Metadata Schema for the German Human Genome-Phenome Archive

Authors:
Anandhi Iyappan[1] (https://orcid.org/0000-0002-5571-4962),
Karoline Mauer[2,3] (https://orcid.org/0000-0002-9454-7941),
Paul Menges[4,10] (https://orcid.org/0009-0001-5687-4298),
Bilge Sürün,[5]
Galina Tremper[4,6,7],
Simon Parker[4,11] (https://orcid.org/0000-0001-9993-533X),
Koray Kırlı[4] (https://orcid.org/0000-0002-2289-0652),
Florian Kraus[5],
Deepak Unni[1] (https://orcid.org/0000-0002-3583-7340),
Joachim L. Schultze[2,3,8] (https://orcid.org/0000-0003-2812-9853),
Peer Bork[1] (https://orcid.org/0000-0002-2627-833X),
Thomas Ulas[2,3,8] (https://orcid.org/0000-0002-9785-4197),
Sven Nahnsen[5,9] (https://orcid.org/0000-0002-4375-0691)
for the GHGA Consortium

[1] Structural and Computational Biology Unit, European Molecular Laboratory (EMBL), Heidelberg, Germany
[2] Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Bonn, Germany
[3] German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, University of Bonn, Germany
[4] German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ), Heidelberg, Germany
[5] Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[6] Federated Information Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany
[7] Complex Data Processing in Medical Informatics, University Medical Center Mannheim, Germany
[8] Life and Medical Sciences Institute (LIMES), University of Bonn, Bonn, Germany
[9] Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen, Germany
[10] Core Facility Omics IT and Data Management (ODCF, W610), German Cancer Research Center (DKFZ), Heidelberg, Germany
[11] 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), Thomas Ulas (tula@uni-bonn.de), Anandhi Iyappan (anandhi.iyappan@embl.de) or Karoline Mauer (karoline.mauer@dzne.de).

Version: 1.0, 8. Sept. 2023

Table of Contents

List of Abbreviations

DFG	Deutsche Forschungsgemeinschaft
EJP-RD	European Joint Programme on Rare Disease
ENA	European Nucleotide Archive
FAIR	Findable, Accessible, Interoperable, Reusable
(F)EGA	(Federated) European Genome-Phenome Archive
GA4GH	The Global Alliance for Genomics and Health
GDI	European Genomic Data Infrastructure Project
GDPR	General Data Protection Regulation
GEO	Gene Expression Omnibus
GHGA	German Human Genome-Phenome Archive
MII	Medical Informatics Initiative / Medizininformatik Initiative
NFDI	Nationale Forschungsdateninfrastruktur
SPHN	Swiss Personalized Health Network
YAML	Yet Another Markup Language

List of Ontologies and Data Standards

Standard Name	Name	Link
BAO	BioAssay Ontology	https://bioportal.bioontology.org/ontologies/BAO
BTO	BRENDA Tissue Ontology	https://obofoundry.org/ontology/bto.html
CL	Cell Ontology	https://bioportal.bioontology.org/ontologies/CL
CLO	Cell Line Ontology	https://bioportal.bioontology.org/ontologies/CLO
DOID	Human Disease Ontology	https://bioportal.bioontology.org/ontologies/DOID
DUO	Data Use Ontology	https://obofoundry.org/ontology/duo.html
EDAM	Bioscientific Data Analysis Ontology	https://bioportal.bioontology.org/ontologies/EDAM
EFO	Experimental Factor Ontology	https://www.ebi.ac.uk/efo
GENEPIO	Genomic Epidemiology Ontology	https://genepio.org/ontology-details/genepio-technical-design/
GENO	Genotype Ontology	https://bioportal.bioontology.org/ontologies/GENO
GSSO	Gender, Sex, and Sexual Orientation Ontology	https://obofoundry.org/ontology/gsso.html
HANCESTRO	Human Ancestry Ontology	https://obofoundry.org/ontology/hancestro
HOOM	HPO - ORDO Ontological Module	https://bioportal.bioontology.org/ontologies/HOOM
HPO	Human Phenotype Ontology	https://obofoundry.org/ontology/hp
ICD-10	International Classification of Diseases	https://www.who.int/classifications/classification-of-diseases
ICD-O	International Classification of Diseases for Oncology	https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology
ICF	International Classification of Functioning, Disability and Health	https://bioportal.bioontology.org/ontologies/ICF

MAXO	Medical Actions Ontology	https://bioportal.bioontology.org/ontologies/MAXO
MedDRA	Medical Dictionary for Regulatory Activities	https://www.ich.org/page/meddra
MIAME	Minimum Information about a Microarray Experiment	https://pubmed.ncbi.nlm.nih.gov/11726920/
minSCe	Minimum Information about a Single-Cell Experiment	https://doi.org/10.1038/s41587-020-00744-z
MINSEQE	Minimum Information about a high-throughput Nucleotide Sequencing Experiment	https://www.fged.org/projects/minseqe/
MONDO	Mondo Disease Ontology	https://obofoundry.org/ontology/mondo
NCIt	National Cancer Institute Thesaurus	https://obofoundry.org/ontology/ncit.html
OBI	Ontology for Biomedical Investigations	https://bioportal.bioontology.org/ontologies/OBI
OMIM	Online Mendelian Inheritance in Man	https://www.omim.org/
OMIT	Ontology for MicroRNA Target	https://bioportal.bioontology.org/ontologies/OMIT
ORDO	Orphanet Rare Disease Ontology	https://www.ebi.ac.uk/ols/ontologies/ordo
PATO	The Phenotype And Trait Ontology	https://bioportal.bioontology.org/ontologies/PATO
phenopackets	Phenopackets	https://github.com/phenopackets/phenopacket-schema
SIO	Semanticscience Integrated Ontology	https://github.com/MaastrichtU-IDS/semanticscience
SNOMED CT	Systematic Nomenclature of Medicine Clinical Terms	https://www.ebi.ac.uk/ols/ontologies/snomed
SO	Sequence Types and Features Ontology	https://bioportal.bioontology.org/ontologies/SO
UBERON	UBER-anatomy Ontology	http://obophenotype.github.io/uberon/

Introduction

Introduction to GHGA

The German Human Genome-Phenome Archive (GHGA) (https://www.ghga.de) is an initiative that aims to address the needs for establishing a national data infrastructure for highly sensitive human omics and health data under a coherent ethical-legal framework. GHGA is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)[1] and is part of the German National Research Data Infrastructure (NFDI)[2].

GHGA promotes data exchange and secondary uses of research and clinical data. As part of its efforts, it also serves as a national node for the federated European Genome-Phenome Archive (FEGA)[3]. By doing this, GHGA can maintain compliance with national laws governing data privacy and protection while also coordinating with global infrastructures and initiatives for data sharing. This increases the discoverability, accessibility, and usability of datasets in GHGA for both national and international research.

Metadata and its Significance

For any research data, the availability of comprehensive metadata is crucial for research reproducibility and especially for any project involving data sharing. By quantitatively and qualitatively describing the data, the metadata aids in others' comprehension of the data's nature. To ensure that data generated for research or clinical purposes is reusable in other research contexts and to maximize the usability of data, there is a need to annotate the data concerning e.g. technologies used for collection, conditions under which it was generated or documentation of the disease status of the data subject. For human omics data, the conditions under which secondary use of the data is allowed are another important meta information for a given dataset.

Any information that characterizes data can be considered as metadata that is relevant to understand the underlying research and to leverage data to new insights. A clear overarching structure for organizing the metadata is essential to transform it into a useful resource. The schema, which serves as a blueprint for how research data should be organized, must offer a way to structure the metadata that must be gathered for various data aspects. A metadata schema also provides a point of harmonization for aligning metadata elements with other similar or adjacent data sharing initiatives, knowledge bases, and databases.

Metadata in GHGA

The goal of GHGA is to support all types of human omics data, most prominently, high throughput sequencing data. Depending on the type of study and the type of experiment, there can be different metadata properties that are relevant and need to be captured. Based on existing international standards and the existing European Genome-Phenome Archive (EGA) metadata

[1] https://www.dfg.de/download/pdf/foerderung/programme/nfdi/absichtserklaerungen_2019/2019_ghga.pdf
[2] https://www.nfdi.de/consortia-ghga/?lang=en
[3] https://ega-archive.org/federated

schema[4], the Metadata Workstream of the GHGA Consortium has developed the GHGA Metadata Schema to provide a systematic and standardized way of representing metadata by adopting and adhering to the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles[5].

