 **ISO/IEC JTC 1/SC 29/ WG 11 N19563**

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

**Coding of moving pictures and audio**

**Convenorship: Japan (JISC)**

|  |  |
| --- | --- |
| **Document type:** | Approved WG 11 document |
| **Title:** | **Joint Call for Proposals for extensions and improvements of ISO/IEC 23092 series** |
| **Status:** | Approved |
| **Date of document:** | 2020-07-03 |
| **Source:** | Convenor, ISO/IEC JTC 1/SC 29/WG 11 |
| **No. of Pages:** | 11 |
| **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 JTC 1/SC 29/WG 11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC 1/SC 29/WG 11 N19563**

**Geneva, Switzerland (Virtual Meeting) – June - July 2020**

|  |  |
| --- | --- |
| **Title:** | **Joint Call for Proposals for extensions and improvements of ISO/IEC 23092 series** |
| **Source:** | **Requirements** |

**Joint Call for Proposals for extensions and improvements of ISO/IEC 23092 series**

# Introduction

The development of high-throughput sequencing technologies has paved the way for the usage of genomic information as everyday practice in several fields, but the growing volume of data generated has become a serious obstacle to a wider diffusion [1] [2]. An appropriate representation and an efficient compression of genomic data has been addressed in a previous joint call in May 2016 between JTC1/SC29/WG11 (MPEG) and TC276/WG5 whose results have led to the development of the ISO/IEC 23092 standards [3] series. Another call has been issued in July 2019 for the compressed representation of genomic annotations, MPEG-G Part 6 and the proposed technology is undergoing core experiments and validation phases with the objective of finalizing a standard specification in 2021.

Current ISO/IEC 23092 series provide a file and transport format, compression technology, metadata specifications, protection support and standard APIs for the access of sequencing data and annotations in native compressed formats will be soon available with the planned standardization of MPEG-G Part 6.

ISO/IEC JTC 1/SC 29/WG 11 (MPEG) has the mission of developing standards for the coded representation and compression of digital audio, video, related data and metadata including also other types of data beyond media, such as genomic data, requiring efficient compression and processing. In its 30 years of activity MPEG has developed many generations of compression standards.

ISO/TC 276 works on standardization in the field of biotechnology processes that include analytical methods (Working Group 3) and data processing and integration (Working Group 5).

By combining their respective expertise, ISO/TC 276 and MPEG have the possibility to further develop and improve the ISO/IEC 23092 standard series by including new technologies and features.

With the objective of increasing the performance and extend the functionality set of the current MPEG-G technologies, preliminary investigations have shown the evidence possible improvement and extensions. New algorithms providing increases of current achievable compression performance, processing speed and more appropriate support of long read sequencing data have been submitted for consideration to the MPEG committee.

Thus, this open joint call intends to solicit the submission of technologies addressing the following items:

* low complexity coding modes
* searches for genomic data and metadata in the compressed domain

Both technologies are expected to improve critical performance aspects and remove some of current limitations occurring when processing genomic data.

In particular, low complexity coding modes are expected to directly improve the speed of encoding and decoding, thus providing faster access to data and reduced latencies for transport and storage of massive amounts of data. In another dimension, lower complexity is also expected to increase the parallelization potential of current implementations and to provide a wider range of performance improvements for both encoder and decoder implementations.

The possibility of applying genome sequence searches directly in the compression domain provides new perspectives for efficient queries of both aligned and unaligned data. The searches, combined with the annotation information currently under standardization in ISO/IEC23092-6 will provide a unified framework for browsing genomic data from raw reads up to tertiary analysis results.

ISO/IEC provides a framework for the improvement of the ISO/IEC 23092 standards series based on the following steps:

* an open call to any party possessing technologies satisfying all or a subset of identified requirements
* a fair assessment of the performance of each submission
* the identification of the most promising technologies
* their combination in a Test Model as a platform for collaboration
* the progressive improvement of the Test Model via Core Experiments
* the approval of the standard following the established ISO/IEC procedure([[1]](#footnote-1))
* the public availability of an informative software to associate coded information and annotations to genomic sequencing data information and a normative software to decompress genomic information and annotations.

# Call for Proposals background

## Compression as a critical element in genomic information processing

The sequencing of the genetic information of human genome has become affordable due to high-throughput sequencing technology. This opens new perspectives for the diagnosis and successful treatment of cancer and other genetic illnesses. However, there remain challenges, scientific as well as computational, that need to be addressed for this technology to find its way into everyday practice in healthcare and medicine. These challenges are recognized by initiatives such as the Genomic Standards Consortium (GSC), https://gensc.org. The first challenge is to cope with the flood of sequencing data. For instance, a database covering the inhabitants of a small country like Switzerland would need to store a staggering amount of data, about 2'335'740 Terabytes. The second challenge is the ability to process such a volume 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 the I/O bandwidth required to connect the database with processing facilities.

