DRAFT INTERNATIONAL STANDARD
ISO/IEC DIS 23092-4

ISO/IEC JTC **1**/SC **29**

Secretariat: **JISC**

Voting begins on:
2020-01-22

Voting terminates on:
2020-04-15

Information technology — Genomic information representation —

Part 4:
Reference software

ICS: 35.040.99

Reference number
ISO/IEC DIS 23092-4:2020(E)

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 the clinic. As a consequence, an extraordinarily growing volume of generated data has been recorded 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, ISO/IEC 23092 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 to address 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 meta-data 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 the standard 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 specification.

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 Part 2 of this Specification. 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 to develop 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.

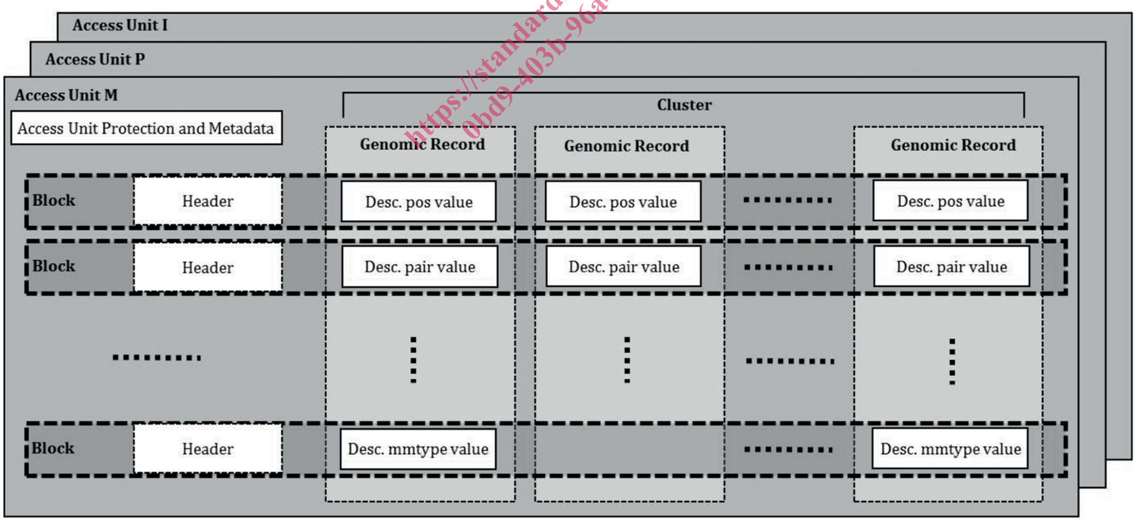

Figure 1 — Access units, blocks and genomic records

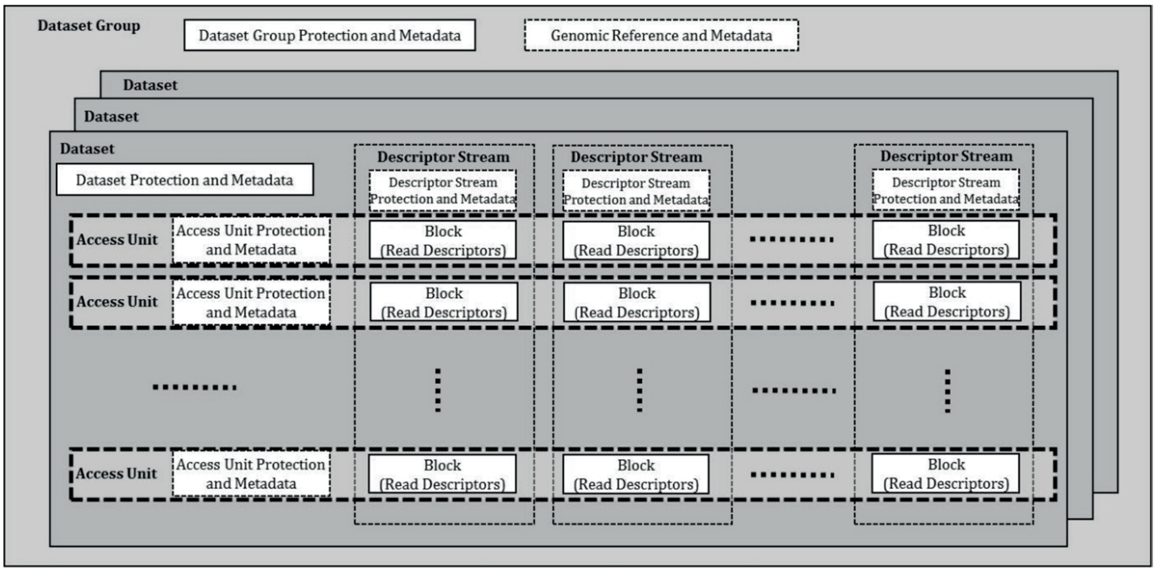

Figure 2 — High-level data structure: datasets and dataset group

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.