GHGA Metadata Schema

The GHGA Metadata Schema is designed to support domain-specific requirements for representing details about data produced by various research communities, starting with the cancer and rare disease communities (see Fig. 1). The schema is built incrementally to ensure that (1) the schema has a basic core that is robust and (2) extensions to the schema can be added as new use cases arise. The schema captures real world objects, such as individuals or samples, as "Entities", which are further described by "Properties", such as sex, age, or phenotypic features. With the help of this schema, it is possible for the data submitters to customize the amount of data they would like to provide depending on the availability and scalable volume of metadata. To develop our schema, we adopted the EGA Metadata Schema, the established best-practice in Europe, and customized the template to meet the requirements of our communities. GHGA's commitment to be a national node in the Federated EGA network requires the schema to be easily translated to the current EGA metadata schema with minimum loss of information.

Modeling Framework

The GHGA Metadata Schema is implemented using LinkML[6], a modeling language and a framework that can be used to build a schema along with semantics. The schema exists as YAML, a human and machine readable file format that is used to define the metadata schema. Using the LinkML framework and the model definition as a YAML file, we generate technology-specific artifacts which are used throughout the GHGA software stack.

[4] https://github.com/EbiEga/ega-metadata-schema
[5] https://doi.org/10.1038/sdata.2016.18
[6] https://linkml.io

Schema Overview

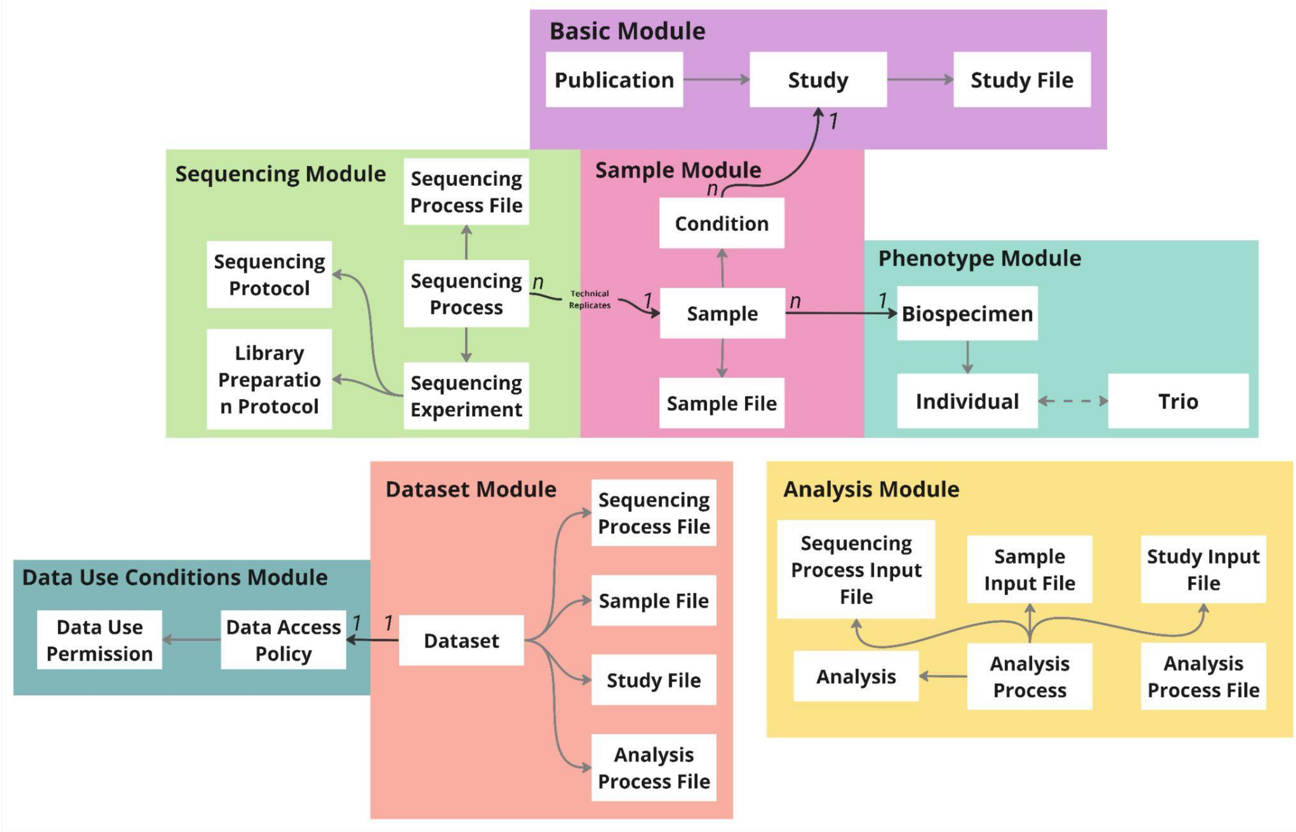


Figure 1: An overview of the GHGA Metadata Schema v1.0. Entities in the schema (white boxes) are grouped into seven modules depending on the type of information they represent (e.g. sequencing data, sample data, phenotype data). Entities and therefore the modules are connected to each other, except for *Analysis*, *Dataset* and *Data Use Conditions Module.* The modules are highlighted in different colors. Modules can be grouped together to enable different kinds of submissions, depending on the data the data submitter is able to submit to GHGA.

The structure of both, the GHGA Metadata Schema and the EGA Metadata schema[7] share the overall representation of different types of metadata and certain aspects of the linkage between metadata containers. In comparison to the EGA schema, the GHGA schema contains three additional entities (‘*Publication*’, ‘*Condition*’ and ‘*Trio*’). These extensions were considered necessary to accurately represent publications related to a study and to group samples based on experimental conditions. These entities enhance the FAIRness of the data stored in GHGA by making datasets more discoverable and traceable.The manner in which files can be appended to a submission also differs between the two data portals. While EGA’s *Run* element is used to link files, experiments and samples, research data can be added as *Study File*, *Sample File*, or *Sequencing Process File* in the GHGA Metadata Schema, as highlighted in Fig. 1.

In the GHGA Metadata Model, entities are organized into higher level structures, so-called “Modules”, depending on the type of information that is being represented. The schema consists of seven modules: *Basic*, *Sample*, *Sequencing*, *Phenotype*, *Data Use Conditions*, *Dataset*, and *Analysis*. Thus, we establish a modular metadata model that can be adapted to specific use cases.

- **Basic Module:** The *Basic Module* is the fundamental module in the GHGA Metadata Schema. It covers the minimal amount of information that must be included in a successful submission.

- **Data Use Conditions Module:** The *Data Use Conditions Module* captures in granular detail what restrictions and use conditions are associated with a Data Access Policy. This section also captures the Data Access Committee that enforces the Data Access Policy requirements.

- **Dataset Module:** The *Dataset Module* contains the *'Dataset'* entity, which is a collection of one or more Files from one or more Modules. All Files within the *Dataset Module* are subject to the Data Access Policy that is captured in the *Data Use Conditions Module*. One *Dataset Module* can only be linked to one *Data Use Conditions Module*.

- **Sample Module:** Every *Basic Module* can be linked to one *Sample Module* with one or more '*Samples'*. This module contains information relating to one or more samples sequenced in a sequencing experiment.

- **Phenotype Module:** One *Sample Module* can have one *Phenotype Module* with one or more *'Phenotypes'*. This module can be used when one or more samples originate from the same *'Biospecimen'* or *'Individual'* and thus allows to group several *'Samples'* within the *Sample Module* based on their common origin. In addition, the *Phenotype Module* captures detailed information about phenotypes or individual demographics.

- **Sequencing Module:** One *Sample Module* can also be linked to one *Sequencing Module*. The *Sequencing Module* captures information about the *'Sequencing Process'*, such as the sequencing and library preparation protocols.

- **Analysis Module:** A dataset can have one *Analysis Module* where each *Analysis Module* links to one or more files as input to the Analysis, one or more files as output to the Analysis, and the *'Analysis Process'* that captures how the analysis was performed.

As indicated in Fig. 1, the modules *Basic*, *Sample*, *Sequencing*, and *Phenotype* are linked to each other using the *'Condition'* and the *'Sample'* entities, while *Dataset* and *Data Use Conditions* are linked using the *'Dataset'* entity. The *'Analysis'* module stands independent of the others.

The modules differ in their respective requisiteness. The *Basic*, *Dataset* and *Data Use Conditions Modules* are required, while all others are optional additions. Files can therefore be attached to a dataset via multiple linkage points, one in every module. In Fig. 1, file attachment points are indicated as *'"Module" File'*.

