 ISO/IEC JTC 1/SC 29/WG 8 N0040

**ISO/IEC JTC 1/SC 29/WG 8**

**MPEG GENOMIC CODING   
Convenorship: SNV (SWITZERLAND)**

**Document type:** Output Document

**Title:** Requirements for the Joint Call for Proposals on New Advanced Features and New Technologies for MPEG-G

**Status:** Approved

**Date of document:** 2021-05-1

**Source:** ISO/IEC JTC 1/SC 29/WG 8

# Expected action: None

**Action due date:** None

**No. of pages:** 5 (with cover page)

**Email of Convenor:** marco.mattavelli@epfl.ch

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg8>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 8**

**MPEG GENOMIC CODING**

**ISO/IEC JTC 1/SC 29/WG 8 N0040**

**Online – April 2021**

|  |  |
| --- | --- |
| **Title** | **Requirements for the Joint Call for Proposals on New Advanced Features and New Technologies for MPEG-G** |
| **Source** | **WG 8, MPEG GENOMIC CODING** |
| **Serial Number** | **20375** |

# Introduction Executive summary

This document lists the requirements that responses to the Final Joint Call for Proposals for extensions and improvements of ISO/IEC 23092 series shall meet.

# MPEG documents

All MPEG documents cited in this document are available on the MPEG documents repository (<http://wg11.sc29.org/>) as output documents issued during past MPEG meetings.

# Terminology

Definitions for most concepts appearing in the following text, and concerning the production and analysis of sequencing data, can be found in the Terms and Definitions sections of ISO/IEC 23092 (MPEG-G) Parts 1, 2 and 3.

## ISO/IEC 23092 (MPEG-G) hierarchy

Sequencing reads of different lengths (which depend on the technology used to generate them) can be localized at one or more points on the DNA molecule they originate from. They are the basic tokens of information at the foundation of all high-level biological experiments based on sequencing. It is hence only natural for the ISO/IEC 23092 (MPEG-G) hierarchy to be based on reads, which get organized in terms of, from bottom to top, Access Units, Datasets, and Dataset Groups. The proposed technology should be as compatible as possible with the current hierarchy and components of ISO/IEC 23092 series because the objective is to create a new profile providing higher performance and new functionality.

## Relevant data types

All data types supported by the ISO/IEC 23092 series are relevant for this call as well as all data present in the MPEG-G database (N0024).

# Requirements

This section introduces the definition of the requirements that an appropriate genomic information representation should meet when the relevant data types are considered. In addition, support for transport should also be provided – the relevant requirements apply to all kinds of data, and are considered in a separate section.

## Third Generation Sequencing (Long Read) Coding Mode and Improved Compression of Genome Sequences, Quality Scores and Metadata

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirements** | **Rationale** |
| 4.1.1 | The new proposed technology shall represent data in compressed form and offering better encoding or decoding performance than the current ISO/IEC 23092 series for at least one data item or type of sequencing data in terms of complexity or compression ratio. | To improve encoding and decoding speed for some type of data and use cases. |
| 4.1.2 | The solution shall be compatible and interoperable with the ISO/IEC 23092 series. | To enable integration with the existing parts of ISO/IEC 23092. |
| 4.1.3 | Lossless compression supporting the 16 symbol alphabet (A, C, G, T, R, Y, S, W, K, M, B, D, H, V, N, -) |  |
| 4.1.4 | Lossless compression of quality scores shall be supported. |  |
| 4.1.5 | Lossy compression of quality scores shall be supported. Lossy compression should respect metrics that MPEG needs to identify with their associated boundaries. | Providing defined error characteristics (such as rate distortion function) is preferred |
| 4.1.6 | Compressed data should be structured so that parallel processing is enabled and compression efficiency is preserved | This is intended to enable efficient parallel processing of data without affecting compression efficiency. |
| 4.1.7 | Compressed data should be structured so that efficient querying of data is enabled and compression efficiency is preserved | Clustering applications could benefit from efficient querying. |
| 4.1.8 | Compressed data streaming should be supported. | This implies that data consumption shall be possible before data transfer completion. |
| 4.1.9 | The association among headers, nucleotides and quality scores must be preserved |  |