The current trends in sequencing data generation show clearly that the storage and transfer (bandwidth) costs are comparable to the costs of sequencing [1] [2]. In addition, the lack of interoperability between compressed data storage facilities and the cost of accessing increasing volumes of data represent a major obstacle to the implementation of economical and efficient genome analysis applications for personalized medicine, early diagnostics and drugs discovery. The picture is made even more complex by the proliferation of a number of solutions and file formats, each one addressing a particular use case or step of bioinformatics data analysis workflows for high-throughput sequencing data. While some formats such as SAM [4] are supported by compressed versions such as BAM or CRAM, they do not offer the full functionality needed to support the entire data chain from raw data to medical reports. That implies a need for frequent format conversions, especially when data moves from one processing step to the next one, when it is used for a different application than the original one, or when it is transmitted from one user to another. Thus, the absence of a unified exchange format able to handle the majority of data structures relevant to high-throughput sequencing data analysis hampers the possibility of writing streamlined analysis workflows; it stifles data reuse; and it impairs efficient communication between different actors in the field.

The ISO/IEC 23092 series strives to address most of these issues and enforce best practices in the field. It has been designed with interoperability and meaningful data sharing in mind; it offers automatic advantages such as transparent transport and a robust configurable protection layer. The incorporation of novel use cases in the standard will result in an even more attractive solution to store and exchange high-throughput data in the field of bioinformatics.

# Open standard development process

The process that will be followed by ISO/IEC JTC 1/SC 29/WG 11 (MPEG) for the development of the extensions of the ISO/IEC 23092 standard series, is based on the well-established and successful approach refined during the last three decades:

* A Call for Proposals is issued (this document) which is open to any interested party ready to accept ISO/IEC IP policy (see later in this document); acceptable proposals can satisfy all or only a subset of the requirements; interested parties which are not members of MPEG have the possibility to address any questions and requests of documentation to a dedicated contact person mentioned below in section 7;
* The evaluation criteria and process elaborated by ISO/TC 276 and MPEG experts and approved by the delegates have been published in parallel to the Final Call for Proposals [5]; furthermore, proponents may provide comments on the requirements and the evaluation process in the context of the application scenarios covered by the Call;
* The assessment of the received proposals will identify either a proposal that will become the Test Model, or a set of the most promising technologies that will be combined into the Test Model implementing the extensions to the current ISO/IEC 23092 series.;

The Test Model is intended as an initial step and a platform for further collaboration; throughout the working period that will follow the assessment of the proposals, the Test Model will be progressively improved through the specification of several Core Experiments; this process allows integrating further relevant enhancements proposed and identified during the whole period. This working period usually lasts several months;

* At the end of the process described above, the new standard will be approved according to the established ISO/IEC procedure described at <https://www.iso.org/developing-standards.html> and <https://www.iso.org/stages-and-resources-for-standards-development.html> ;
* In MPEG "standard" typically means a description of the normative sequence of operations to be performed on compressed and/or transported data in order to reconstruct data into their original (uncompressed) form. This is what MPEG calls "decoding" process. A reference implementation of a "decoder" is typically developed and actually assumes a "normative" status. Conversely one or more entire or partial implementations of "encoders" are also developed as reference examples, but they assume only an "informative" status.

# Technology Solicited by this Call

Responses are solicited that propose technologies for Coding of Genomic Sequencing data covering any of these aspects:

* lower complexity coding modes for reads, quality values and read names in comparison to the technology included in current ISO/IEC 23092 series specification
* lower complexity entropy coding modes for descriptors representing reads, quality values and read names in comparison to the technology included in current ISO/IEC 23092 series specification
* technologies for compressing and indexing aligned read features
* technologies for compressing and indexing nonaligned reads features
* technologies supporting efficient genome sequences searches in the compressed domain for both unaligned and aligned reads
* technologies supporting advanced sequencing data and metadata querying, searching, filtering and associated APIs

For the detailed requirements to be fulfilled by responses to this CfP, refer to the requirements document N19564 [6].