Entities

Entities represent real world objects that capture certain aspects of metadata, and are organized in a hierarchy based on their semantics. Following is a list of entities as referenced in Fig. 1:

- **Study:** Studies are experimental investigations of a particular phenomenon. It involves a detailed examination and analysis of a subject to learn more about the phenomenon being studied.

- **Publication:** A publication refers to a journal article, book, or conference paper that presents original research or scholarly findings in a specific field of study, contributing to the collective knowledge and understanding of that discipline.
- **Condition:** A condition specifies which special characteristics and treatments apply to a sample.
- **Sample:** A sample is a limited quantity of a biospecimen to be used for testing, analysis, inspection, investigation, demonstration, or trial use. A sample is prepared from a biospecimen (isolate or tissue).
- **Biospecimen:** A biospecimen is any natural material taken from a biological entity (usually a human) for testing, diagnostics, treatment, or research purposes. The biospecimen is linked to the Individual from which the biospecimen is derived.
- **Individual:** An Individual is a person who is participating in a study.
- **Trio:** A trio is a study design that involves the genetic analysis of three individuals within a family unit. It consists of a child and their biological parents.
- **Sequencing Process:** The sequencing process captures the technical parameters that were used to produce sequencing output from a sample.
- **Sequencing Experiment:** An experiment is an investigation that consists of a coordinated set of actions and observations designed to generate data with the goal of verifying, falsifying, or establishing the validity of a hypothesis.
- **Protocol:** A plan specification which has sufficient level of detail and quantitative information to communicate it between investigation agents, so that different investigation agents will reliably be able to independently reproduce the process. There can be additional types of protocols depending on their application.
    - **Library Preparation Protocol** captures information about library preparation for an experiment.
    - **Sequencing Protocol** captures information about parameters and metadata associated with a sequencing experiment.
- **File:** A file is an object that contains information generated from a process, either an experiment or an analysis.
- **Analysis:** An analysis is a transformation that transforms input data into output data.
- **Analysis Process:** An analysis process captures the workflow steps that were performed to analyze data obtained from sequencing experiments.
- **Dataset:** A dataset is a collection of files that is prepared for distribution and is tied to a Data Access Policy.
- **Data Access Policy:** A Data Access Policy specifies under which circumstances, legal or otherwise, a user can have access to one or more datasets belonging to one or more Studies.
- **Data Access Committee:** A group of members that are delegated to grant access to one or more datasets after ensuring the minimum criteria for data sharing has been met, and request for data use does not raise ethical and/or legal concerns.

Properties

Each entity in the schema has one or more properties that are unique to itself. The table below shows all the properties that are required for entities. Please refer to GHGA Metadata Schema Documentation[7] or the GHGA Metadata Submission Spreadsheets[8] for a full list of properties that are part of the metadata schema.

Table 1: Required properties for each entity within the GHGA Metadata Schema. Each entity has a unique set of required properties, which capture minimal information ensuring findability and reusability of the dataset.

Entity	Required Properties
Study	title, description, type, affiliations
Publication	doi, study
Condition	description, name, disease or healthy, case control status, mutant or wildtype
Sample	name, description, condition
Biospecimen	individual, tissue, age at sampling
Individual	sex
Trio	mother, father, child
Sequencing Process	description, name, sequencing experiment, sample
Sequencing Experiment	description, sequencing protocol, library preparation protocol
Library Preparation Protocol	description, library name, library layout, library type, library selection, library preparation
Sequencing Protocol	description, instrument model
File	name, format, size, checksum, checksum type, dataset
Analysis	reference genome, reference chromosome

[7] https://ghga-de.github.io/ghga-metadata-schema
[8] https://github.com/ghga-de/ghga-metadata-schema/tree/main/spreadsheets

Analysis Process	analysis, study input files, sample input files, sequencing process input files
Dataset	title, description, types, data access policy
Data Access Policy	name, description, policy text, data access committee
Data Access Committee	email, institute

In addition to required properties, there are several other properties that are relevant for each entity. Rather than a binary classification, required or not, the properties for each entity have been classified along a *Requirement* axis.

On the *Requirement* axis, each property for an entity is classified as,

- **Required:** Properties marked as required are considered to be required for the functionality of the GHGA Metadata Schema (see Table 1). Without these properties, metadata cannot be submitted to GHGA successfully. In the submission spreadsheet, these properties are marked as “mandatory”.
- **Recommended:** Properties that are considered to be of importance for FAIR data sharing.
- **Optional:** Properties that are considered to be optional but providing these properties improves the discoverability and reusability.

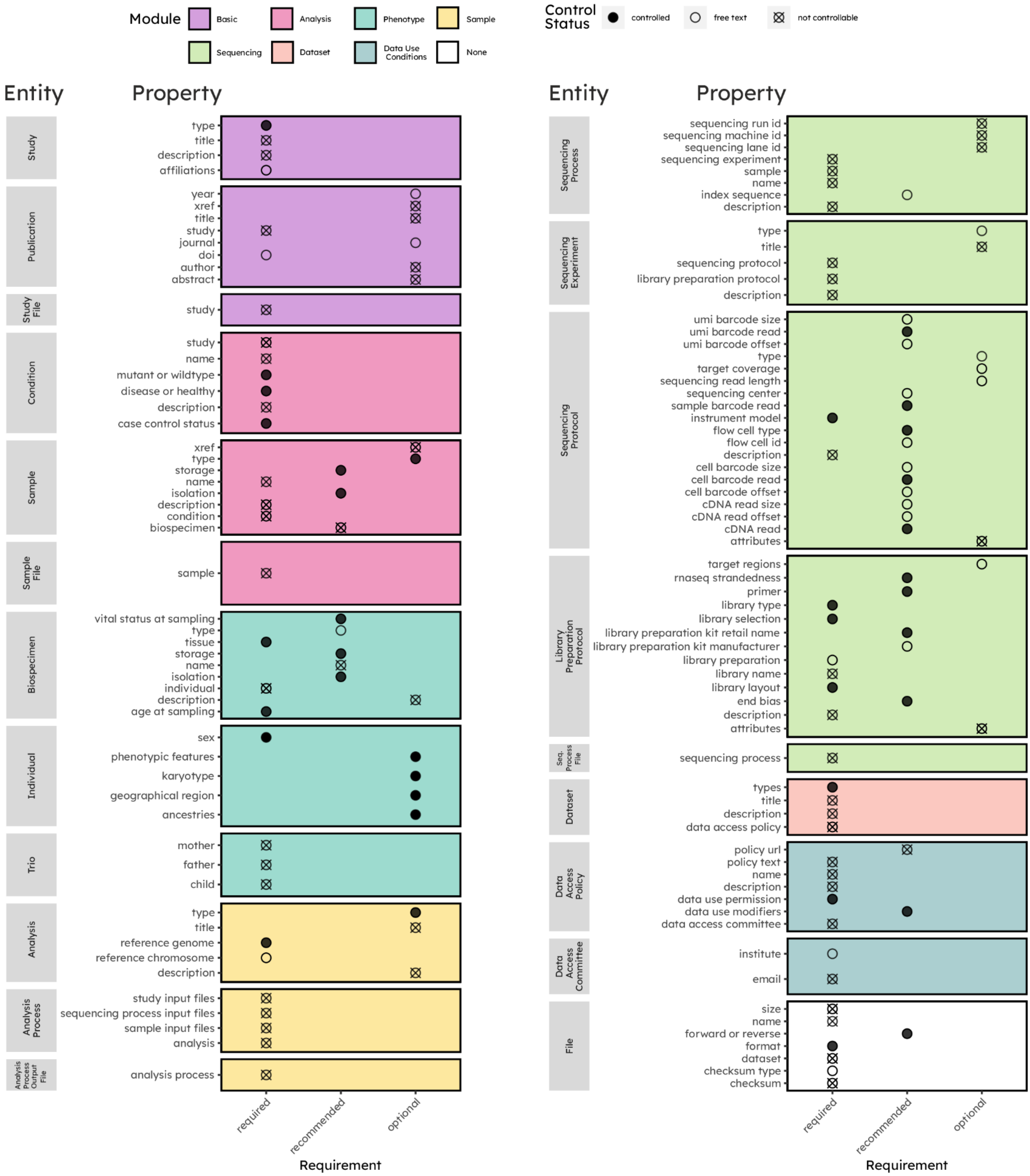
Fig. 2 depicts the classification of properties on the Requirement axis. In total, 84 out of 141 properties are considered to be **required**, 27 **recommended**, and 27 **optional**.