## Machine Learning Coding Mode

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement** | **Rationale** |
| 4.2.1 | The new proposed technology shall represent data in compressed form and offering better encoding or decoding performance than the current ISO/IEC 23092 series for at least one data item or type of sequencing data in terms of complexity or compression ratio. | To improve encoding and decoding speed for some type of data and use cases. |
| 4.2.2 | The solution shall be compatible and interoperable with the ISO/IEC 23092 series. | To enable integration with the existing parts of ISO/IEC 23092. |
| 4.2.3 | Lossless compression of nucleotides sequences supporting a minimum of 5 symbols (A, C, G, T, N) |  |
| 4.2.4 | Lossless compression of quality scores shall be supported. |  |
| 4.2.5 | Lossy compression of quality scores shall be supported. Lossy compression should respect metrics that MPEG needs to identify with their associated boundaries. |  |
| 4.2.6 | Compressed data shall be structured so that parallel processing is enabled and compression efficiency is preserved | This is intended to enable efficient parallel processing of data without affecting compression efficiency. |
| 4.2.7 | Compressed data should be structured so that efficient querying of data is enabled and compression efficiency is preserved | Clustering applications could benefit from efficient querying. |
| 4.2.8 | Compressed data streaming shall be supported. | This implies that data consumption shall be possible before data transfer completion. |
| 4.2.9 | The association among headers, nucleotides and quality scores must be preserved |  |
| 4.2.10 | Enable efficient storage of the machine learning model using an existing standard | An existing standard is preferred |
| 4.2.11 | Enable efficient execution of the machine learning model | An abstract description of the algorithm independent of the library used to implement the model is preferred |

## Graph Reference Genome Support

|  |  |  |
| --- | --- | --- |
| **Req ID** | **Requirement** | **Rationale** |
| 4.3.1 | The new proposed technology shall represent data in compressed form | Support for compressed graph reference genomes. |
| 4.3.2 | The solution shall be compatible and interoperable with the ISO/IEC 23092 series. | To enable integration with the existing parts of ISO/IEC 23092. |
| 4.3.3 | Support for existing graph genome representation | Graph reference genome is still an emerging area and support for more than one representation is desirable |
| 4.3.4 | Lossless compression of nucleotides sequences supporting a minimum of 5 symbols (A, C, G, T, N) | For graph genome, the branches and tree structure must be represented |
| 4.3.5 | Lossless compression of quality scores shall be supported. |  |
| 4.3.6 | Lossy compression of quality scores shall be supported. Lossy compression should respect metrics that MPEG needs to identify with their associated boundaries. |  |
| 4.3.7 | Compressed data shall be structured so that parallel processing is enabled and compression efficiency is preserved | This is intended to enable efficient parallel processing of data without affecting compression efficiency. |
| 4.3.8 | Compressed data should be structured so that efficient querying of data is enabled and compression efficiency is preserved | Clustering applications could benefit from efficient querying. |
| 4.3.9 | Compressed data streaming shall be supported. | This implies that data consumption shall be possible before data transfer completion. |
| 4.3.10 | The association among headers, nucleotides and quality scores must be preserved |  |

## Support for Interchange of Genomic Information with Other Standards

|  |  |  |  |
| --- | --- | --- | --- |
| **Req ID** | **Requirement** | **Rationale** | |
| 4.4.1 | The solution shall be compatible and interoperable with the ISO/IEC 23092 series. | | To enable integration with the existing parts of ISO/IEC 23092. |
| 4.4.2 | Demonstrate MPEG-G metadata populated from FHIR resources | Show how to map elements of FHIR to MPEG-G Dataset metadata and/or annotation tables | |
| 4.4.3 | Demonstrate an MPEG-G API conforming FHIR API to the MPEG-G metadata API | Show how MPEG-G metadata could interoperate with FHIR Resources | |