Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-621

Metadata to Describe Genomic Information

Jaime DELGADO[1], Daniel NARO[1,3], Silvia LLORENTE[1],
Josep Lluís GELPÍ[2,3], Romina ROYO[2]
[1] *DMAG, DAC, Universitat Politècnica de Catalunya (UPC)*
[2] *Barcelona Supercomputing Center (BSC)*
[3] *Dept. Biochemistry and Molecular Biomedicine. Univ. of Barcelona (UB)*

Abstract. Interoperable metadata is key for the management of genomic information. We propose a flexible approach that we contribute to the standardization by ISO/IEC of a new format for efficient and secure compressed storage and transmission of genomic information.

Keywords. Metadata, Representation, Genomic Information, Interoperability

1. Introduction

Metadata in genomic information representation covers a broad set of elements of relevance for the understanding and processing of the information at different levels of granularity (i.e. study, individual, read, etc.). For example, one can include information on the origin of the biological sample, the preparation procedure, sequencing details or even pointers to related studies.

Different sources have proposed metadata schemas. Research centers and repositories like EGA (European Genome-Phenome Archive) [1], Genomic Standards Consortium [2] or NCBI (National Center for Biotechnology Information) [3] define metadata sets which can be referenced (or included) in genomic information repositories. They define XML (eXtensible Markup Language) schemas to facilitate interoperability. On the other hand, companies and researchers dealing with genomic information define their own metadata and include it alongside existing file formats, making interoperability complex, as other researchers may use a different representation of the same information.

In genomic studies on specific medical conditions, it is possible to have common and specific metadata applying to different elements of the study. Some values might be shared by all individuals (e.g. genomic sequencing center) while other values may be different for specific individuals inside the study (e.g. age). Therefore, being able to describe groups of individuals could be desirable for shared metadata information.

This paper addresses the need of defining metadata applied to genomic information in a way that it can be interoperable, extensible and hierarchically defined. The aim of the mechanisms described is to include or reference metadata in genomic information representation formats.

2. Methods: Categorizing metadata

The metadata associated to the reads inside a SAM (Sequence Alignment Map) [4] file is different to the metadata associated to a whole study which involves several individuals. In this case, the concept of metadata inheritance could be relevant to avoid repetition of the same metadata, relevant to the whole study and/or to each individual.

To support metadata applied to genomic information at different levels (study and dataset), we first defined a hierarchical file format called GENIFF (GENomic Information File Format) [5] and proposed it to the MPEG (Moving Picture Experts Group) [6] standardization committee in the context of the currently under definition MPEG-G [7] standard.

GENIFF's underlying idea is to structure genomic information in different layers (study, dataset, genomic data encoding element) in order to be able to apply the specific metadata to the corresponding layer. In this way, it solves several issues related to metadata association to genomic data. First of all, it proposes a well-defined way to include metadata within the genomic data file. Each level includes a metadata information structure, so it is not needed to invent a mechanism to do so. Secondly, it is possible to inherit data from the above level(s). Common metadata to both datasets and study could be defined once at the study level, and the same value would be used in the datasets. When the dataset's value differs from the study's one, the inherited value can be overwritten by providing explicitly a value for the field.

To define the metadata fields for each level, we analyzed initiatives in metadata for genomic information, one of which is summarized in the next subsection.

2.1. European Genome-phenome Archive (EGA)

The European Genome-phenome Archive (EGA) [1] is designed to be a repository for a wide range of sequence and genotype experiments generated by biomedical research projects. EGA provides several metadata schemas [8], describing different aspects related to the study, the datasets, and more. Within the EGA repository, the metadata is stored in a referential manner: as in a relational data base, the metadata for run and analysis, for example, refer to other metadata files such as the sample description.

This solution has clear advantages such as reducing the necessary size, avoiding conflicts in the information, and allowing to keep metadata, usually required to identify and locate specific studies, separated from the data that in the case of EGA is not accessible unless agreement. However, the solution is incompatible with an offline solution where data and metadata should be contained in only one file.

2.2. The MPEG (Moving Pictures Experts Group) work on Genomics

MPEG [6] is a working group of ISO/IEC (International Organization for Standardization / International Electrotechnical Commission) identified as ISO/IEC JTC 1/SC 29/WG 11. Since 1988, the group has produced standards for coded representation of digital audio and video and related data.