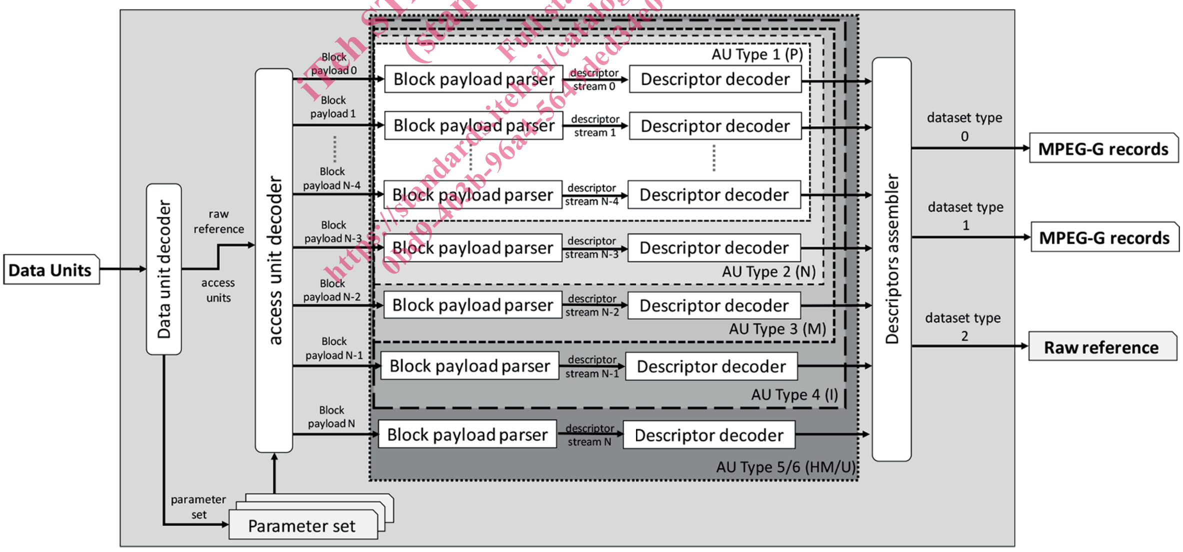

Figure 3 — Decoding process

This document contains the Reference Software for tools defined in parts 1 and 2 of the ISO/IEC 23092 series. This software has been derived from verification models used in the process of developing the standard.

The reference software is available at https://standards.iso.org/iso-iec/23092/-4/ed-1/en.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draws 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 4: Reference software

1 Scope

Reference software is normative in the sense that any conforming implementation of the software, taking the same conformant bitstreams, using the same output file format, will output the same file. Complying ISO/IEC 23092 implementations are not expected to follow the algorithms or the programming techniques used by the reference software. Although the decoding software is considered normative, it cannot add anything to the textual technical description included in parts 1 and 2 of ISO/IEC 23092.

The Genomic Information Representation Reference Software contained in this part of ISO/IEC 23092 series is referred to as Genomic Model (GM) and it comprises the following items:

1. **Decoding Software** is catalogued in subclause 6.3. This software accepts streams encoded according to the normative specification in parts 1 and 2 of ISO/IEC 23092 and decodes the streams into the content types associated with each stream. While this software appears in the normative part of this specification, attention is drawn to the fact that the implementation techniques used in this software are not considered normative – several different implementations could produce the same result – but the software is considered normative in that it correctly implements the decoding processes described in parts 1 and 2 of ISO/IEC 23092.

The Genomic Model is attached to this document in file `w18803_MPEG-G_Reference_SW_1.3.1-DIS.zip`.

Attention is called to the fact that this decoder is provided for the purpose of decoding normative bitstreams. The performance of this decoder should not be taken as indicative of that which can be obtained from implementations where computational optimization is given priority.

Since bitstream encoding software is provided, attention is called to the fact that this encoder is provided for the purpose of creating bitstreams with normative syntax. The performance of this encoder should not be taken as indicative of that which can be obtained from implementations where quality and computational optimization are given priority. The techniques used for encoding are not specified by this specification.

File locations given in this document are expressed relative to this document's location in the source tree.

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

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: Genomic Information Metadata and Application Programming Interfaces (APIs)*