 **ISO/IEC JTC 1/SC 29/ WG 11 N 19562**

**ISO/IEC JTC 1/SC 29/WG 11**

**Coding of moving pictures and audio**

**Convenorship: Japan (JISC)**

|  |  |
| --- | --- |
| **Document type:** | Approved WG 11 document |
| **Title:** | Draft Call for Evidence on new advanced  features and new technologies for MPEG-G |
| **Status:** | Approved |
| **Date of document:** | 2020-07-03 |
| **Source:** | Convenor, ISO/IEC JTC 1/SC 29/WG 11 |
| **No. of Pages:** | 8 |
| **Email of acting convenor** | ostermann@tnt.uni-hannover.de |
| **Committee URL:** | <http://isotc.iso.org/livelink/livelink/open/jtc1sc29> |

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2015/N19562**

**June-July 2020, Geneva On-Line**

|  |  |
| --- | --- |
| **Source:** | Requirements |
| **Status:** | Approved |
| **Title:** | **Draft Call for Evidence on new advanced features and new technologies for MPEG-G** |
| **Editor(s):** | Ray Krasinski, Marco Mattavelli |

**Abstract**

This document is a Call for Evidence (CfE) aiming at assessing the performance of new technologies that can demonstrate that:

* Current compression, transport and indexing technology of ISO/IEC 23092 series can be improved for sequencing data of current and/or emerging sequencing technologies
* Other functionalities can be added to the ones provided by ISO/IEC 23092 series

This is a draft version of a Call that will be issued at the 132nd MPEG meeting in October 2020.

# Introduction

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology [1], [2]. One challenge is to efficiently handle the increasing flood of sequencing data. A second challenge is the ability to process such a deluge of data in order to 1) increase the scientific knowledge of genome sequence information and 2) search genome databases for diagnosis and therapy purposes. High-performance compression of genomic data is required to reduce the storage size, increase transmission speed and reduce the cost of I/O bandwidth connecting the database and the processing facilities. Other important features are non-sequential access, integration with metadata, protection of the data supporting controlled selective access, integration with annotation data and indexing capabilities for raw data, metadata and annotations.

All these functionalities are already part of ISO/IEC 23092 (MPEG-G) standard series. However, new compression technologies, particularly applied to very long reads, can yield higher compression rate, support new functionality or improve performance of other metrics.

In this perspective this call is seeking the evidence, but not limited to this list, of technologies addressing:

* coding modes specialized for “Third Generation Sequencing” (long reads technologies) devices
* coding modes satisfying data access modalities required by machine learning approaches
* coding genome sequences and quality scores with higher compression performance than current ISO/IEC 23092 series
* support for representation and usage of graph genome references
* support of interfaces with existing standards for the interchange of clinical data

Companies and organizations are invited to submit proposals in response to this call.

The results of these tests will be made public, taking into account that no direct identification of any of the proponents will be made (unless it is specifically requested or authorized by a proponent to be explicitly identified). Prior to having evaluated the results of the tests, no commitment to any course of action regarding the proposed technology can be made.

Descriptions of proposals shall be registered as input documents to the proposal evaluation meeting that will take place in Cape Town (SA) on January 2021 (see timeline in section 3 and submission procedure in section 5). Proponents need to attend that meeting to present their proposals. Further information about logistical steps to attend the meeting can be obtained from the listed contact persons (see section 8).

# Purpose

The main purpose of this CfE is to assess whether new technologies can achieve better performance, but additional purposes are to:

* become aware of which additional functionalities (e.g. support of new type of sequencing data, non sequential access, increased indexing capabilities, lossy compression efficiency, etc. ) are provided by these new technologies
* collect information that may be used in drafting a future Call for Proposals

# Timeline

2020/07/03 Draft Call for Evidence

2020/10/11 Availability of test materials

2020/10/16 Final Call for Evidence

2021/01/06Submission of documents (for details of the submission process contact the persons listed in section 8 and see section 5 for the submission details).Submissions of executables is encouraged.

2021/01/07-15 Evaluation of the proposals

# Test Conditions

## Category 1: Compression of raw sequence data

### Anchors and Test Material

For raw sequence data the anchors to be considered are the compressed files using MPEG-G technology.

Test material files are described in output document N19561 produced at the 131st MPEG meeting and might be updated at the 132nd MPEG meeting.

## Category 2: Compression of aligned sequence data

### Anchors and Test Material

Anchors are contained in the revised document N19561 produced at the 131st MPEG meeting and might be updated at the 132nd MPEG meeting.