Following their successful previous experience in audiovisual content compression, a new initiative inside MPEG to provide compression mechanisms for genomic information started in 2014. After a detailed process of obtaining requirements, a call for proposals was launched in July 2016 [9] to provide solutions to the genomic information compression and representation problem. Based on the responses received, Working

Drafts and Committee Drafts have been developed until now. This new standard is known as MPEG-G and, once finished, it will have the official number IS 23092.

MPEG-G is divided in 5 parts: transmission and representation, compression, metadata and APIs, reference software and conformance. The first 3 parts are already in their ballot process as Committee Drafts [10] [11] [12].

3. Results: Metadata in Genomics contributed in MPEG

Partially based on GENIFF, in MPEG-G's file format structure there are 3 main layers: Dataset group, Dataset and Streams/Access units.

The dataset level contains one set of genomic information, freely combining aligned or unaligned records. In other words, the content of a SAM/BAM [4] or FASTQ [13] file is meant to populate one dataset.

Dataset groups are intended to group multiple datasets into one container. For example, a user may want to cluster all datasets sharing some common characteristics such as a common phenotype.

The last hierarchy level is the stream or access unit, depending on the dataset's encoding representation strategy. In the first strategy, each stream corresponds to one of the data streams conforming the dataset (e.g. the first position of each read, or the type of each mutation …). In the second strategy, each access unit corresponds to one genomic region, being more akin to the concept of block in BAM. In both modes (streams or access units), the data is divided in encoding blocks, the difference being the ordering of the blocks. In stream mode, blocks with the same type of information are contiguous, while in access unit mode, blocks encoding the same genomic region are stored together.

This new standard for genomic information representation strives towards a new file format containing all information required to work with the content, from the actual data to the indexing information and to the privacy rules [5]. Furthermore, the file should be usable offline. This has motivated the definition of a suitable metadata strategy, where all necessary metadata can be stored within the file.

3.1. MPEG-G metadata elements

For the two top hierarchy levels in the file, study and dataset, MPEG-G Part 1 defines a metadata box to describe the content. In other words, the metadata element in the dataset group defines the common metadata fields shared between the dataset, and then each dataset is described in its corresponding metadata box.

As we expect to find redundancy between dataset and dataset group metadata, we consider that every field which is not documented at the dataset level is in fact inherited from the dataset group level. For example, if in the dataset group's metadata element the "type" field (see Table 1), indicates "Whole Genome Sequencing", we consider each dataset (see Table 2) within the dataset group to be of this type. However, the dataset can overwrite the inherited value by providing its own.

As MPEG-G could be used in a wide variety of use cases, the schema proposes to use a reduced set of mandatory fields. As such, the following dataset group's metadata fields are mandatory: title, type, and samples.

In the case of Sample and Project center, we do not rely on the basic data types but we propose a specific data type for this field (see Table 3 for Sample type's schema, where only the TaxonId element is mandatory).

The metadata is represented in XML format within the metadata boxes defined in MPEG-G. We can highlight certain benefits of using this format, such as the existence of libraries to parse this information, and the possibility to define schemas for the boxes content, which enables the possibility to assess the validity of the provided metadata.

3.2. Extensions

Certain genomic repositories, such as EGA (see Clause 2.1), require a broader set of fields than the ones provided in the core metadata sets (see Tables 1-3). In order to address these issues, we propose to use the concept of "extensions", which the Genomic Standards Consortium proposal (see Clause 1) also considers. An extension is defined with an information type identifier (akin to the field name in Tables 1-3), a value and a pointer to a resource documenting the semantics of the given information type: this resource provides information for auto-discovery of the extension.

Using this mechanism, we can extend the previous tables to include further fields. For example, the sample metadata core set could be extended to have a field containing the scientific name of the specimen. Table 4 summarizes the envisioned sample metadata set extended to meet EGA's schema.

