

**ISO/IEC JTC 1/SC 29/WG 11 N 18648**

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

**Coding of moving pictures and audio**

**Convenorship: UNI (Italy)**

**Document type: Approved WG 11 document**

**Title: Joint Call for Proposals for Coding of Genomic Annotations**

**Status: Approved**

**Date of document: 2019-07-12**

**Source: Requirements**

**Expected action:**

**No. of pages: 11**

**Email of convenor: leonardo@chiariglione.org**

**Committee URL: mpeg.chiariglione.org**

**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 N18648**

**Gothenburg, Sweden – July 2019**

|  |  |
| --- | --- |
| **Title:** | **Joint Call for Proposals for Coding of Genomic Annotations** |
| **Source:** | **Requirements** |

**Joint Call for Proposals for Coding of Genomic Annotations**

# 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 family of ISO/IEC 23092 standards [3]. They provide a file and transport format, compression technology, metadata specifications, protection support and standard APIs for the access of sequencing data in the native compressed format.

An important element in the usage of sequencing data is the association of the data with the results of the analysis and annotations that are generated by processing pipelines and analysts. At the moment such association happens as a separate step, and there is no way of representing sequencing data and meta-information derived from them in the same format. That is why other requirements have been identified that current raw data and aligned data formats do not fulfill (see documents N18080 [4] and N18647 [5]).

ISO/IEC JTC 1/SC 29/WG 11 (MPEG) has the mission to develop standards for the coded representation and compression of digital audio, video, related data and metadata including also other types of data beyond media 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 the ISO/IEC 23092 standard series by including compressed representations of genomic annotations linked to the compressed representation of raw sequencing data and metadata, thus providing new effective solutions to the stated problem.

ISO/IEC provides a framework for the development of a genome annotation standard 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

## Annotations as a critical element in genomic information processing

While the existing parts 1-5 of the ISO/IEC 23092 (MPEG-G) standard [3] concern themselves with the representation of genomic information derived from the so-called *primary analysis* of sequencing data – sequencing reads and qualities, and their storage and alignment to a reference set of sequences – that is only the first step in a long series. In particular, the results of primary analysis are usually processed further in order to obtain higher-level information.

For instance, the alignments of RNA-sequencing reads falling across a *gene* (i.e., a set of specific intervals of the reference genome) might be counted in order to establish the expression level of such a gene in the given biological condition. Several conditions (and hence several sets of reads coming from different experiments) might need to be compared. In general, the process of aggregating information deduced from single reads and their alignments to the genome into more complex results is called *secondary analysis*.

In most biological studies based on sequencing protocols, the output of secondary analysis is usually represented as different types of *annotations* (meta-information) all associated to one or more *intervals* on the reference sequences.

An interval is typically identified by the name of the *sequence* in the reference, the molecule *strand* (can be forward or reverse), and a lower (5’) and a higher (3’) *positions* specifying the base range.

|  |  |  |  |
| --- | --- | --- | --- |
| **Interval** | | | |
| sequence | strand | lower position | higher position |

Intervals are the natural way to talk about features localized on the genome, be them the number of aligning reads (or read coverage), variants, genes, regions of the genome binding to proteins, regions that perform a specific function in the architecture of the genome, and so on. Intervals can be as short as one single base, but they usually span much larger scales, allowing people to associate meta-information to its correct scale.

Making a parallel with video coding, and thinking of movements along the genome as movements in time, the reads, and their alignments as derived from primary analysis could be seen as video information, while meta-information based on intervals, i.e. the results of secondary analysis, would be similar to subtitles.

A complex study based on sequencing technologies can include a number of experiments, each one based on a different sequencing-derived protocol that probes a different compartment/function of the cell – RNA-sequencing, ChIP-sequencing to explore binding of proteins to the genome, bisulfite sequencing to ascertain genome methylation, and so on. After primary and secondary analysis have been performed for each experiment, the results are usually visualized in graphical form on a *genome browser*, which adopts a set of standardized ways to show the information in relation to its placement on the genomes – there will be different *tracks*, one or more per sequencing experiment, displaying things such as transcript presence and structure, presence of sequence variants in the population, intensity of protein binding to each position of the genome, coverage of RNA-seq reads, and so on.

At the moment each source of information – functional annotation of the reference genome, signal intensity for ChIP-seq, coverage for DNA- or RNA-seq, databases of variants – is represented using a number of completely different formats. In addition, the precise semantics of such formats is usually left unspecified, leaving the field open to the presence of different, and slightly incompatible, formats for each information source. That places onto scientists working on integrative analysis a heavy burden, forcing them to convert information frequently, complicating set-up when sets of experiments need to be visualized together, and making the exchange of information in most cases unnecessarily complicated.