Properties which are considered to be required are necessary for a successful submission of data to GHGA because they establish links between entities and modules, such as the *“study”* property in the ‘*Condition’* entity, which links the ‘*Condition’* to one ‘*Study’* and therefore connects the *Basic Module* and the *Sample Module*.

Other required properties represent the data within the Metadata Schema (e.g. *“name”, “title”*, *“description”*) or are used by potential Data Requesters to identify suitable datasets based on information about the library preparation protocol (e.g. *“library layout”*, *“library selection”*, *“library preparation”*), the sequencing protocol (*“instrument model”*) or the condition under study (*“disease or healthy”, “case control status”*). Properties considered to be required capture minimal information to ensure that the submitted data is findable and reusable.

Thirty properties are classified as recommended using the *Requirement* axis. Among these properties are *“library preparation kit retail name”*, *“vital status at sampling”,* and different barcode offsets, sizes, and reads. Recommended properties are not expected to serve as the primary method by which datasets are identified but instead further describe entities such as ‘*Library Preparation Protocol’* and ‘*Sequencing Protocol’*, support the reusability of datasets and inform data users about data use permissions.

Twenty-seven properties are classified as optional, including a significant proportion of properties captured within the *Sequencing Module* and the *Phenotype Module*. Similar to the properties which are recommended, these properties do not support the user in identifying novel datasets but aid in describing datasets for reuse purposes (e.g. *"geographical region"*, *"karyotype", "target regions" or "target coverage"*).



Module
Basic
Analysis
Phenotype
Sample
Sequencing
Dataset
Data Use Conditions
None
Control Status
controlled
free text
not controllable
Entity
Property
Study
type
title
description
affiliations
Publication
year
xref
title
study
journal
doi
author
abstract
Study File
study
Condition
study
name
mutant or wildtype
disease or healthy
description
case control status
Sample
xref
type
storage
name
isolation
description
condition
biospecimen
Sample File
sample
Biospecimen
vital status at sampling
type
tissue
storage
name
isolation
individual
description
age at sampling
Individual
sex
phenotypic features
karyotype
geographical region
ancestries
Trio
mother
father
child
Analysis
type
title
reference genome
reference chromosome
description
Analysis Process
study input files
sequencing process input files
sample input files
analysis
Analysis Process Output File
analysis process
required
recommended
optional
Requirement
Sequencing Process
sequencing run id
sequencing machine id
sequencing lane id
sequencing experiment
sample
name
index sequence
description
Sequencing Experiment
type
title
sequencing protocol
library preparation protocol
description
Sequencing Protocol
umi barcode size
umi barcode read
umi barcode offset
type
target coverage
sequencing read length
sequencing center
sample barcode read
instrument model
flow cell type
flow cell id
description
cell barcode size
cell barcode read
cell barcode offset
cDNA read size
cDNA read offset
cDNA read
attributes
Library Preparation Protocol
target regions
rnaseq strandedness
primer
library type
library selection
library preparation kit retail name
library preparation kit manufacturer
library preparation
library name
library layout
end bias
description
attributes
Seq. Process File
sequencing process
Dataset
types
title
description
data access policy
Data Access Policy
policy url
policy text
name
description
data use permission
data use modifiers
data access committee
Data Access Committee
institute
email
File
size
name
forward or reverse
format
dataset
checksum type
checksum
required
recommended
optional
Requirement

Figure 2: Detailed view of the GHGA Metadata Schema with all modules, entities, and properties. Properties in the GHGA Metadata Schema are classified on the *Requirement* axis. On the requirement axis, properties can either be optional, recommended or required. The corresponding table can be found in the supplement (Supplementary table 1). Properties are grouped into entities (grey), which in turn are grouped into modules (purple: *Basic*, pink: *Sample*, turquoise: *Phenotype*, blue: *Data Use Conditions*, yellow: *Analysis*, green: *Sequencing*, red: *Dataset*). Each property can further be classified as controlled, not controllable or free text, depending on whether terms for this property are controlled with either ontology terms or controlled vocabularies, or whether submitters can enter free text (see also Table 2).

Ontologies and Controlled Vocabularies

Properties in the GHGA Metadata Schema have values that can be free text, numbers, dates, or enumerated values. Certain properties within the schema are controlled using either common ontology terms or controlled vocabularies specific to GHGA. Harmonization and standardization of vocabularies helps to ensure consistent communication among individuals and organizations. It allows for efficient data sharing, particularly in fields like healthcare and research where precise language is crucial. Additionally, it can help to minimize errors and misunderstandings. As indicated by Fig. 2 and Table 2, 71% of the required properties captured in GHGA are controlled either with ontology terms or controlled vocabularies. The same is true for 59% of the recommended and 46% of the optional properties. In total, 80 properties fall under the not controllable category as they consist of titles, names or descriptions which can only be provided as free text, or they serve as linkage between entities.

Ontologies

To ensure that the metadata that is collected in GHGA is of high quality, we support a selection of ontologies for certain properties where their values can be one or more concept terms from these ontologies. The ontologies were chosen based on their suitability as well as coverage to represent the knowledge specific to genomic medicine. To evaluate this, we reviewed the detail and context of terms captured by the respective ontology. Ontologies were only included when they can be used to represent more than one property or controlled vocabulary within a property. We also take care to integrate ontologies that are well-maintained and regularly updated. The chosen ontologies have a wide adoption and community support, which increases their interoperability and reusability.

Bioscientific data analysis ontology

Bioscientific data analysis ontology (EDAM)[9] is a straightforward ontology that consists of commonly known and widely used concepts in the field of bioinformatics, such as data types and identifiers, data formats, operations, and topics. It offers a collection of terms that come with definitions and synonyms, all organized into an easily understandable hierarchy for convenient usage. We recommend the use of concepts from EDAM to represent file formats. For example, instead of using free text ‘FASTQ’ to represent a file in FASTQ-format, we recommend using the appropriate concept **Format:1930 FASTQ**.

[9] https://bioportal.bioontology.org/ontologies/EDAM

BRENDA Tissue Ontology

The BRENDA Tissue Ontology (BTO)[10] provides a structured controlled vocabulary to describe the source of an enzyme. The ontology contains terms to represent tissues, cell lines, cell types and cell cultures. These terms span uni- and multicellular organisms. We recommend the use of concepts from BTO to represent anatomical location/site associated with a Biospecimen and/or a Sample. For example, instead of using free text 'heart tissue' to represent the site from which a Biospecimen was derived from, we would recommend using the appropriate concept **BTO:0004293 heart endothelium**.

Data Use Ontology

Endorsed by GA4GH, the Data Use Ontology (DUO)[11] allows users to tag datasets with usage restrictions, allowing them to become automatically discoverable based on a health, clinical, or biomedical researcher's authorization level or intended use. We recommend the use of concepts from DUO to represent the use restrictions associated with a Dataset. For example, instead of having use restrictions as free text in a Data Access Policy, we would recommend using the appropriate concepts from DUO to better represent the granularity of use conditions and restrictions.

Experimental Factor Ontology

The Experimental Factor Ontology (EFO)[12] provides a systematic description of many experimental variables available in databases like those from the EBI. EFO combines parts of several biological ontologies, such as UBERON anatomy, ChEBI chemical compounds, and Cell Ontology. EFO is endorsed by EMBL-EBI[13], EGA[14], and ENA[15]. We recommend the use of concepts from EFO to represent experimental factors that are typically associated with studies. For example, instead of using free text 'Exome sequencing' to signify the type of an Experiment, we would recommend using the appropriate concept **EFO:0005396 Exome sequencing**.

The Gender, Sex, and Sexual Orientation Ontology

The Gender, Sex, and Sexual Orientation Ontology (GSSO)[16] offers terms to describe gender, sex, and sexual orientation. It is aimed at interdisciplinary research in the biomedical and related sciences. We recommend the use of concepts from GSSO to represent the biological sex of an individual. For example, instead of using free text 'female' to represent the sex of an individual, we would recommend using the appropriate concept **GSSO:011317 female sex for clinical use.**

Genomic Epidemiology Ontology

The Genomic Epidemiology Ontology (GenEpiO)[17] is an open-source application ontology designed to offer a universally accessible collection of terms for databases and software interfaces.