Table 1: Base dataset group's metadata core set

Element name	Element type
Title	String
Type	Controlled vocabulary
Abstract	String
Project center name	Project center type
Description	String
Samples	List of sample types
Extensions	List of extension types

Table 2: Base dataset's metadata

Element name	Element type
Title	String
Type	Controlled vocabulary
Abstract	String
Project centers	Project center type
Description	String
Samples	List of type sample
Extensions	List of extensions

Table 3: Sample's metadata core set

Field name	Field type
TaxonId	Integer
Title	String
Extensions	List of extensions

Table 4: Sample's metadata extensions for EGA's specification

Field name	Field type
Sample Name – scientific	String
Sample Name – common name	String
Sample Name – anonymized name	String
Sample Name – individual name	String
Description	String
Links	Uri

3.3. Profiles

We have seen how the schemas for metadata elements can be adapted to different needs using the extension mechanism. However, to aid auto discoverability, improve interoperability and simplify the development of tools, we also propose the use of profiles. When a metadata profile is active, certain extensions are mandatorily present. In the line of Table 4, and for the case of EGA, it means that the required extensions are present to extend the core set to reach EGA's set.

However, our discussion with the EGA team has highlighted that extensions are not enough to guarantee compatibility. We also need to add certain constraints. For example, in the case of the sample's taxonomy, the value has to be equal to that of the human species, or the description of a dataset group box has to be mandatory. This is possible with profiles. As with the core set, MPEG-G's standard will define an XML schema for the core sets of elements and for every different profile.

4. Discussion, conclusions and future work

Certain issues still have to be addressed. For example, certain extensions in the future could document properties only relevant for the dataset group, such as the number of datasets in the original dataset group. Although the how is not yet defined, the idea is to indicate for every extension if its value is inherited or not.

Other issues cannot be addressed in the standard, because intrinsic differences in the metadata management might require ad-hoc solutions for the download and upload of MPEG-G solutions. In the case of EGA's repository for genomic information [1], when downloading one or more genomic resources in the form of an MPEG-G file, multiple metadata resources have to be combined in the different metadata elements. Similarly, when uploading an MPEG-G file, its different levels have to be separated to perform the action. This task is especially challenging when the information being uploaded is meant to be added to existing content, or even to modify certain metadata.

The increasing existence of permanent, public repositories in the Life Sciences domain [14], allows to envision the possibility of allowing the metadata to reference an external resource compatible with the proposed schemas. By doing so we lose the ability to use only one file or use it offline, but this also simplifies sharing metadata between multiple MPEG-G instances, for example if for a given study one single file would become too overwhelming in size.

ISO/IEC 23092 is progressing taking into account the results presented in this paper. Specifically, the metadata issues are being integrated in Part 3. The proposal concerns only the data representation, and no restrictions are imposed on the implementation: whether a relational or a graph database or another strategy is the best is an open question.

Acknowledgements

This work is partly supported by the Spanish Government (GenCom, TEC2015-67774-C2-1-R and TEC2015-67774-2-R). We also thank the EGA team for their valuable comments.

References

[1] EGA, https://ega-archive.org/
[2] Genomic standards consortium, http://gensc.org/
[3] NCBI, https://www.ncbi.nlm.nih.gov/
[4] Sequence Alignment / Map (SAM) Format Specification, https://samtools.github.io/hts-specs/
[5] Jaime Delgado et al., Protecting Privacy of Genomic Information, Studies in Health Technology and Informatics, 2017. Volume 235(318-322).
[6] MPEG https://mpeg.chiariglione.org/
[7] MPEG-G https://mpeg.chiariglione.org/standards/mpeg-g
[8] Read domain XML 1.5 metadata format, https://www.ebi.ac.uk/ena/submit/read-xml-format-1-5
[9] ISO/IEC JTC 1/SC 29/WG 11 - ISO/TC 276/WG 5 MPEG2016/N16320, Joint Call for Proposals for Genomic Information Compression and Storage, June 2016.
[10] MPEG, ISO/IEC CD 23092-1, Transport and Storage of Genomic Information, November 2017.
[11] MPEG, ISO/IEC CD 23092-2, Coding of Genomic Information, November 2017.
[12] MPEG, ISO/IEC CD 23092-3, Genomic Information Metadata and APIs, February 2018.
[13] Peter Cock et al., The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleir Acids Research, 38(6), April 2010.
[14] https://www.elixir-europe.org/platforms/data/core-data-resources