That is why integrating interval-based information into the ISO/IEC 23092 (MPEG-G) standard [3] would be extremely beneficial to the field. The addition of such a capability would make it possible to perform all the stages of the analysis of sequencing-based experiments – sequencing, primary and secondary analysis, visualization – in the same format, with further stages of the analysis resulting in straightforward additions to an already existing MPEG-G container. MPEG-G containers may also contain only annotational data.

## Elaboration of requirements

Since October 2018 [4] work on the identification of the most important types of annotation data and the definition of a list of requirements for their efficient compressed data representation has been carried out. This work focused on identifying file formats that are used by the community to store biological and quantitative information relevant to genomic intervals; on categorizing such information as reported in (N18647 [5], table 2) and ISO CD 20691 [6]; and on selecting a representative set of example files. Requirements related to the efficient transport of compressed annotation data have been identified as well.

Most of the information associated with biological features is either some kind of classification based on existing biological ontologies [6][7] or some kind of numerical data. Depending on the scope and scale of the data (visualization of data on a genomic browser, which is the typical end point of such formats, can happen at different resolution levels) some of the information, in particular pre-computed large-scale snapshot, might be encoded with different level of accuracy. Therefore requirements for solutions addressing compression of genomic metadata have been identified, according to the nature of the metadata.

Experts in several domains including bioinformatics, biology, information theory, telecommunication, data compression, data storage and information security have participated in this process.

The requirements identified so far are listed in detail in the output document N18647 [5].

# Open standard development process

The process that will be followed for the development of a new open standard on Coding of Genomic Annotations is based on the well-established and successful approach refined during the last three decades:

* A Draft 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 will be published in parallel to the Final Call for Proposals [8]; 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;
* 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 Annotations covering at least one of these seven aspects:

1. Mapping statistics
2. Quantitative browser tracks
3. Variants
4. Genome functional annotations
5. Expression data associated to genome features
6. Hi-C-like experiments results
7. Transport of and access to data belonging to the previous categories.

For the detailed requirements to be fulfilled by responses refer to the requirements document N18647 [5].

# 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 certain parts of their technology. 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.

# Timetable and Procedures

The following estimated milestones are planned:

* Draft CfP Issued - Geneva, Switzerland, 29th March 2019
* CfP Issued - Goteborg, Sweden, 12th July 2019
* CfP Responses: Deadline for Submissions 8th January 2020.
* Technology identification and selection - Brussels, Belgium, 17th January, 2020
* Test Model 0 and preliminary Working Draft – Brussels, Belgium 17th January 2020
* Committee Draft – Geneva, Switzerland, 3rd July 2020
* Draft International Standard – Cape Town, South Africa 15th January 2021
* Final Draft International Standard – July 2021

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 definition of the annotation format addressed
* 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
* The self-assessment template filled with the relevant information for the proposed technology that can be found in the Evaluation Criteria document (N18649) [8]
* 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:**

* **6th January 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/)**)**
* **8th January 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 (N18649)** [8]**.**

Such items must be available for evaluation on Saturday 11th January 2020 at the beginning of the AhG meeting prior to the 129th 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. Proposals are not required to meet all requirements.

An evaluation procedure document (N18649) has been issued at the 127th MPEG meeting and the criteria defined therein will be used in the technology selection process.

## Participation

Respondents to the CfP are required to attend the AhG meeting in Brussels (Belgium) to present and discuss details of their proposals. This meeting will be held on the weekend (11th – 12th January 2020) before the main MPEG meeting (Monday 13th January 2020 to Friday, 17th January 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 in 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

To register for this call please send an e-mail to the main MPEG Genome Compression mail reflector ([genome\_compression@listes.epfl.ch](mailto:genome_compression@listes.epfl.ch)) before the document registration deadline on the **6th January 2020 23:59 CET**.

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] | MPEG Requirements, "ISO/IEC JTC 1/SC 29/WG 11 MPEG2018/N18135 - Thoughts on future standardisation activities in the domain of genomic representation," in *124th MPEG Meeting*, Macao, October 2018. |
| [5] | MPEG Requirements, "ISO/IEC JTC 1/SC 29/WG 11 MPEG2019/N18647 - Requirements for ISO/IEC 23092-6 Coding of Genomic Annotations Call for Proposals," in *127th MPEG meeting*, Gothenburg, July 2019. |
| [6] | ISO 20691 Biotechnology — Requirements for data formatting and description in the life sciences for downstream data processing and integration workflows, 2019. |
| [7] | The Gene Ontology Consortium: "The Gene Ontology project in 2008". Nucleic Acids Research. 36 (Database issue): D440–4. doi:10.1093/nar/gkm883, January 2008. |
| [8] | ISO/IEC JTC 1/SC 29/WG 11, "N18649 - Evaluation Procedure for the Call for Proposals for ISO/IEC 23092-6," in *127th MPEG meeting* , Gothenburg, July 2019. |

**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)