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/M41731
October 2017, Macau, China

Source: **DMAG-UPC et al.**
Status: **Proposal**
Title: **Genomic Information Representation Metadata. Revision after 119th meeting**
Authors: **Jaime Delgado, Daniel Naro, Silvia Llorente (Distributed Multimedia Applications Group - Universitat Politècnica de Catalunya)**

Table of contents

1 Background

Metadata in genomic information representation covers a broad set of elements (or fields) of relevance for the understanding and processing of the information at different levels. Examples include information on how the sample was obtained, what the experiment was about or even pointers to other relevant studies.

Input contribution [1] already introduced a deep analysis of metadata to be included in a Genome Information File Format. That document presented several alternatives, examples and analysis of options. [1] is the basis of the current draft standard's metadata specification. Furthermore, an output document [2] from the 119th meeting was produced to be discussed with relevant experts.

The objective is to specify metadata elements with, for example, the same expressiveness than EGA's metadata (https://ega-archive.org), and the ability to use relevant attributes such as those listed in MIAME (http://fged.org/projects/miame/) or by the Genomic Standards Consortium (http://gensc.org/). Some of the information already available in SAM files is also relevant.

The current draft standard's part dealing with metadata has considered the mentioned, and other, metadata sets in combination with the MPEG-G requirements and current WDs. The results of discussions after 119th meeting have been incorporated into this slightly revised version. Some issues on the discussion with other experts are described in the new input document [3].

An example on how to handle different approaches is related to EGA's metadata. In this case, there are different foreign keys in use to reference metadata, but this information is not needed in the case of a file format where the relationships are implicit in the structure of the file. Additionally, we avoid having to repeat redundant information: for example, if all samples in a study undergo the same experiments, we do not repeat that information.

As already described in current WD of Part 1, the MPEG-G file format considers metadata for two levels of information: *Dataset group* and *Dataset*. Although the standard format identifies more levels, only these two are considered as candidates for incorporating metadata elements.

The metadata schema aims at:

- being modular (at creation time, it can be decided to add more or less description elements, from almost none to as complete as EGA, MIxS or NCBI attributes);
- using the file format to avoid using foreign keys;
- avoiding repetitions: for those elements which are of relevance at the dataset level, the dataset group elements' value are inherited, but they can be updated if needed.

2 Introduction

Different usages of the genomic data require different sets of metadata elements. The specified structure is adaptable to different requirements.

The metadata structure and the set of elements is specified using XML.

The standard defines a minimum core set of metadata elements, which can then be extended by users and applications by including extra information elements. Sets are specified for a Dataset Group and for a Dataset.

Extensions to (new elements for) the metadata set specified in this standard are represented with an information type identifier, a value and a pointer to a resource documenting the semantics of the given information type.

Profiles are specific metadata sets specified using mechanisms provided in the standard. A profile corresponds to a well-known metadata set specified or used out of this standard, such as those from EBI or NCBI. This allows easy interoperability with already existing systems.

A profile includes a subset of core elements described in this standard, and a set of new elements specified with the extensions mechanism (see Clause 5).

The rest of clauses of this standard specify Dataset group metadata (clause 3), Dataset metadada (clause 4), Extensions (clause 5) and Profiles (clause 6).

3 Dataset group metadata

Dataset group metadata is associated to a genomic study. Table 1 presents the core set information in a dataset group metadata box. Those elements that are necessary to identify and process the dataset group are marked as mandatory.

Table 1: Base dataset group's metadata core set

Element name	Element type	Mandatory
Title	String	Yes
Type	Controlled vocabulary	Yes
Abstract	String	No
Project centre name	String	No
Description	String	No
Samples	List of sample types	Yes
Extensions	List of extension types	No

This table is readily translatable into a XML schema: each row is translated into one element of the type indicated in the element type column, with a maximum occurrence of one and a minimum occurrence depending on the mandatory nature of the element. In the case where the type is controlled vocabulary, the XML schema represents the data as a

string, but all words not included in the list of controlled vocabulary are considered as ill-formed.

