Contribution ID: **94** Type: **TALK**

Streamlined Submission of Human Omics Data via the GHGA Metadata Model

Monday 4 November 2024 12:25 (20 minutes)

The German Human Genome Phenome Archive (GHGA) is a national infrastructure that promotes the secure storage, exchange, and management of access-controlled human omics data. To facilitate user-friendly and comprehensive data submissions, we developed the GHGA metadata model. The standardized model aims at maximizing the amount of collected metadata on the submitter side, enabling reusable submissions of different types of -omics data into GHGA. This metadata model is embedded in a robust ethico-legal framework addressing sensitive data and can satisfy the heterogeneous needs of submitters while maintaining FAIR principles, interoperability with European Genome Archive (EGA) and facilitating streamlined user journeys.

The GHGA metadata schema models a bottom-up experimental approach from sample collection via omics experiment procedure to bioinformatic analysis. The model allows capturing information about submitter-specified datasets, access restrictions, and inherent studies. To appropriately model this approach, the schema consists of classes that comprise both research and administrative metadata. The research metadata resembles the skeleton of the metadata model, based on the central EGA (cEGA) model, which are: Experiment, Analysis, Sample and Individual. Other classes such as Experiment Method and Analysis Method capture specifications that are tailored to perform different experiment or analysis types based on the submission type. Furthermore, the metadata model controls submitted Files through three different classes, namely Research Data File, Process Data File, and Supporting File. These file types differ with regard to the information they contain. A Research Data File holds the raw data that is the basis for further processing and analyses, which will result in a Process Data File. A Supporting File can be submitted for Experiment Method, Analysis Method and the Individual and may contain additional (un-) structured details, such as protocols or phenopackets. The administrative metadata captures information related to governance, access controls, and data use policies (Data Access Committee, Data Access Policy, Study, Publication).

The GHGA metadata schema is equipped to enable metadata exchange between GHGA and central EGA, as well as between GHGA and NFDI4Health, a partner consortium in the national research data infrastructure project, and the model project genome sequencing (MV GenomSeq, §64e SGB V). Information about the type of data collected, the methodology used, the purpose, and the governing entities are required for GHGA functionality. Details regarding downstream analysis are only required when submitting processed files. Classes in the metadata schema are further explained using properties, such as 'sex'or 'phenotypic feature'for Individual or 'instrument model'and 'library type'for Experiment Method. These properties can either be restricted, recommended, or optional, highlighting their importance in FAIRifying omics metadata. Further, we make use of community-accepted ontologies to control the content of submitted properties, promoting the significance of standardizing metadata collections and limiting the number of free-text fields in our model to an absolute minimum.

To summarize, we have developed the GHGA metadata model to be an openly accessible resource with an easy-to-use and streamlined data submission process. The model is designed to be FAIR, flexible, and interoperable to address diverse community needs, while maintaining data subject anonymity.

Please specify "other"

In addition, please add 3 to 5 keywords.

FAIR data, metadata model, metadata quality, human omics metadata, biomedicine

Please specify "other"

For whom will your contribution be of most interest?

Researchers

Please assign yourself (presenting author) to one of the following groups.

Data professionals and stewards

Primary authors: Dr IYAPPAN, Anandhi (EMBL Heidelberg); Ms MAUER, Karoline (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany)

Co-authors: Mr MENGES, Paul (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Ms SÜRÜN, Bilge (Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany); Ms TREMPER, Galina (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Dr KIRLI, Koray (German Center for Cancer Research (DKFZ), Heidelberg, Germany); Dr ULAS, Thomas (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Germany, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany); Dr NAHNSEN, Sven (Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany); Prof. SCHULTZE, Joachim L. (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Bonn, Germany, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany, Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany); Prof. BORK, Peer (European Molecular Biology Laboratory (EMBL), Heidelberg, Germany); CONSORTIUM, GHGA (German Human Genome-Phenome Archive (GHGA, W620), Deutsches Krebsforschungszentrum, Heidelberg, Baden-Württemberg, Germany)

Presenters: Dr IYAPPAN, Anandhi (EMBL Heidelberg); Ms MAUER, Karoline (Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V, German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany)

Session Classification: Session B1

Track Classification: Connecting research data: 6. Interoperable semantics at domain and application level