[10] https://obofoundry.org/ontology/bto.html
[11] https://obofoundry.org/ontology/duo.html
[12] https://www.ebi.ac.uk/efo
[13] https://www.ebi.ac.uk/services
[14] https://ega-archive.org
[15] https://www.ebi.ac.uk/ena/browser
[16] https://obofoundry.org/ontology/gsso.html
[17] https://genepio.org/ontology-details/genepio-technical-design/

Leveraging OWL capabilities, GenEpiO organizes terms in a hierarchy based on Basic Formal Ontology (BFO) and Ontology for Biomedical Investigations (OBI) schemas, ensuring compatibility with other ontologies in the OBOFoundry group. Primarily focused on measurements, observables, and data in laboratory practice, genomics, epidemiology, and clinical records, GenEpiO resides under the "information content entity" category within BFO/OBI, aiming for logical coherence, interoperability, and simplicity. Geared towards outbreak investigations, food safety, and environmental pathogen surveillance, GenEpiO encompasses shared aspects of human and animal infectious disease outbreaks. We recommend the use of concepts from GenEpiO to represent the library selection method for a sequencing experiment. For example, instead of using free text “RT-PCR” to represent the usage of reverse transcription PCR, we would recommend using the appropriate concept **GENEPIO:0001959 RT-PCR method.**

Human Ancestry Ontology

The Human Ancestry Ontology (HANCESTRO)[18] provides a systematic description of the ancestry concepts. HANCESTRO was originally built for NHGRI-GWAS Catalog and has since then been used by other consortia like the GA4GH, and the Human Cell Atlas[19]. We recommend the use of concepts from HANCESTRO to represent the ancestry of an Individual. For example, instead of using ‘European ancestry’ to represent the ancestry of an Individual, we would recommend using the appropriate concept **HANCESTRO:0005 European**.

Human Phenotype Ontology

The Human Phenotype Ontology (HPO)[20] provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. HPO is used by various consortia like the GA4GH[21], Solve-RD[22], and IRDiRC[23]. We recommend the use of concepts from HPO to represent phenotypic abnormalities that characterize a Biospecimen and/or an Individual. For example, instead of using free text ‘Heart attack’ to represent an Individual who has suffered from a heart attack, we would recommend using the appropriate concept **HP:0001658 Myocardial infarction**.

International Classification of Diseases

The International Classification of Diseases (ICD)[24] is widely used across the world and is a crucial source of information on the prevalence, causes, and outcomes of human disease and mortality. Through the use of standardized coding, clinical information can be collected and recorded using ICD in primary, secondary, and tertiary care settings, as well as on death certificates. These data form the foundation for disease surveillance and statistical analysis, which inform healthcare planning, payment systems, quality control, and research. In addition, ICD's diagnostic categories facilitate consistent data collection and enable large-scale research studies. We recommend the use of classifications from ICD to represent a diagnosis associated with an Individual. For example, instead of using free text ‘Malignant neoplasm of thymus’ to represent that an Individual suffers

[18] https://obofoundry.org/ontology/hancestro
[19] https://www.humancellatlas.org
[20] https://obofoundry.org/ontology/hp
[21] https://www.ga4gh.org
[22] https://solve-rd.eu
[23] https://irdirc.org
[24] https://www.who.int/classifications/classification-of-diseases

from thymic carcinoma, we would recommend using the appropriate concept **C37 Malignant neoplasm of thymus**.

Mondo Disease Ontology

The Mondo Disease Ontology (Mondo)[25] provides a unified disease terminology that yields precise equivalences between disease concepts across various terminologies like OMIM, Orphanet, EFO, and DOID. Mondo is used by several consortia like GA4GH, ClinGen[26], and Gabriella Miller Kids First[27]. We recommend the use of concepts from Mondo to represent diseases associated with a Biospecimen and/or an Individual. For example, instead of using free text 'Myocardial infarction' to represent an Individual who has suffered from a heart attack, we would recommend using the appropriate concept **MONDO:0005068 Myocardial infarction**.

National Cancer Institute Thesaurus

The National Cancer Institute Thesaurus (NCIt)[28] is a reference terminology covering the cancer domain, including diseases, abnormalities, anatomy, drugs, genes, and more. It provides granular and consistent terminology in certain areas like cancer diseases and combination chemotherapies. The terminology is a combination from numerous cancer research domains and enables integration of information through semantic relationships. We recommend the use of concepts from NCIt to represent the case or control status associated with a Sample. For example, instead of using free text 'True Case Status' to represent the case status of a sample, we would recommend using the appropriate concept **NCIT:C99269 True Case Status**.

Orphanet Rare Disease Ontology

Orphanet and the EBI have collaborated to develop the Orphanet Rare Disease Ontology (ORDO)[29], which provides a well-organized and structured vocabulary for rare diseases. This ontology captures the relationships between diseases, genes, and other relevant features, and it serves as a valuable resource for the computational analysis of rare diseases. The Orphanet database, which is a multilingual database dedicated to rare diseases that is populated from literature and validated by international experts, serves as the basis for the ORDO. The ORDO incorporates a nosology, which is a classification system for rare diseases, as well as relationships such as gene-disease connections and epidemiological data. Additionally, the ORDO is connected to other terminologies like Medical Subject Headings (MeSH)[30], Unified Medical Language System (UMLS)[31], and Medical Dictionary for Regulatory Activities (MedDRA)[32], databases like OMIM[33], UniProtKB[34], HGNC[35],

[25] https://obofoundry.org/ontology/mondo
[26] https://clinicalgenome.org
[27] https://kidsfirstdrc.org
[28] https://obofoundry.org/ontology/ncit.html
[29] https://www.ebi.ac.uk/ols/ontologies/ordo
[30] https://www.nlm.nih.gov/mesh/meshhome.html
[31] https://www.nlm.nih.gov/research/umls/index.html
[32] https://www.ich.org/page/meddra
[33] https://www.omim.org/
[34] https://www.uniprot.org/
[35] https://www.genenames.org/

Ensembl[36], Reactome[37], The International Union of Basic and Clinical Pharmacology (IUPHAR)[38], GenAtlas[39], and classifications like ICD10. We recommend the use of concepts from ORDO to represent phenotypic features that characterize a Biospecimen and/or Individual. For example, instead of using free text 'Duchenne muscular dystrophy' to represent an Individual with Duchenne, we recommend using the appropriate concept **ORPHA:98896 Duchenne muscular dystrophy**.

Systematic Nomenclature of Medicine Clinical Terms

SNOMED Clinical Terms (SNOMED CT)[40] is a computerized repository of medical terms that are systematically organized for easy processing. This collection includes codes, terms, synonyms, and definitions that are commonly used in clinical documentation and reporting. We recommend the use of concepts from SNOMED CT to represent a sampling method associated with a Biospecimen and/or a Sample. For example, instead of using free text 'Bone marrow sampling' to represent the method used to isolate a sample, we would recommend using the appropriate concept **SNOMEDCT:234326005 Bone marrow sampling**.

Semanticscience Integrated Ontology

The Semanticscience Integrated Ontology (SIO)[41] is a biomedical ontology for knowledge discovery. It describes diverse objects, processes, and attributes (real or hypothetical) using simple design patterns. SIO has extensions for chemistry, biology, biochemistry, and bioinformatics. It underpins the Bio2RDF linked data project and aids semantic integration for SADI-based web services. To unambiguously indicate the meaning of properties in the GHGA Metadata Schema they are linked to SIO terms, e.g. **SIO:000089 dataset.**

Uber-Anatomy Ontology

Uber-Anatomy Ontology (Uberon)[42] is an integrated cross-species anatomy ontology representing a variety of entities classified according to traditional anatomical criteria such as structure, function and developmental lineage. Uberon is used in various databases like ENA, EGA, and EBI BioSamples[43]. We recommend the use of concepts from Uberon to represent anatomical location/site associated with a Biospecimen and/or a Sample. For example, instead of using free text 'heart tissue' to represent the site from which a Biospecimen was derived from, we would recommend using the appropriate concept **UBERON:0008307 heart endothelium**.

Controlled Vocabularies

Not all properties in the GHGA Metadata Schema can be represented using ontologies, either because one ontology is not ranged broadly enough or because there is no ontology covering our use case. Besides the ontologies listed previously, we defined and support a list of controlled vocabularies for several properties. These controlled vocabularies were selected based on the data

36 https://www.ensembl.org/index.html
37 https://reactome.org/
38 https://www.guidetopharmacology.org/
39 https://bio.tools/genatlas
40 https://www.ebi.ac.uk/ols/ontologies/snomed
41 https://github.com/MaastrichtU-IDS/semanticscience
42 http://obophenotype.github.io/uberon/
43 https://www.ebi.ac.uk/biosamples

dictionaries from state-of-the-art genomic data portals, such as EGA, combined with feedback from external data submitters and GHGA data hubs. As the schema will continue to evolve according to the community needs, the list of controlled vocabularies will also be extended or pruned from time to time. Property values can be one or more terms from these lists of controlled vocabularies. The table below shows all the properties represented with an ontology (ONT) or manually defined controlled vocabularies (CV). Please refer to GHGA Metadata Schema Documentation[44] or the GHGA Metadata Submission Spreadsheets[45] for a full list of controlled vocabularies.

