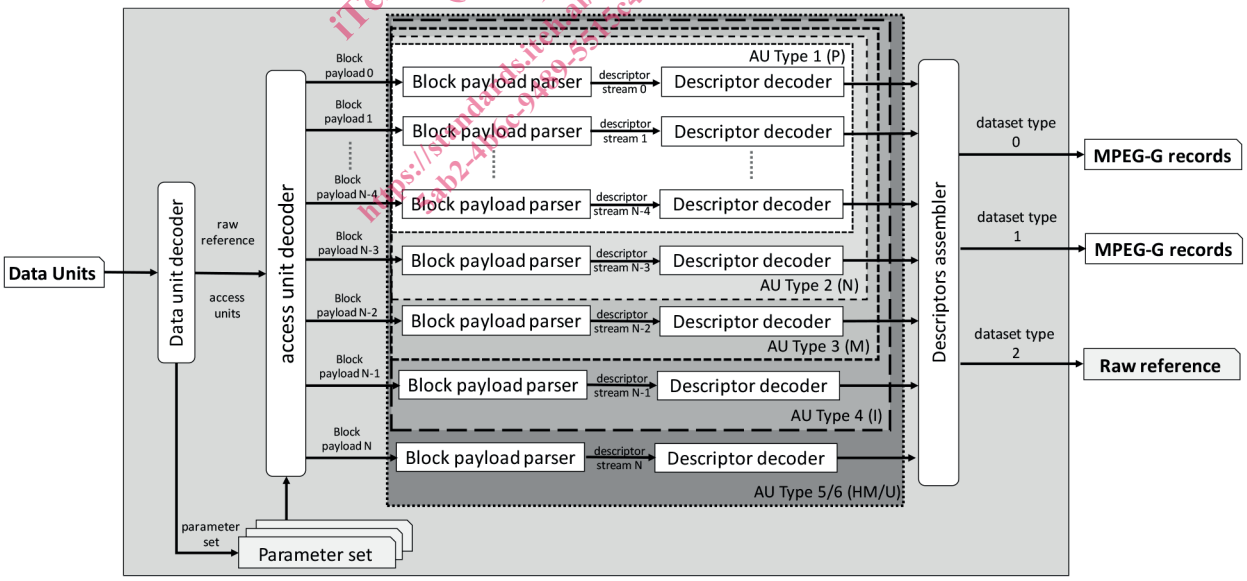

Figure 3 — Decoding process

This document defines a set of test procedures designed to verify whether bitstreams and decoders meet requirements specified in Parts 1 and 2 of ISO/IEC 23092. In this Part of ISO/IEC 23092 encoders are not addressed.

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:

GenomSys SA
EPFL Innovation Park Building C
CH-1015 Lausanne
Switzerland
info@genomsys.com

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified above. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — Genomic information representation —

Part 5: Conformance

1 Scope

This Part of the Standard specifies a set of test procedures designed to verify whether bitstreams and decoders meet requirements specified in Parts 1 and 2 of the ISO/IEC 23092 series.

Procedures are described for testing conformity of bitstreams and decoders to the requirements that are fully determined in Parts 1 and 2 of ISO/IEC 23092. This Part identifies those requirements, associates them to functionality under test and defines how conformity with them can be tested. Test bitstreams implemented according to those functionalities are provided in electronic form as specified in clause 6.

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 23092-1:2019, *Information technology — Genomic information representation — Part 1: Transport and storage of genomic information*

ISO/IEC 23092-2:2019, *Information technology — Genomic information representation — Part 2: Coding of genomic information*

3 Terms and definitions

For the purposes of this document, the terms and definitions in ISO/IEC 23092-1 and the following 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/

4 ISO/IEC 23092 Part 1 Conformance

4.1 *Definition of Part 1 Conformance*

4.1.1 Assumptions

In the sections of this Part of ISO/IEC 23092 describing conformity tests for Part 1 of ISO/IEC 23092, the following assumptions are made:

The term ‘file’ means ISO/IEC 23092-1 file; the term ‘transport’ means ISO/IEC 23092-1 transport.

The term ‘decapsulator’ means ISO/IEC 23092-1 decapsulator, i.e. an implementation of the parsing and demultiplexing processes specified by ISO/IEC 23092-1. A decapsulator operates on data structures that

are specified in clause 6 of ISO/IEC 23092-1. An ISO/IEC 23092-1 decapsulator is also interchangeably called a “Part 1 decoder” in this specification.

If any statement made in this section accidentally contradicts a statement or requirement in ISO/IEC 23092-1, the text of ISO/IEC 23092-1 prevails.

The following subclauses specify the normative tests to verify the conformity of files and decapsulators. Those normative tests make use of normative test data (test files and reference outputs), made available as specified in clause 6, and it makes use of the reference software specified in ISO/IEC 23092-4, with source code available as described in Part 4 of this standard.

This Part of ISO/IEC 23092 does not specify normative tests to verify the conformity of transport.

4.1.2 Definition of ISO/IEC 23092 File conformity

An ISO/IEC 23092-1 file is a file that conforms to the specification defined by the normative sections of ISO/IEC 23092-1.

A conformant file shall meet all the requirements and implement all the restrictions in the syntax specified in ISO/IEC 23092-1.

Subclause 4.3 of this document defines the normative test that a file shall pass successfully in order to be claimed in conformity with this specification.

4.1.3 Definition of Part 1 decoder conformity

An ISO/IEC 23092-1 decoder, or decapsulator, is an implementation of the processes necessary to parse and demultiplex the normative data structures of ISO/IEC 23092-1 and to perform operations associated to these data structures.

A conformant Part 1 decoder shall meet all the requirements and implement all the restrictions in the syntax defined by the ISO/IEC 23092-1 specification.

Subclause 4.4 of this Part defines the normative tests that a decoder shall pass successfully in order to be claimed in conformity with this specification.

A conformant Part 1 decoder shall implement parsing and decapsulation procedures that are equivalent to the ones specified in ISO/IEC 23092-1 and meet all the general requirements defined in ISO/IEC 23092-1.

Fundamental requirement areas for Part 1 decoders and their mapping to functionality under test are listed in the following subclause.

4.2 *Requirements and functionality under test*

Table 1 — Requirement Areas for Part 1

Requirement Area	Functionality
Dataset Group	Dataset Extraction from dataset group
Reference	Get reference with checksum calculation
Indexing by positions	Selective access by position ranges
Indexing by signatures	Selective access by signatures for non-aligned content (signature decoding)
Labels	Selective access by labels (single dataset)
Non-indexed content	Content extraction without Indexing Table
DSC and AUC storage mode	Access in AUC and DSC mode
Ordered Blocks	Content extraction with and without ordered blocks