 ISO/IEC JTC 1/SC 29/WG 08 N00166

**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:** 2023-04-27

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

**No. of pages:** xx(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 N00166**

**April 2023, Antalya (Turkey)**

|  |  |
| --- | --- |
| **Title** | **Approaches under discussion for the MPEG-G support of provenance information** |
| **Source** | **WG 08, MPEG Genomic Coding** |
| **Status** |  |
| **Serial Number** | **22798** |

Table of Contents

[1 Context 2](#_Toc133490083)

[2 Discussion 2](#_Toc133490084)

[3 Conclusion 2](#_Toc133490085)

[3.1 Method 1.1: Link to an external resource of encoded W3C-PROV data 3](#_Toc133490086)

[3.2 Method 1.2: Link to an external W3C-PROV service 3](#_Toc133490087)

[3.3 Method 2: Provenance data is fully included in the MPEG-G file 3](#_Toc133490088)

[3.4 Method 3: Provenance data is provided in a file external to the MPEG-G file 4](#_Toc133490089)

# Context

There are currently revisions to MPEG-G part 3 intending to include some provenance information, more specifically geared towards the computation steps which were used to reach the data stored in the MPEG-G file. Through a collaboration between WG8 and ISO/TC 276, we are now considering how to achieve a mutual support between MPEG-G and their standard on provenance (ISO 23494). Part 2 of ISO 23494 (Common Provenance Model) relies on W3C-PROV and gives support for new types of provenance information such as authority delegation, or provenance of the tools used in the computations.

Input document M63434 discusses the interaction strategies between MPEG-G and ISO 23494. The document describes two main lines: how to integrate ISO 23494 within MPEG-G, and how to reference MPEG-G data from ISO 23494.

Regarding the first point, the input document proposes to use the W3C-XML representation of ISO 23494 to integrate it within MPEG-G, as this would allow to leverage some technologies already present within MPEG-G, as XML encryption and signatures. W3C-XML is one of the current or intended ways to store or interact with ISO 23494 data.

The second point, dedicated on how to reference MPEG-G data from ISO 23494, describes which identifiers are required to point to the correct content in an MPEG-G file, depending on the intent (e.g. pointing to one dataset, or a collection of datasets).

The input document also discusses how to maintain the provenance data up-to-date within the MPEG-G file: new outputs can be generated from the content stored in the MPEG-G, and updating the provenance information can be important for certain use cases requiring completeness.

# Discussion

The mentioned input document was discussed during the meeting. The main points of discussion concerned the context in which the provenance data contained in MPEG-G is going to be used. As MPEG-G can be used in different settings, the requirements will differ between them. The most impactful requirements identified were:

* Whether MPEG-G file is going to be used in a setting allowing access to remote resources (i.e. accessible on the internet), access to other local files only (e.g. in a cluster environment), or even if all the data must be self-contained in a unique archive
* Whether the provenance data must be up-to-date

Another point of discussion was at what level of the hierarchy of elements within MPEG-G the linkage should be established. This point of discussion has implications on which parts of the MPEG-G standards are affected, and on the complete roadmap for provenance information.

# Conclusion

The group has identified three possible paths covering different answers to the identified requirements. In particular, four (one has two options) alternatives are initially considered on how to add provenance information to MPEG-G. They are described in the next subsections.

## Method 1.1: Link to an external resource of encoded W3C-PROV data

Provenance data is referenced through a link pointing to a remote resource. This resource is an encoding of the W3C-PROV data, using one of the defined encodings of W3C-PROV data such as PROV XML. The link is stored in the MPEG-G file. The Provenance data is owned and maintained by the original authors of the provenance data.

The semantics is provided in ISO 23494, the syntax is provided by W3C Prov-XML, and the container storing the link by MPEG-G Part 3 (and possibly Part 1).

In any case, an initial solution could allow any external encoding identified with an URI. Therefore, a new formal container would not be necessary, but just an additional element in the information metadata box at the dataset group and/or dataset level, specified in MPEG-G part 3.

**Use cases:**

* Always up-to-date information
* Requires external access

**Limitations:**

* All bio data in one MPEG-G is linked to one PROV context
* It relies on W3C encoding (or an external encoding, in general)

## Method 1.2: Link to an external W3C-PROV service

Provenance data is referenced through a link pointing to a remote service. The service complies with PROV-AQ. The link is stored in the MPEG-G file. The Provenance data is owned and maintained by the original authors of the provenance data.

The semantics is provided in ISO 23494, the syntax is provided by W3C Prov-XML, and the container storing the link by MPEG-G Part 3 (and possibly Part 1).

Again, an initial solution could allow any external encoding identified with an URI, meaning that minor changes are needed in MPEG-G, as previously indicated.

**Use cases:**

* Always up-to-date information
* Requires external access

**Limitations:**

* All bio data in one MPEG-G is linked to one PROV context
* It relies on W3C encoding (or an external encoding, in general)

## Method 2: Provenance data is fully included in the MPEG-G file

Provenance data is fully included in the MPEG-G file, as an XML document consisting of a sequence of PROV-XML elements. The semantics is provided in ISO 23494, the syntax is provided by W3C Prov-XML, and the container by MPEG-G Part 3 (and possibly Part 1).

**Use cases:**

* Simplifies transfer of data
* No access to resources outside of the file is required
* No guarantee that the provenance is up-to-date

**Limitations:**

* All bio data in one MPEG-G is linked to one PROV context
* It relies on W3C encoding, XML specifically

## Method 3: Provenance data is provided in a file external to the MPEG-G file

Provenance data is stored in a file external to the MPEG-G file, as an XML document consisting of a sequence of PROV-XML elements. The semantics is provided in ISO 23494, the syntax is provided by W3C Prov-XML, and the container by MPEG-G Part 3 (and possibly Part 1).

**Use cases:**

* Allows to have one central point, which is local to the computation environment, containing the provenance data
* No guarantee that the provenance is up-to-date

**Limitations:**

* All bio data in one MPEG-G is linked to one PROV context
* It relies on W3C encoding, XML specifically