Table 2: Properties controlled with ontology terms (ONT) or controlled vocabularies (CV). This table expands on the state-of-the art ontologies and terminologies used in the GHGA metadata schema and highlights whether a property is controlled by a standardized ontology or with manually selected controlled vocabularies.

Entity	Properties controlled with Ontologies (ONT) or Controlled Vocabularies (CV)
Study	**CV:** type
Condition	**CV:** disease or healthy, mutant or wildtype **ONT:** case control status
Sample	**ONT:** isolation **CV:** type, storage
Biospecimen	**ONT:** vital status at sampling, age at sampling, tissue, isolation **CV:** storage
Individual	**ONT:** sex, phenotypic feature, geographical region, ancestry **CV:** karyotype
Sequencing Experiment	**–**
Sequencing Process	**–**
Library Preparation Protocol	**CV:** library layout, library type, library selection, library kit retail name, RNA strandedness, primer, end bias
Sequencing Protocol	**CV:** instrument model, flow cell type, UMI barcode read, cell barcode read, cDNA barcode read, sample barcode read
File	**ONT:** format **CV**: forward or reverse

44 https://ghga-de.github.io/ghga-metadata-schema
45 https://github.com/ghga-de/ghga-metadata-schema/tree/main/spreadsheets

Analysis	**CV:** type, reference genome
Dataset	**CV:** type
Data Access Policy	**ONT:** data use permission, data user modifier
Data Access Committee	-

Data Privacy

Within the GHGA project, metadata are considered to be either Personal or Non-personal based on the definitions of those terms as given by Art. 4 GDPR[46] and with respect to Recital 26[47]. The GHGA Metadata Schema has been designed to collect non-personal Metadata that can be made publicly available, with minimal risk of re-identification of the data subjects, whilst still being scientifically useful and supporting the FAIR use of the underlying Research Data. Personal Metadata is defined as any information that is used to explain Research Data, or aid in its understanding, that would be considered to be personal within the meaning given in Art. 4 GDPR. This would include any metadata beyond that required by the GHGA Metadata Schema that is provided by the Data Submitter and is only accessible upon approval by the Data Controller.

Although no directly identifiable information is collected by the GHGA Metadata Schema, it is important for the Metadata Submitter to also consider how properties which are non-disclosive in isolation could be misused in combination resulting in the re-identification of a data subject. For example, the biological sex of a Data Subject and their phenotypic information are unlikely to separately reveal much about the Data Subject's identity. However, if the occurrence of a particular phenotype is heavily skewed by sex, sex and phenotype could be sufficient information to (re-)identify the Data Subject if an unusual combination is presented as metadata. As such Metadata Submitters may choose to submit markers for suppressed values as mandatory properties to reduce the risk to the Data Subjects.

Alongside the GHGA Metadata Schema, metadata submitters are provided with guiding information on aspects to consider when completing the schema. This guidance may include top-coding and banding of properties such as 'Individual: Age' or more general considerations where particular combinations of properties may lead to potential disclosure issues. Through this approach we have sought to avoid prohibiting properties which are scientifically valuable and may be safe in certain circumstances, whilst also being mindful that because the GHGA Metadata Schema describes individual-level data there remains a small risk to the data subjects.

46 https://gdpr-info.eu/art-4-gdpr/
47 https://gdpr-info.eu/recitals/no-26/

Conforming with the FAIR Data Principles

Published in Scientific Data in 2016[5], the FAIR Data Principles are a set of principles proposed by a group of scientists and organizations to support the reusability of digital assets. These guiding principles enable both humans and machines to find, access, interoperate, and reuse digital assets without having the need for human intervention.
Following are the ways in which the GHGA Metadata Schema conforms to FAIR Data Principles.

Findable

F1: (Meta)data are assigned a globally unique and persistent identifier

Metadata elements in the GHGA Metadata Schema are assigned globally unique identifiers (UUIDs). In addition to UUIDs, we will also assign GHGA Accessions to entities that are publicly referenced and shared. All the entities within GHGA, namely '*File*', '*Study*', and '*Dataset*' have been assigned a public-facing ID. There is a clear distinction in storing data and metadata in an encrypted state. The decryption keys are also stored in separate vaults. GHGA collects publicly available – as well as sensible – metadata but never data linking directly to the data subject such as names or addresses.

F2: Data are described with rich metadata

The GHGA Metadata Schema captures information about context, characteristics, and condition about various entities. Each entity comes with standard description and referenced with examples. The schema is constantly evolving and new properties are being added to increase the quality of metadata captured and thus improve discoverability and reusability.

F3: Metadata clearly and explicitly include the identifier of the data they describe

Each metadata entity describes itself and makes reference to other entities that give context. Each module of the schema contains entities that describe itself and are linked with various properties. Additionally, modules are connected to other modules. For example, the *'Study'* entity of the *Basic* module is connected to the 'Sample' entity of the *Sample* module via '*Condition*'. Similarly, the *Phenotype* module is connected with the *Sample* module via *'Biospecimen'*.

F4: (Meta)data are registered or indexed in a searchable resource

Properties that are submitted via the GHGA Metadata Schema will be available for search via the GHGA Archive to all users.

Accessible

A1: (Meta)data are retrievable by their identifier using a standardized communications protocol

Metadata elements and entities that the metadata elements refer to are accessible via HyperText Transfer Protocol (Secure) (HTTP(S)) protocol where any client capable of understanding the HTTP(S) protocol can fetch resources from GHGA. These protocols are open, free, secure and universally implementable across various clients.

Overall, the GHGA Archive will be built in such a way that the client will be notified if the resource that they request requires authentication and authorization, as is the case with sensitive objects like File and Sample.

A2: Metadata are accessible, even when the data are no longer available

The GHGA Metadata Schema is built to keep track of:

- Deprecation status (and deprecation date)
- What is the replacement (replaced by)

The schema can also be extended to track the lifecycle of metadata entities in a more granular manner as needed.

Interoperable

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

The GHGA Metadata Schema is built using the LinkML framework which provides translations of the YAML file that is used to describe the schema into various forms including Resource Description Framework (RDF) and Web Ontology Language (OWL).
These machine readable formal representations of the schema can be indexed in external terminology services and provide more context on how metadata entities in GHGA relate to concepts and entities from other terminologies, thesauri, and controlled vocabularies.

I2: (Meta)data use vocabularies that follow the FAIR principles

We map entities and properties in the GHGA Metadata Schema to other vocabularies, terminologies, and ontologies. Currently, we map our entities to SIO[48], EFO, Mondo, HPO, NCIt, EDAM, and DUO. Ontology terms are used to precisely describe properties defined in our model. These vocabularies and ontologies are evaluated using the FAIRsharing resource[49]. FAIRsharing[50] provides a searchable web-portal containing crowd-sourced standards, data policies and databases. This helps GHGA to ensure that used vocabularies and ontologies are well maintained, recommended and released in a stable version.

I3: (Meta)data include qualified references to other (meta)data

In the GHGA Metadata Schema we have meaningful links between metadata entities such that each entity is better described in the context of other entities.

Reusable

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

The GHGA Metadata Schema provides a catalog of entities and properties that facilitates representation of:

---

[48] https://github.com/MaastrichtU-IDS/semanticscience
[49] https://fairsharing.org
[50] https://doi.org/10.1038/s41587-019-0080-8

- Scope
- License
- Protocol
- Type of data
- Data use conditions

The schema also utilizes existing domain standards and reporting guidelines for expressing metadata. For example, the schema reuses attributes from MIAME, MINSEQE and minSCe[51] for representing information about high throughput sequencing experiments.
The data provenance is reflected under the Data Use Condition, which identifies the data controller of the research data. The data controller is defining the usage rights of the data. In the GHGA Metadata Schema, conditions for the usage of data are defined using DUO permissions.