As previously introduced, an extensions type is the combination of three fields: the value, the identifier of the extension, and a link to a resource documenting the interpretation of the field. In the XML schema, this is translated as an element with two attributes: the identifier (of type string) and the resource (a URL); the value is represented as the element's text as UTF-8 characters text (in case of binary information, Base64 encoding is used). Additionally, a Boolean attribute of the element indicates if the extension is only relevant to the dataset group, or if the dataset also inherits it. The resource documentation might be human readable, and the extensions parsing is not required.

As Table 1 indicates, certain elements can be described with basic types, but other element types require more complex descriptions, such as the sample type. For those elements, their respective core set of fields is provided. Table 2 provides those for the *sample* type, and Table 3 for *project centre*.

Table 2: Sample's metadata core set

Field name	Field type	Mandatory
TaxonId	Integer	Yes
Title	String	No
Extensions	List of extensions	No

[Note: See section on extensions examples to learn how this information can be adapted to EBI-EGA or NCBI's BioSample usage.]

Table 2: Project centre's metadata core set

Field name	Field type	Mandatory
ProjectcentreId	Integer	Yes
Title	String	No
Extensions	List of extensions	No

4 Dataset metadata

Dataset metadata is associated to a genomic analysis. A dataset metadata overwrites those elements whose values differ from the ones indicated at the dataset group level (i.e., the new value in the dataset is a specialization of the value at the dataset group level).

Table 4 specifies the core element names for dataset metadata. No elements are mandatory since they could be inherited from the dataset group metadata.

For example, we might have datasets for patients A, B and C; therefore the dataset group's metadata includes a list of samples representing A, B and C. The datasets then provide only one sample description (respectively of A, B or C). The base set of elements in the dataset's metadata is the same as for the dataset group, but the elements are not mandatory (so there is no need to repeat them), since per default their values are considered equal to

the values indicated in the dataset group except for those cases that have the inheritance parameter set to false. [Note: Inheritance parameter to be specified.]

As in the case of the dataset group metadata, the information is represented as an XML document, the schema of which is derived from Table 4, using the previously described methodology.

Table 4: Base dataset's metadata

Element name	Element type	Description	Mandatory
Title	String		No
Type	Controlled vocabulary		No
Abstract	String		No
Project centres	List of type project centres	Contact information of centres participating in the generation of the described study's data.	No
Description	String		No
Samples	List of type sample	Identification of the samples, based on taxonomy/scientific name, common name or anonymized name and further attributes defined in a controlled library.	No
Extensions	List of extensions		No

Also as in the case of the dataset group, the extension mechanism is available to include new attributes where necessary. See section on extensions for an example in the case of dataset's metadata.

5 Extensions

A mechanism for adding new elements to the different core metadata sets (dataset group and dataset levels) is provided.

An extended element consists of:

- information type identifier;
- value;
- pointer to a resource documenting the semantics of the given information type: this resource provides information for auto-discovery of the extension.

5.1 Examples of extensions

This clause presents two examples:

- For dataset group, based on currently existing metadata sets, as those from EGA, NCBI or others.
- For dataset, using the concept of *label* from Part 1.

5.2 Example for Dataset group metadata extensions

In order to support the broad sets of attributes used by EGA or NCBI, the extensions mechanism could be used. For example, in the case of EGA's sample metadata [4], we could give as semantic reference to the extension a link to an extra schema which would define the elements presented in Table 5 (taken over the schema provided by EGA). In this case, the fact that the semantic specification is provided as an XML schema would simplify an automatic integration of the content. The value of the extension would be a string containing the XML file.

Table 5: Sample's metadata extended to EGA's specification

Field name	Field type	Mandatory
Sample Name – taxon id	Integer	No
Sample Name – scientific	String	No
Sample Name – common name	String	No
Sample Name – anonymized name	String	No
Sample Name – individual name	String	No
Description	String	No
Links	Uri	No

In the case of NCBI's BioSample metadata, the specification is split in multiple cases [5]. Each of these subtypes is a different extension, the definition of which is constructed on the same principles: one xml element per attribute, using the data types indicated by NCBI.

