

# Metadata Schema for the German Human Genome-Phenome Archive (GHGA): Update for v2.0

## Authors:

Anandhi Iyappan<sup>1</sup> (<https://orcid.org/0000-0002-5571-4962>),  
 Karoline Mauer<sup>2,3</sup> (<https://orcid.org/0000-0002-9454-7941>),  
 Paul Menges<sup>4,10</sup> (<https://orcid.org/0009-0001-5687-4298>),  
 Bilge Sürün,<sup>5</sup> (<https://orcid.org/0009-0004-7799-8494>),  
 Galina Tremper<sup>4,6,7</sup> (<https://orcid.org/0009-0009-3607-9279>),  
 Simon Parker<sup>4,11</sup> (<https://orcid.org/0000-0001-9993-533X>),  
 Koray Kırılı<sup>4</sup> (<https://orcid.org/0000-0002-2289-0652>),  
 Joachim L. Schultze<sup>2,3,8</sup> (<https://orcid.org/0000-0003-2812-9853>),  
 Peer Bork<sup>1</sup> (<https://orcid.org/0000-0002-2627-833X>),  
 Thomas Ulas<sup>2,3,8</sup> (<https://orcid.org/0000-0002-9785-4197>),  
 Sven Nahnsen<sup>5,9</sup> (<https://orcid.org/0000-0002-4375-0691>)  
 for the GHGA Consortium

<sup>1</sup> Structural and Computational Biology Unit, European Molecular Laboratory (EMBL), Heidelberg, Germany

<sup>2</sup> Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Bonn, Germany

<sup>3</sup> PRECISE Platform for Single Cell Genomics and Epigenomics, German Center for Neurodegenerative Diseases (DZNE), University of Bonn and West German Genome Center, Bonn, Germany

<sup>4</sup> German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ), Heidelberg, Germany

<sup>5</sup> Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany

<sup>6</sup> Federated Information Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany

<sup>7</sup> Complex Medical Informatics, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

<sup>8</sup> Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany

<sup>9</sup> Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen, Germany

<sup>10</sup> Core Facility Omics IT and Data Management (ODCF, W610), German Cancer Research Center (DKFZ), Heidelberg, Germany

<sup>11</sup> Juristische Fakultät, Universität Heidelberg, Heidelberg, Germany

Correspondence: For any questions, comments, suggestions, or concerns regarding the GHGA Metadata Schema, please feel free to reach out to the GHGA Metadata Workstream by sending an email to Sven Nahnsen ([sven.nahnsen@uni-tuebingen.de](mailto:sven.nahnsen@uni-tuebingen.de)), Thomas Ulas ([tula@uni-bonn.de](mailto:tula@uni-bonn.de)), Anandhi Iyappan ([anandhi.iyappan@embl.de](mailto:anandhi.iyappan@embl.de)) or Karoline Mauer ([karoline.mauer@dzne.de](mailto:karoline.mauer@dzne.de)).

**Version:** 2.0, 2. Juli 2025

# Introduction to GHGA

The German Human Genome-Phenome Archive (GHGA) serves as a national platform for the secure storage, management, and dissemination of human omics data. Central to GHGA's mission is the provision of a robust metadata schema that ensures data are Findable, Accessible, Interoperable, and Reusable (FAIR). This document presents an updated overview of the GHGA Metadata Model, reflecting recent developments and current best practices in metadata management.

*For a comprehensive explanation of the foundational GHGA metadata concepts and design rationale, please refer to the original GHGA metadata white paper on Zenodo: <https://zenodo.org/records/8341224>*

## Modeling Framework

GHGA employs the Linked Data Modeling Language (LinkML) to define its metadata schema, enabling machine-readable and semantically rich metadata structures. The framework incorporates external ontologies and controlled vocabularies—selected based on stability, community adoption, and availability in FAIRsharing.org—to maximize reusability and interoperability.