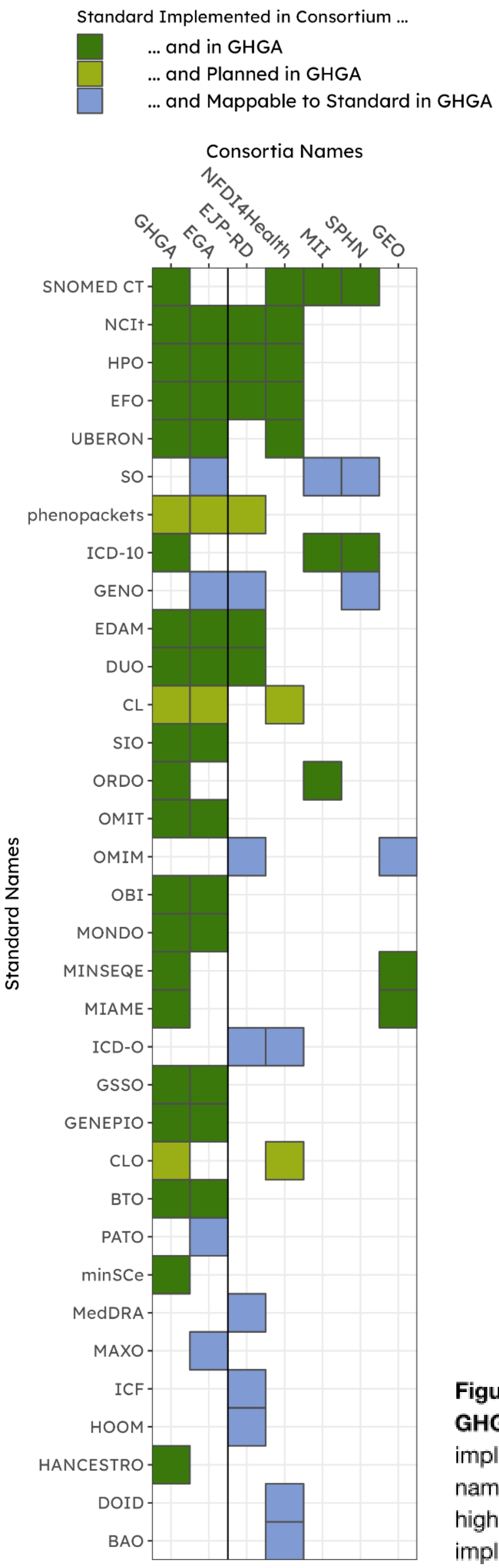
Integration of Standards and Comparison to (Inter-) National Consortia

The management of massive amounts of research data is becoming an increasingly significant aspect of non- as well as academic research. Within Germany, the NFDI consortia aim to enable long term research data management with interested communities. One of the important aspects of GHGA's design is to make the shared data reusable, interoperable and sustainable. For this purpose, we have incorporated state-of-the-art ontologies and terminology systems to make data exchange possible within the national spectrum as well as international platforms. Our model is based on the long-standing metadata model developed by the EGA. We have utilized the fundamental aspects of the EGA model and further extended it to suit our specific needs while maintaining compatibility with the current EGA Metadata Model. Subsequently, EGA and GHGA share a wide range of accepted ontologies, such as DUO, which is a standard developed by GA4GH, an international policy-framing and technical standards-setting organization. Beyond that, to facilitate easy exchange of data between GHGA and other NFDI consortia, such as NFDI4Health[52], or the Medizininformatik Initiative (Medicine Informatics Initiative, MII)[53], we have adapted common ontology and terminology services such as SNOMED CT, HPO, NCIt, EFO and ICD. The usage of well-established data standards enables the responsible, voluntary, and secure sharing of genomic and health-related data.

GHGA Shares Standard Ontologies with other Consortia

In order to validate our alignment with other consortia, we did a survey on existing research communities and data portals dedicated towards genomic medicine in Germany, Switzerland, the United States, and those operating across Europe. Fig. 3 highlights the shared standards between GHGA and other similar consortia, namely the MII and NFDI4Health in Germany, SPHN in Switzerland, GEO in the US, and EJP-RD and EGA in Europe. For this comparison, we evaluated

[51] https://doi.org/10.1038/s41587-020-00744-z
[52] https://www.nfdi4health.de/en/
[53] https://www.medizininformatik-initiative.de/en/start



standards focused on phenotype descriptions, medical terminologies and omics experiments. The outcome of this analysis reflects that GHGA is well aligned with comparable consortia involved in genome research. Quantifying all ontologies that are already implemented in GHGA (dark green boxes), planned to be implemented (light green boxes), or mappable to a standard implemented in GHGA (blue boxes), it can be observed that there is a complete overlap between GHGA and all examined consortia.

The most abundant terminology standard across all analyzed consortia are SNOMED CT, which is implemented by all consortia operating in (partially) German-speaking countries, as well as NCIt, HPO, and EFO, which are implemented by GHGA, EGA, EJP-RD, and NFDI4Health. Our analysis revealed the usage of standards and ontologies that are already in use in other consortia and still need to be included in GHGA, such as phenopackets, CLO, and CL. Additionally, we identified ontologies that are mappable to those used in GHGA: i.e. ICD-O, which is implemented by EJP-RD and NFDI4Health, can be represented using ICD-10 and SNOMED-CT, both of which are in use at GHGA. Other instances for mappable ontologies are SO, GENO, OMIM, PATO, MedDRA, MAXO, ICF, HOOM, DOID, and BAO, all of which map to either OBI, EFO, UBERON or HPO. With DUO and phenopackets, we incorporate both standards set by GA4GH which are relevant for omics medical research.

As expected, the largest intersection in terms of data standards is between GHGA and EGA. Both consortia completely share thirteen ontologies and additional two are planned to be implemented in GHGA. EGA uses four ontologies that are mappable to one or more ontologies implemented in GHGA. In comparison to GHGA, EGA does not incorporate SNOMED CT, ORDO, CLO, and HANCESTRO, as well as the minimum information standards MINSEQE, MIAME, and minSCe.

Figure 3: Overview of ontologies and standards shared between GHGA and other consortia. The x-axis displays standards implemented in GHGA and other consortia; the y-axis shows the names of the consortia. The implementation status of a standard is highlighted using different colors (dark green: standard is implemented by both the consortium and GHGA; light green: standard is implemented by the consortium and implementation into GHGA is planned; blue: standard is implemented by the consortium and mappable to one or more standards implemented by GHGA).

Conclusion and Future Outlook

The GHGA Metadata Schema[57] is built to aid in the storage and representation of non-personal data arising from omics experiments. Its structure and makeup is inspired by EGA's metadata schema, since GHGA's position as a national node in the FEGA network requires compatibility with EGA.

The GHGA schema is divided into seven modules, depending on the type of information that is being represented. Each module captures a set of unique entities, which in turn are represented by properties. Properties in GHGA can be either required, recommended or optional, depending on their significance for basic GHGA functionality. In line with the FAIR principles, property submission is limited and controlled with the usage of ontology terms or controlled vocabularies, which are highlighted in the submission spreadsheets as well as the YAML schema itself. GHGA incorporates common ontologies which are widely accepted among the omics medicine community. Additionally, ontologies and other data standards used in GHGA are implemented by the majority of other national and international consortia, such as the MII and NFDI4Health in Germany, or EGA and EJP-RD in Europe.

As research in omics medicine is evolving, so is our schema. We have built it incrementally on the basis of a robust core, allowing the schema to be changed flexibly whenever new technologies or experiment setups arise that need to be accommodated. Therefore, the schema is open to change in the future and the white paper will be versioned accordingly.

Future work concerns the extension of functionality of the GHGA schema to other omics data layers. We will also work on extending the metadata model to facilitate the integrated modeling of human multi-omics data in the near future. As upcoming phases of GHGA will enable standardized data analysis, e.g. using nf-core[54] pipelines such as sarek[55], the GHGA Metadata Schema will be extended to capture workflow metadata. This includes metadata needed for the workflow to run (e.g. filepaths, IDs)[56] and metadata produced by the workflow (e.g. software versions, execution reports, quality control report)[57]. Furthermore, the GHGA metadata schema is designed in such a way that it can accommodate the requirements of the future phases of the project such as integration of multiple omics modalities and connecting omics data to phenotypic data for the 'Atlas' phase and community specific large scale omics data for the 'Cloud' phase respectively.

54 https://nf-co.re/
55 https://nf-co.re/sarek/3.2.3/
56 https://nf-co.re/sarek/3.2.3/docs/usage/
57 https://nf-co.re/sarek/3.2.3/results/sarek/results-ed1cc8499366dcefea216fe37e36c6189537d57b/germline_test/

Acknowledgements

This work was done by the GHGA Metadata Workstream as part of the GHGA Consortium (www.ghga.de). GHGA is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Grant Number 441914366) and is part of the National Research Data Infrastructure (NFDI).