Although BioSample provides compatibility with the Minimum Information about any (x) Sequence (MIxS) [6], extensions dedicated to MIxS could be also specified, once again using the same strategy.

5.3 Example for Dataset metadata extensions

Part 1 of the standard introduces the concept of dataset's labels: they allow giving unique names to regions of the data. As such, this does not allow documenting what that region represents. A possible extension to the sample metadata would be a translation tool from the label name to an ontology term, using the information indicated in Table 6.

Table 6: Sample's metadata extended with labels

Field name	Field type	Mandatory
Label name	Integer	Yes
Ontology term	URN	Yes

6 Profiles

Profiles are specific metadata sets. They are specified using the mechanisms provided in clause 6.1 [Note: Mechanisms to be specified]. Clauses 6.2 to 6.x provide formalised profiles. [Note: Currently, clause 6.1 provides an example of a possible profile.]

A profile corresponds to a well-known metadata set specified or used out of this standard, such as those from EGA or NCBI. This allows easy interoperability with already existing systems.

A profile includes a subset of the core elements described in this standard, and a set of new elements specified with the extensions mechanism (see Clause 5).

6.1 Example of profile: EGA's metadata schema

The MPEG-G dataset metadata shares characteristics with both EGA's run and analysis. We here introduce a MPEG-G profile ("EGArun") that maps the EGA's run schema (ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.run.xsd).

Table 7 presents the set of elements included in this profile. Some of them are already part of the core metadata set, and the rest are the extensions needed to match the EGA's run schema.

Table 7: Metadata elements of the *EGArun* profile

Element name		Element type	Extension description
Title		String	
Type		Controlled vocabulary	
Abstract		String	
Project centres		List of type project centres	
Description		String	
Samples		List of type sample	
Extensions		List of extensions	
	Spot	SRA.common.xsd/ SpotDescriptorType	*The SPOT_DESCRIPTOR specifies how to decode the individual reads of interest from the monolithic spot sequence. The spot descriptor contains aspects of the experimental design, platform, and processing information. There will be two methods of specification: one will be an index into a table of typical decodings, the other being an exact specification.*
	Platform	SRA.common.xsd/ PlatformType	*The PLATFORM record selects which sequencing platform and platform-specific runtime parameters. This will be determined by the Center.*
	Processing	SRA.common.xsd/ ProcessingType	

	Related links	SRA.run.xsd/ RUN_LINKS	*Links to resources related to this RUN or RUN set (publication, datasets, online databases).*
	Attributes	SRA.run.xsd/ RUN_ATTRIBUTES	*Properties and attributes of a RUN. These can be entered as free-form tag-value pairs. For certain studies, submitters may be asked to follow a community established ontology when describing the work.*

The descriptions of the extensions are taken from ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.run.xsd and ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.common.xsd.

This profile extends the MPEG-G dataset core metadata to match EGA's run schema. The file element (that points towards the file from within the metadata) from EGA's run schema is not needed in MPEG-G because the metadata is placed within the element it refers to. Further constraints need to be applied to ensure the compatibility between EGA's metadata and this profile: namely the description element in the dataset group's metadata schema has to be mandatory, and in this profile the TaxonID field can only be equal to 9096 (human).

6.1.1 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.

6.1.1.1 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.

6.1.1.2 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.

7 References

[1] Jaime Delgado, Silvia Llorente et al., ISO/IEC JTC 1 SC 29/WG 11 M39175, GENIFF (GENomic Information File Format), a proposal for a Secure Genomic Information Transport Layer (GITL) based on the ISO Base Media File Format. Chengdu, China, October 2016.

[2] ISO/IEC JTC1 SC29/WG11 N17079, Genomic Information Representation Metadata. Torino, Italy, July 2017.

[3] Jaime Delgado, Daniel Naro et al., ISO/IEC JTC1 SC29/WG11M41730, MPEG-G metadata: Issues and mapping with EGA profiles. Macau, China, October 2017.

[4] EGA Sample's metadata schema. ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/SRA.sample.xsd

[5] NCBI's BioSample types and attributes. https://submit.ncbi.nlm.nih.gov/biosample/template/?action=definition

[6] MIxS http://gensc.org/mixs/