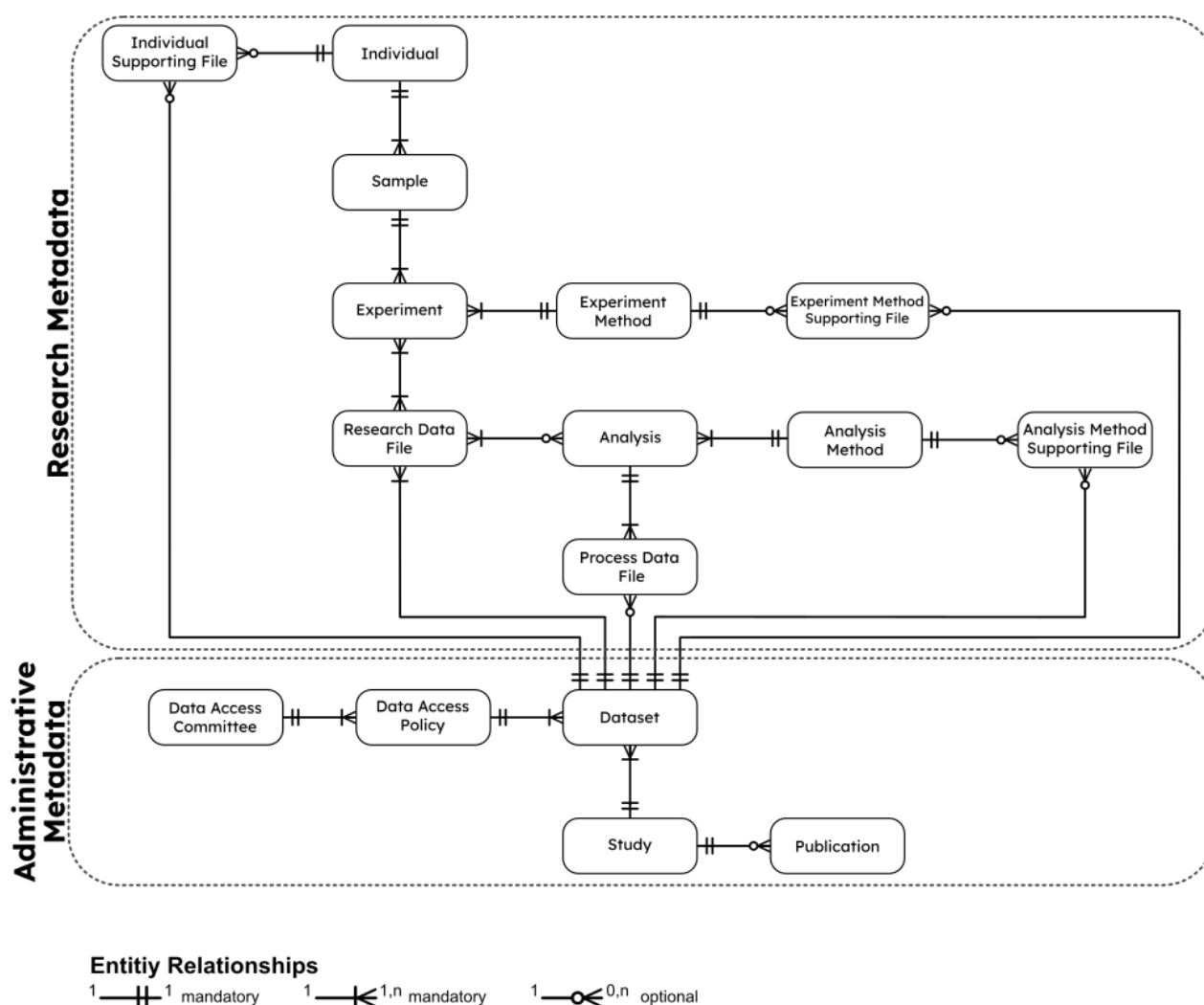
*For detailed information on the GHGA modeling framework and the use of LinkML, consult the Zenodo article linked above.*

The GHGA Metadata Model ensures non-personal data by avoiding linkage to low-level geographic information, banded age groups, and aggregated ICD-10 codes, preserving individual privacy. Data protection is achieved without compromising findability through targeted reductions in precision. By using these measures, the model maintains individual-level metadata for better discoverability while safeguarding personal privacy.

## GHGA Metadata Schema

The GHGA Metadata Model captures classes from two major information domains, research metadata and administrative metadata, as shown in Fig. 1. The model design follows a bottom-up experimental approach, facilitating a metadata registration process that is both user friendly and easy to understand.

Classes in the metadata model are either required or optional, while properties are required, recommended, or optional. Wherever possible, property entries are controlled using defined vocabulary lists or domain-specific ontologies, such as the Human Phenotype Ontology (HPO), BRENDA Tissue Ontology (BTO), or the International Classification of Diseases (ICD).



**Figure 1: Overview of the GHGA Metadata Model.** The metadata model includes 16 classes, which can be required or optional. All linkages follow a one-to-many relationship, except *Experiment* and *Research Data File*, which are connected by a many-to-many cardinality to allow submission of multiplexed files.

## Research Metadata

This domain captures the biological and experimental context of data generation. It is necessary to ensure data findability, interoperability and reusability of data submitted to GHGA. Metadata registration begins at the level of the individual subject of the study, proceeds with descriptions of the samples used in the experiment, and continues with detailed documentation of the experimental procedures. When available, users can also register metadata related to data analysis and processing.

- **Individual:** Records non-identifiable data on the person from whom the sample was derived (e.g., sex, diagnosis).
- **Sample:** Describes biospecimen characteristics such as tissue type, sample preparation, and collection method.

- **Experiment & Experiment Method:** Includes protocols and instruments used in data generation, such as sequencing platforms.
- **Analysis & Analysis Method:** Covers computational workflows, pipelines, and tools used in data processing.

## Administrative Metadata

This domain governs data provenance, access control, and publication references. It ensures that users of the GHGA data portal are informed about data accessibility, contact information and related publications.

- **Dataset:** Groups data files under a unified access policy.
- **Data Access Policy & Committee:** Defines data usage conditions and the responsible reviewing body.
- **Study & Publication:** Provides bibliographic and contextual information linked to a research project or publication.

## File Type Classification

To aid clarity and organization, GHGA classifies submitted files into:

- **Research Data Files:** Raw experimental output.
- **Process Data Files:** Derived files from computational analyses.
- **Supporting Files:** Documents like protocols, metadata sheets, or QC reports that enhance data understanding.

## Access, Submission, and Tooling

The GHGA Metadata Schema is publicly available via [GitHub](#), with downloadable schema definitions, sample YAML instances, and auto generated Excel templates for easy data entry. Metadata services ([transpiler](#) and [validator](#)) are publicly available on the GHGA Github. Please note that the validation currently does not include ontologies. The official [GHGA documentation site](#) provides an in-depth guide to using these tools.

*For best practices, file examples, and metadata inheritance logic, the Zenodo white paper remains a valuable technical companion.*

## Conclusion

GHGA's updated Metadata Model supports the scalable and standardized description of human omics datasets, advancing interoperability across national and international archives. By staying aligned with FAIR principles and leveraging community-endorsed modeling standards, GHGA fosters open science, reproducibility, and collaborative research.