# Source Code and IPR

Proponents are advised that, upon acceptance into the standardization process, they may be required to make available source code software for both encoders, decoders as well as any other part of the technology that is necessary to finalize a complete version of the reference software. This code will be included in the standard as Reference Software to be released under the Reference Software Copyright License in annex (N15898). If proponents feel that any aspects of their technology should not be made available in source code, they should clearly state which aspects and why.

Furthermore, proponents are advised that this Call is being made under the auspices of ISO/IEC, and as such, subject to the ITU-T/ITU-R/ISO/IEC Intellectual Property Rights Policy as approved by the ISO, IEC and ITU councils[[2]](#footnote-2).

In that respect proponents are invited to submit together with the response to the call an IP declaration as suggested in Annex A of this document.

In order to encourage the widest responses to the Call, we encourage “no-license” or “type-1” contributions. With “type-1” we refer to the option mentioned as box 1 in Annex A.

In the case alternative solutions achieve an equivalent level of satisfaction of requirements and present equivalent performance according to core experiment results, “no-license” or “type-1” solutions would be preferred.

Proponents are advised that, upon acceptance for further evaluation, it will be required that certain parts of any technology proposed to be made available in source code format to participants in the core experiments process and for potential inclusion in the prospective standard as reference software. When a particular technology is a candidate for further evaluation, commitment to providing such software is a condition of participation. The software shall produce identical results to those submitted to the test. Additionally, submission of improvements (bug fixes, etc.) is encouraged.

# Timetable and Procedures

The following estimated milestones are planned:

* Draft CfP Issued - Brussels, Belgium, 17th January 2020
* CfP Issued – Geneva Virtual Meeting, 3rd July 2020
* CfP Responses: Deadline for Submissions –Test Model 0 and preliminary Working Draft – Rennes, France, 12th October 2020.
* Technology identification and selection – Committee Draft – Cape Town, South Africa 15th January 2021
* Draft International Standard – April 2022

All communications concerning the Call and responses thereto should be addressed to the CfP Contacts (listed below), and communications are preferred in electronic form, via email.

Interested parties should approach the CfP Contacts for assistance regarding all aspects of their submission and subsequent attendance at ISO meetings, which may involve explaining how they can become accredited to attend the meetings.

## Registration

There is no need to register interest in responding to this call.

## Items to be submitted

CfP respondents should submit the following:

* A description of the technology having sufficient detail to permit technical discussions.
* A description of how the proposed technology can be implemented as backward compatible extension of the ISO/IEC 23092 series.
* A list of satisfied requirements
* A description of the encoding technology
* Evidence of the performance of the technology, including compression factor and any other meaningful metrics the proponents deem appropriate
* A self-assessment template filled with the relevant information for the proposed technology that include a comparison on how the proposed technology outperforms the current standard performance.
* If appropriate, the proponent may provide comments on the requirements and the evaluation process in context of the application scenarios or use cases.

The proponent’s documents should be provided in Microsoft Word format.

**Important dates to answer this Call for Proposals are:**

* **5th October 2020 23:59 CET: Proponents must register documentation related to their submission as MPEG input contribution on the MPEG documents repository (**[**http://wg11.sc29.org/**](http://wg11.sc29.org/)**)**
* **7th October 2020 23:59 CET: Proponents must upload the complete documentation to the MPEG documents repository (**[**http://wg11.sc29.org/**](http://wg11.sc29.org/)**)**

**To support evaluation tests proponents should submit:**

* **Executables – Decoders and encoders must be delivered to the CfP Contact as executables on either the Linux/Intel or Windows platforms (statically linked libraries are required for all non-standard libraries). All executables should preferably have command-line interface (i.e. no GUI).**
* **Bitstreams - Compressed data must be supplied corresponding to each individual test item as described in the Evaluation Criteria document (N19565)** [5] **that has been issued at the 131th MPEG meeting.**

Such items must be available for evaluation on 8th October 2020 at the beginning of the AhG meeting prior to the 132nd MPEG meeting (see section 6.4 for more details).

## Evaluation Criteria

Evaluation is based on the fulfillment of the requirements and measurable degree of fulfillment if applicable [6]. Proposals are not required to meet all requirements.

An evaluation procedure document (N19565) has been issued at the 131st MPEG meeting and the criteria defined therein will be used in the technology selection process [5].

## Participation

Respondents to the CfP are required to attend the AhG meeting in Rennes (France) to present and discuss details of their proposals. This meeting will be held on 8th – 9th October 2020 before the main MPEG meeting (Monday 12th October 2020 to Friday, 16th October 2020).

