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
MPEG2017/M41730
October 2017, Macau, China

Source: **DMAG-UPC et al.**
Status: **Proposal**
Title: **MPEG-G metadata: Issues and mapping with EGA profiles**
Authors: **Jaime Delgado, Daniel Naro, Silvia Llorente (Distributed Multimedia Applications Group - Universitat Politècnica de Catalunya),**
Josep Lluís Gelpí, Romina Royo (Barcelona Supercomputing Center),
Audald Lloret (Center for Genomic Regulation (CRG), Barcelona, EGA team member), Jordi Rambla (Center for Genomic Regulation (CRG), Barcelona, EGA Project Head at CRG)

Table of contents

1 Issues in metadata in MPEG-G

1.1 Introduction

Document [1] is an output document, from the last meeting in July 2017, which presents the current status of the approach to metadata in the Genomic Information Representation standard MPEG-G. Metadata will be specified in part 3 of ISO/IEC 23092.

The standard will specify a set of metadata elements associated to different levels of information (Datasets group and Dataset). These elements will compose a core set of elements, being mandatory only a part of them. In addition, the standard will specify how to include new elements that might be required in specific cases. Therefore, it would be possible to convert many different genomic information metadata sets into MPEG-G metadata.

However, since there are currently well-known and widely used sets of metadata elements, the standard will also provide "profiles"; i.e. combinations of metadata elements from the core and new ones that could be mapped to those well-known datasets.

EGA [2] is a good example of available widely used metadata sets. Therefore, it is important to verify that the MPEG-G profile intended for interoperability with some EGA metadata sets works properly. But there are others, like NCBI [3], to mention one example.

In order to validate the approach, several experts in the area are being contacted. This is an input document that reflects the discussions with part of the EGA Team.

As officially described, the European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA contains exclusive data collected from individuals whose consent agreements authorise data release only for specific research use or to bona fide researchers. The EGA provides the necessary security required to control access, and maintain patient confidentiality, while providing access to those researchers and clinicians authorised to view the data.

In order to have a discoverable set of data, EGA mandates that each submission of genomic information is properly documented in a collection of metadata documents. Each genomic file being uploaded is described with information such as a sample description, the library, or the center which has generated the file. Additionally, each submission is classified first in datasets (as collection of files), and then in studies (as collections of datasets). This also allows adding one file/dataset to multiple datasets/studies if it is relevant for every one of those. The metadata's schema can be found at [7].

The EGA Team at the CRG [4] co-manages the European Genome and phenome Archive together with the EGA Team at the European Bioinformatics Institute [5]. The EGA Team at the CRG is the one participating in this analysis.

After the joint analysis of metadata mapping, a revision of document [1] has been produced [6] and input to the next MPEG meeting in October 2017.

1.2 Terminology of metadata's structure

When addressing the mapping between EGA's (and other's) metadata specification and MPEG-G's, one needs to be aware of the changes in terminology. While in the case of EGA we find three levels (Study, Dataset, and Analysis/Run, from higher to lower level), in MPEG-G we only have two (Datasets group and Dataset). Furthermore, MPEG-G's Dataset is closer to EGA's Analysis/Run as it stores information for the set of records (either aligned or unaligned) for one patient.

MPEG-G's stream level has been left out of this comparison as it does not contain information usable without accessing the other streams and headers in the dataset it belongs to.

The mapping between the two structures can be summarized in Table 1.

EGA's metadata structure	**MPEG-G's metadata structure**
Study	Dataset group
Dataset	
Run/Analysis	Dataset
	Stream

Table 1: EGA and MPEG-G metadata structures mapping.

1.3 Differences in approach

In EGA's metadata approach, external resources are only referenced, not included. For example, in the case of an Analysis, the Sample's metadata is not included, but rather pointed to with a Reference Object.

In the case of MPEG-G, the objective is to define a file which can be used independently of external resources. Therefore, the schemas contain the description of the otherwise referenced information. However, as this might be introducing redundancies, the MPEG-G's proposed inheritance mechanism allows specifying the information only once. The side effect of this is that certain elements, such as the Sample's metadata, might be placed one level earlier, although inherited in the corresponding level. For example, if we create a file comparing the result of sequencing one individual using different platforms (i.e. one study and multiple, one per platform, datasets), by describing the sample at the study level we avoid having to repeat it for each dataset. In the meantime, through value inheritance each dataset is bounded to have the required information, thus satisfying the constraint defined by EGA.

2 Mapping

2.1 Constraints to be introduced

In order to improve the mapping between EGA's metadata and MPEG-G's EGA profile we need to add further constraints. At the moment, the identified requirements are:

- The description element of dataset group has to be mandatory.

- The TaxonID can only be 9606 (limiting to human Data). This modification does not restrict MPEG-G use cases when using other profiles.

Regarding the mandatory nature in EGA of certain fields in the dataset's metadata, even if the fields are not marked as mandatory, since their values are mandatory at the dataset group's level and the dataset level inherits them, the information is provided and no further constraints are needed.

2.2 Interoperation of EGA's repository and MPEG-G files

We here suppose that EGA's repository uses a format or metadata data base different to the MPEG-G, and that the user wants to download or upload in MPEG-G format. As such, a strategy for conversion is required from, or to, EGA's metadata. For example, the genomic data could be stored in BAM files, and the metadata values stored alongside in one or more XML documents. In this case, for the user to download an MPEG-G file, the BAM's content should be converted, and the metadata fields should be populated with the values stored in the XML documents.

2.2.1 Downloading an MPEG-G file from EGA

In this situation, all MPEG-G metadata fields (using the EGA profile) are set to the values contained in EGA's metadata. The differences analysed in clause 1.3 should not be an impediment: if a run/analysis used a certain sample, the run/analysis' metadata and the samples metadata are combined and mapped to MPEG-G's dataset metadata.

If only one sample or another field was used, the inheritance mechanism can be used advantageously to reduce the metadata size in the MPEG-G file.

2.2.2 Uploading an MPEG-G file to EGA

In case where the uploaded information does not yet have any counterpart in the EGA's database, MPEG-G metadata can be unpacked in the different metadata files used by EGA, and the upload process is thus the same as currently.

In case where part of the information was already present, for example uploading a new run/analysis to an existing study, the duplication of information should be avoided. One solution might be to include in the MPEG-G's EGA profile a field for the reference. If during the upload process the reference is already set, the relevant policy can be applied to the metadata. If the reference is not set, the record is created and the reference is send to the uploader to be used in future uploads.

3 Conclusions

The metadata approach proposed for MPEG-G is flexible enough to interoperate with other several metadata elements sets, such as the ones from EGA.

The presented specific analysis, done in collaboration with EGA experts, validates the mapping, taking into account the described differences. Nevertheless, the collaboration between the MPEG-G and EGA teams will continue in order to assure the maximum interoperability in the future.

Input document [6] reflects the considerations presented here.

4 References

[1] M17079, Genomic Information Representation Metadata, July 2017.

[2] European Genome-Phenome Archive, http://ega-archive.org/

[3] NCBI metadata's schema, https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=xml_schemas

[4] CRG http://www.crg.eu/

[5] EBI https://www.ebi.ac.uk/

[6] M41731, Genomic Information Representation Metadata. Revision after 119th meeting, October 2017.

[7] EGA metadata's schema, ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5