We thank the European Genome-Phenome Archive (EGA) for their continuous support and feedback during the development of the GHGA Metadata Schema and the Helmholtz Metadata Collaboration (HMC) for providing valuable input throughout the implementation process.

Supplement

Supplementary Table 1: Property classification on the *Requirement* axis.

Property	Requirement	Entity	Module	Control_Status
type	optional	Analysis	Analysis	controlled
title	optional	Analysis	Analysis	not controllable
description	optional	Analysis	Analysis	not controllable
reference chromosome	required	Analysis	Analysis	free text
reference genome	required	Analysis	Analysis	controlled
analysis	required	Analysis Process	Analysis	not controllable
study input files	required	Analysis Process	Analysis	not controllable
sample input files	required	Analysis Process	Analysis	not controllable
sequencing process input files	required	Analysis Process	Analysis	not controllable
analysis process	required	Analysis Process Output File	Analysis	not controllable
individual	required	Biospecimen	Phenotype	not controllable
isolation	recommended	Biospecimen	Phenotype	controlled
storage	recommended	Biospecimen	Phenotype	controlled
vital status at sampling	recommended	Biospecimen	Phenotype	controlled
description	optional	Biospecimen	Phenotype	not controllable
type	recommended	Biospecimen	Phenotype	free text
name	recommended	Biospecimen	Phenotype	not controllable
tissue	required	Biospecimen	Phenotype	controlled
age at sampling	required	Biospecimen	Phenotype	controlled
description	required	Condition	Sample	not controllable
name	required	Condition	Sample	not controllable
study	required	Condition	Sample	not controllable

disease or healthy	required	Condition	Sample	controlled
case control status	required	Condition	Sample	controlled
mutant or wildtype	required	Condition	Sample	controlled
email	required	Data Access Committee	Data Use Conditions	not controllable
institute	required	Data Access Committee	Data Use Conditions	free text
name	required	Data Access Policy	Data Use Conditions	not controllable
description	required	Data Access Policy	Data Use Conditions	not controllable
policy text	required	Data Access Policy	Data Use Conditions	not controllable
data access committee	required	Data Access Policy	Data Use Conditions	not controllable
data use modifiers	recommended	Data Access Policy	Data Use Conditions	controlled
policy url	recommended	Data Access Policy	Data Use Conditions	not controllable
data use permission	required	Data Access Policy	Data Use Conditions	controlled
title	required	Dataset	Dataset	not controllable
description	required	Dataset	Dataset	not controllable
data access policy	required	Dataset	Dataset	not controllable
types	required	Dataset	Dataset	controlled
name	required	File	None	not controllable
size	required	File	None	not controllable
checksum	required	File	None	not controllable
dataset	required	File	None	not controllable
forward or reverse	recommended	File	None	controlled
checksum type	required	File	None	free text
format	required	File	None	controlled

karyotype	optional	Individual	Phenotype	controlled
geographical region	optional	Individual	Phenotype	controlled
ancestries	optional	Individual	Phenotype	controlled
phenotypic features	optional	Individual	Phenotype	controlled
sex	required	Individual	Phenotype	controlled
target regions	optional	Library Preparation Protocol	Sequencing	free text
library preparation kit retail name	recommended	Library Preparation Protocol	Sequencing	controlled
description	required	Library Preparation Protocol	Sequencing	not controllable
library name	required	Library Preparation Protocol	Sequencing	not controllable
primer	recommended	Library Preparation Protocol	Sequencing	controlled
end bias	recommended	Library Preparation Protocol	Sequencing	controlled
rnaseq strandedness	recommended	Library Preparation Protocol	Sequencing	controlled
attributes	optional	Library Preparation Protocol	Sequencing	not controllable
library preparation kit manufacturer	recommended	Library Preparation Protocol	Sequencing	free text
library preparation	required	Library Preparation Protocol	Sequencing	free text
library layout	required	Library Preparation Protocol	Sequencing	controlled
library type	required	Library Preparation Protocol	Sequencing	controlled
library selection	required	Library Preparation Protocol	Sequencing	controlled
study	required	Publication	Basic	not controllable
year	optional	Publication	Basic	free text
journal	optional	Publication	Basic	free text
title	optional	Publication	Basic	not controllable
abstract	optional	Publication	Basic	not controllable
author	optional	Publication	Basic	not controllable
xref	optional	Publication	Basic	not controllable
doi	required	Publication	Basic	free text
type	optional	Sample	Sample	controlled

name	required	Sample	Sample	not controllable
description	required	Sample	Sample	not controllable
condition	required	Sample	Sample	not controllable
isolation	recommended	Sample	Sample	controlled
storage	recommended	Sample	Sample	controlled
xref	optional	Sample	Sample	not controllable
biospecimen	recommended	Sample	Sample	not controllable
sample	required	Sample File	Sample	not controllable
type	optional	Sequencing Experiment	Sequencing	free text
description	required	Sequencing Experiment	Sequencing	not controllable
sequencing protocol	required	Sequencing Experiment	Sequencing	not controllable
library preparation protocol	required	Sequencing Experiment	Sequencing	not controllable
title	optional	Sequencing Experiment	Sequencing	not controllable
description	required	Sequencing Process	Sequencing	not controllable
name	required	Sequencing Process	Sequencing	not controllable
sequencing experiment	required	Sequencing Process	Sequencing	not controllable
sample	required	Sequencing Process	Sequencing	not controllable
sequencing run id	optional	Sequencing Process	Sequencing	not controllable
sequencing lane id	optional	Sequencing Process	Sequencing	not controllable
sequencing machine id	optional	Sequencing Process	Sequencing	not controllable
index sequence	recommended	Sequencing Process	Sequencing	free text
sequencing process	required	Sequencing Process File	Sequencing	not controllable
type	optional	Sequencing Protocol	Sequencing	free text
sequencing read length	optional	Sequencing Protocol	Sequencing	free text
target coverage	optional	Sequencing Protocol	Sequencing	free text
description	required	Sequencing Protocol	Sequencing	not controllable

flow cell type	recommended	Sequencing Protocol	Sequencing	controlled
umi barcode read	recommended	Sequencing Protocol	Sequencing	controlled
cell barcode read	recommended	Sequencing Protocol	Sequencing	controlled
cDNA read	recommended	Sequencing Protocol	Sequencing	controlled
sample barcode read	recommended	Sequencing Protocol	Sequencing	controlled
sequencing center	recommended	Sequencing Protocol	Sequencing	free text
flow cell id	recommended	Sequencing Protocol	Sequencing	free text
umi barcode offset	recommended	Sequencing Protocol	Sequencing	free text
umi barcode size	recommended	Sequencing Protocol	Sequencing	free text
cell barcode offset	recommended	Sequencing Protocol	Sequencing	free text
cell barcode size	recommended	Sequencing Protocol	Sequencing	free text
cDNA read offset	recommended	Sequencing Protocol	Sequencing	free text
cDNA read size	recommended	Sequencing Protocol	Sequencing	free text
attributes	optional	Sequencing Protocol	Sequencing	not controllable
instrument model	required	Sequencing Protocol	Sequencing	controlled
title	required	Study	Basic	not controllable
description	required	Study	Basic	not controllable
affiliations	required	Study	Basic	free text
type	required	Study	Basic	controlled
study	required	Study File	Basic	not controllable
analyses	required	Submission	Submission	not controllable
analysis process output files	required	Submission	Submission	not controllable
analysis processes	required	Submission	Submission	not controllable
biospecimens	required	Submission	Submission	not controllable
conditions	required	Submission	Submission	not controllable
data access committees	required	Submission	Submission	not controllable

data access policies	required	Submission	Submission	not controllable
datasets	required	Submission	Submission	not controllable
individuals	required	Submission	Submission	not controllable
library preparation protocols	required	Submission	Submission	not controllable
publications	required	Submission	Submission	not controllable
sample files	required	Submission	Submission	not controllable
samples	required	Submission	Submission	not controllable
sequencing experiments	required	Submission	Submission	not controllable
sequencing process files	required	Submission	Submission	not controllable
sequencing processes	required	Submission	Submission	not controllable
sequencing protocols	required	Submission	Submission	not controllable
studies	required	Submission	Submission	not controllable
study files	required	Submission	Submission	not controllable
trios	required	Submission	Submission	not controllable
mother	required	Trio	Phenotype	not controllable
father	required	Trio	Phenotype	not controllable
child	required	Trio	Phenotype	not controllable