## Preliminary Evaluation

At the kickoff meeting, the group will conduct a preliminary evaluation of submissions to check their compliance with the Requirements outlined in this document and procedures described in the associated evaluation procedure document. Submissions that are compliant will undergo further evaluation.

## Selection of Technology

At the kickoff meeting, the final selection of the proponent technology that will become Test Model Zero (TM0), and which will be the start of the standardization phase, will be based on the judgment and consensus of the experts in the joint working group.

## Test Model and Core Experiments

Two working tools play a major role in the collaborative development phase that follows the initial competitive phase: the Test Model and Core Experiments (CE).

The best technology, as identified in the evaluation process, will be selected as TM0 and be the basis for subsequent core experiments. Proponents whose technology is selected as TM0 and all proponents participating in the subsequent core experiment process shall supply a detailed description of their technology.

### Test Model

A Test Model is a complete framework such that an experiment performed by multiple independent parties will produce essentially identical results. The TM enables the checking of the relative performance of different tools, as well as improving the performance of selected tools. The TM will be built after screening the proposals answering the CfP. The first TM will not the best proposal, but a combination of the best tools, independently of the proposal that they belonged to. The TM will include normative and non-normative tools to create the “common framework” that allows performing adequate evaluation and comparison of tools targeting the continuous improvement of the technology included in the TM. After the establishment of the first TM new tools can be proposed and evaluated inside the TM following a core experiment procedure. The TM will evolve through versions as core experiments verify the inclusion of new techniques, or prove that included techniques should be substituted. At each TM version, only the best performing tools will be part of the TM. If any part of a proposal will be selected for inclusion in the TM, the proposer will have to provide the corresponding source code for integration into the TM software under the conditions specified by the ISO/IEC Intellectual Property Rights Policy.

### Core Experiments

The improvement of the TM will start with a first set of core experiments defined at the conclusion of the evaluation of the proposals. The core experiments process allows for multiple, independent, directly comparable experiments to be performed to determine whether or not a proposed tool has merit. Proposed tools target the substitution of a tool in the TM or the direct inclusion in the TM to provide a new relevant functionality or improved performance. Improvements and additions to the TMs will be decided based on the results of core experiments.

A core experiment has to be completely and uniquely defined, so that the results are unambiguous.

In addition to the specification of the tool to be evaluated, a core experiment also specifies the conditions to be used, again so the results can be compared. A core experiment is proposed by one or more experts and is accepted by consensus, providing that two or more independent experts agree to perform the experiment.

It is important to realize that the Core Experiments will not end up in the standard itself, as these are just working tools to ease the development process.

# Call for Proposals Contact information

For any other questions about the call, test conditions, required software or test sequences please contact:

**Joern Ostermann, MPEG Requirements Group Chair**

**Email: ostermann@tnt.uni-hannover.de**

# Bibliography

|  |  |
| --- | --- |
| [1] | S. D. Kahn, "On the Future of Genomic Data," *Science,* vol. 331, pp. 728-729, 2011. |
| [2] | Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron and G. E. Robinson, "Big Data: Astronomical or Genomical?," *PLOS Biology,* 2015. |
| [3] | ISO/IEC SC29WG11, ISO/IEC 23092 Series, ISO/IEC, 2019. |
| [4] | The SAM/BAM Format Specification Working Group, "https://samtools.github.io/hts-specs/SAMv1.pdf," 2020. [Online]. |
| [5] | ISO/IEC JTC 1/SC 29/WG 11, "MPEG2020/N19565 - Evaluation procedure for the Call for Proposals for extensions and improvements of ISO/IEC 23092 series," in *Published at the 131th MPEG meeting*, Geneva - Virtual Meeting, July 2020. |
| [6] | ISO/IEC JTC 1/SC 29/WG 11 , "MPEG2020/N19564 - Final Requirements for ISO/IEC 23092 series extensions," in *131th MPEG meeting*, Geneva - Virtual Meeting, July 2020. |

**Annex A:   
Example of declaration of readiness to grant a license**

XYZ Organization may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, would be prepared to (check at least one of the following items):

* make them available free of charge (per box 1 of the ISO/IEC patent statement and licensing declaration form)
* grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ISO/IEC International Standard (per box 2 of the ISO/IEC patent statement and licensing declaration form)

However, XYZ Organization is aware that other entities may also have current or pending patent rights relating to the technology described in this contribution.

1. http://www.iso.org/iso/home/standards\_development/resources-for-technical-work/support-for-developing-standards.htm [↑](#footnote-ref-1)
2. http://www.itu.int/en/ITU-T/ipr/Pages/policy.aspx [↑](#footnote-ref-2)