 ISO/IEC JTC 1/SC 29/WG 08 N00179

**ISO/IEC JTC 1/SC 29/WG 08  
MPEG Genomic Coding  
Convenorship: SNV (Switzerland)**

**Document type:** Output Document

**Title:** Approaches under discussion for the MPEG-G support of provenance information

**Status:** Approved

**Date of document:** 2023-07-28

**Source:** ISO/IEC JTC 1/SC 29/WG 08

**No. of pages:** 3 (with cover page)

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

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

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 08 MPEG GENOMIC CODING**

**ISO/IEC JTC 1/SC 29/WG 08 N00179**

**July 2023, Geneva (Switzerland)**

|  |  |
| --- | --- |
| **Title** | **Approaches under discussion for the MPEG-G support of provenance information** |
| **Source** | **WG 08, MPEG Genomic Coding** |
| **Status** | **Approved** |
| **Serial Number** | **22915** |

# Context

Output document N00166 summarized the group reflections on how to establish the link between ISO 23494 and MPEG-G. Four paths were described:

* Link from an MPEG-G file to an external resource of encoded W3C-PROV data
* Link from an MPEG-G file to an external W3C-PROV service
* Provenance data is fully included in the MPEG-G file
* Provenance data is provided in a file external to the MPEG-G file

Input document M63434 discussed the use cases for provenance data in the context of genomic files:

* Use of provenance data as a tool for identifying relevant genomic data for research
* Use of provenance data as an auditing input

From the discussion in the meeting, a further use case was introduced:

* Use of provenance data as a documentation unit integrated to the file, to make the file self-contained and suitable to a medical setting.

Input document M63434 also described the integration of the provenance data in a GraphQL database, which provides access to the data as a service with a range of filtering criteria as wide as possible.

# Discussion

Depending on the use case, one of the previously approaches is preferable over the others. Thus, it seems preferable to keep the different routes as open as possible.

When using provenance data to identify relevant genomic data for a research purpose, it is important to be able to query this information prior to downloading the genomic data: downloading genomic data has a cost in terms of storage capacity, bandwidth and time required. Thus, for this context, a solution where the provenance data is separated from the MPEG-G file(s), is more in line with the requirements of the use case.

Instead, if the provenance data is to be used as an auditing tool, or if it must be stored within the genomic file to ensure medical relevance, then a solution where the provenance data is stored within the file is preferable.

Whereas the second use case must be addressed with the provenance data represented as a serialized self-contained unit, the first one can either be based on the same approach, or as query point, as described in the input document M63434.

A (still) unanswered question is whether it is realistic to pursue both the strategy of provenance as a service and provenance as a serialized self-contained unit. By leveraging a well-defined schema, it should be feasible to easily transition from one to the other: for example, by defining the service as returning XML content, and the serialized unit as being XML based.

# Conclusion

TC276 is currently standardizing the core set of provenance data. The objective of the collaboration between TC276 and WG8 is to ensure the applicability of the extension to the provenance data specialized for genomic information.

Once the technical requirements are identified, the best approach to addressing the use cases should be derived from it.