For aligned sequence data the anchors to be considered are the compressed files in MPEG-G format.

For the assessment of lossy compression of quality scores, details on the applications in scope, references datasets and additional test material is contained in the output document N19561.

## Category 3: Representation of graph genome references

Test material files information will be provided in an updated output document at the 132nd MPEG meeting.

# Requirements on Submissions

## Submission categories and details

Please note that a submission shall include the following mandatory elements

* Filled spreadsheets provided in Annex A with the following information
  + Input item
  + Output bitstream size for lossless compression
  + Compression factor
  + Weight (%) of the main classes of data in the compressed bitstream (in case a specific implementation does not support this partitioning, the submission of any information related to this metric is encouraged)
    - Reads headers/identifiers
    - Sequence reads
    - Quality scores
    - Any other metadata (identified as “Auxiliary data”)
  + Encoding time
  + Peak and average memory usage
  + A description of the platform used to run the decoding process (HW, OS, etc.)
* Encoded bit streams and – if available - decoder executable together with documentation on how to run the decoding process. Please add any other relevant information (e.g. which OS, any HW requirements?). Proponents need to
  + be present at the 133rd MPEG meeting to present their submissions
  + bring the encoded bitstreams on a physical storage device at the 133rd MPEG meeting
* Technical descriptions described below:
  + Proponents shall provide a description of the platform and methodology used to determine the time.  To help evaluation, a description of software libraries (e.g. low level libraries) used, if any, is encouraged.
  + A technical description of the proposal sufficient for the conceptual understanding and generation of equivalent performance results by experts and for conveying the degree of optimization required to replicate the performance is desirable.
  + Supported platforms (HW requirements and OS)
  + Configuration parameters (if any)
  + Expected resources (e.g. RAM, disk space etc.) usage of the decoder
  + Any other information needed to run the decoding process
  + Support for any feature such as:
    - Non sequential access
      * Example: extract all the reads (and the associated metadata such as QS or identifiers) that map within a given interval of the reference genome
    - Support of more than 5 symbols (A, C, G, T, N) alphabets
    - Encoding of additional metadata (extensibility)
    - Lossy compression of
      * Quality scores
      * Reads identifiers

For the assessment of lossy compression of quality scores, details on the applications in scope, references datasets and additional test material is contained in the output document N15739.

Any document included in the submissions shall be registered according to the usual MPEG submission procedure, i.e.:

**For MPEG members.**

Please follow the usual input documents submission procedure

* Online submission portal: <http://wg11.sc29.org/>
  + Click on “Next meeting” (username and password required, ask your national Head of Delegation for credentials)
  + Click on “Register a new document”
  + Please use the [input document template](http://wg11.sc29.org/Templates/mxxxx.dot) available on the MPEG portal

**For non MPEG members**

Please use the [input document template](http://wg11.sc29.org/Templates/mxxxx.dot) available on the MPEG portal and contact the reference persons listed in section 8.

For both MPEG members and non MPEG members the deadlines are the following:

* Document registration deadline: Monday 2021-01-04 23:59:59 CET
* Document upload deadline: Wednesday 2021-01-06 23:59:59 CET

## Binaries and bitstreams

* Proponents are encouraged to allow other committee participants to have access, on a temporary or permanent basis, to their encoded bitstreams and binary executables (if available).
* Proponents are encouraged to submit a statement about the programming language in the software is written, e.g. C/C++ and platforms on which the binaries were compiled.

# IPR

Proponents are advised that this call is being made subject to the common patent policy of ITU-T/ITU-R/ISO/IEC (http://www.itu.int/en/ITU-T/ipr/Pages/policy.aspx) and other established policies of these standardization organizations. The persons named below as contacts can assist potential submitters in identifying the relevant policy information.

# Fees

None.

# Contact(s)

Joern Ostermann (ostermann@TNT.UNI-HANNOVER.DE), Marco Mattavelli (marco.mattavelli@epfl.ch)

# References

|  |  |
| --- | --- |
| [1] | S. D. Kahn, “On the Future of Genomic Data,” *Science,* vol. 331, pp. 728-729, 2011. |
| [2] | S. Wandelt, M. Bux and U. Leser, “Trends in Genome Compression,” *Journal of Current Bioinformatics,* 2013. |

**ANNEX A**

Spreadsheets to be completed and submitted by the respondents are provided in the attached Excel file.

**Attachments:**

1. Performance measurement spreadsheet.