

The GHGA Metadata Model: A Framework for National Data Sharing

Karoline Mauer^{1,2,7}, Anandhi Iyappan^{3,7}, Deepak Unni^{3,7}, Florian Kraus^{4,7}, Simon Parker^{5,7}, Galina Tremper^{5,7}, Bilge Sürün^{4,7}, Paul Menges^{5,7}, Koray Kirli^{5,7}, Joachim L. Schultze^{1,2,6,7}, Thomas Ulas^{1,2,6,7}, Sven Nahnsen^{4,7}

¹ Systems Medicine, German Center for Neurodegenerative Diseases (DZNE) e.V., Venusberg-Campus 1/99, 53127 Bonn, Germany,

² German Center for Neurodegenerative Diseases (DZNE), PRECISE Platform for Genomics and Epigenomics at DZNE, and University of Bonn, Bonn, Germany, ³ European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, ⁴ Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany,

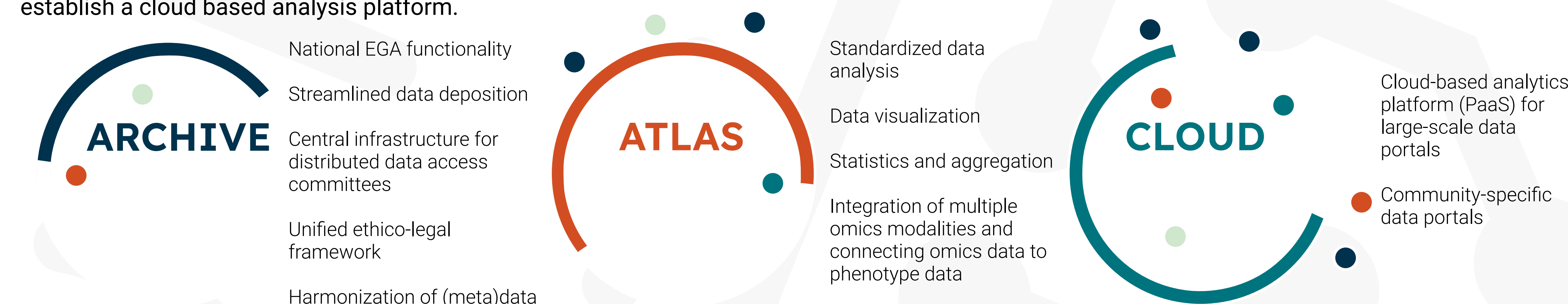
⁵ German Center for Cancer Research (DKFZ), Heidelberg, Germany, ⁶ Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany, ⁷ German Human Genome-Phenome Archive (GHGA), Heidelberg, Germany

About GHGA

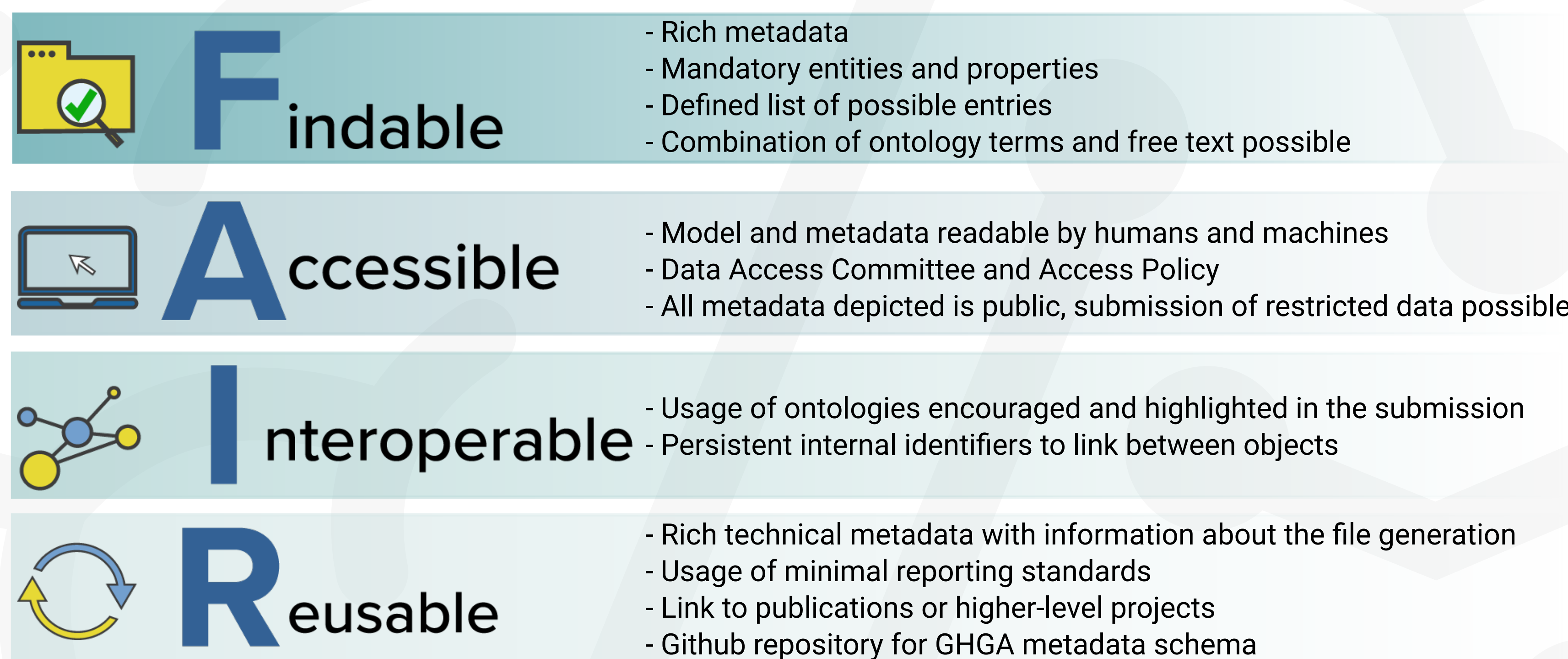
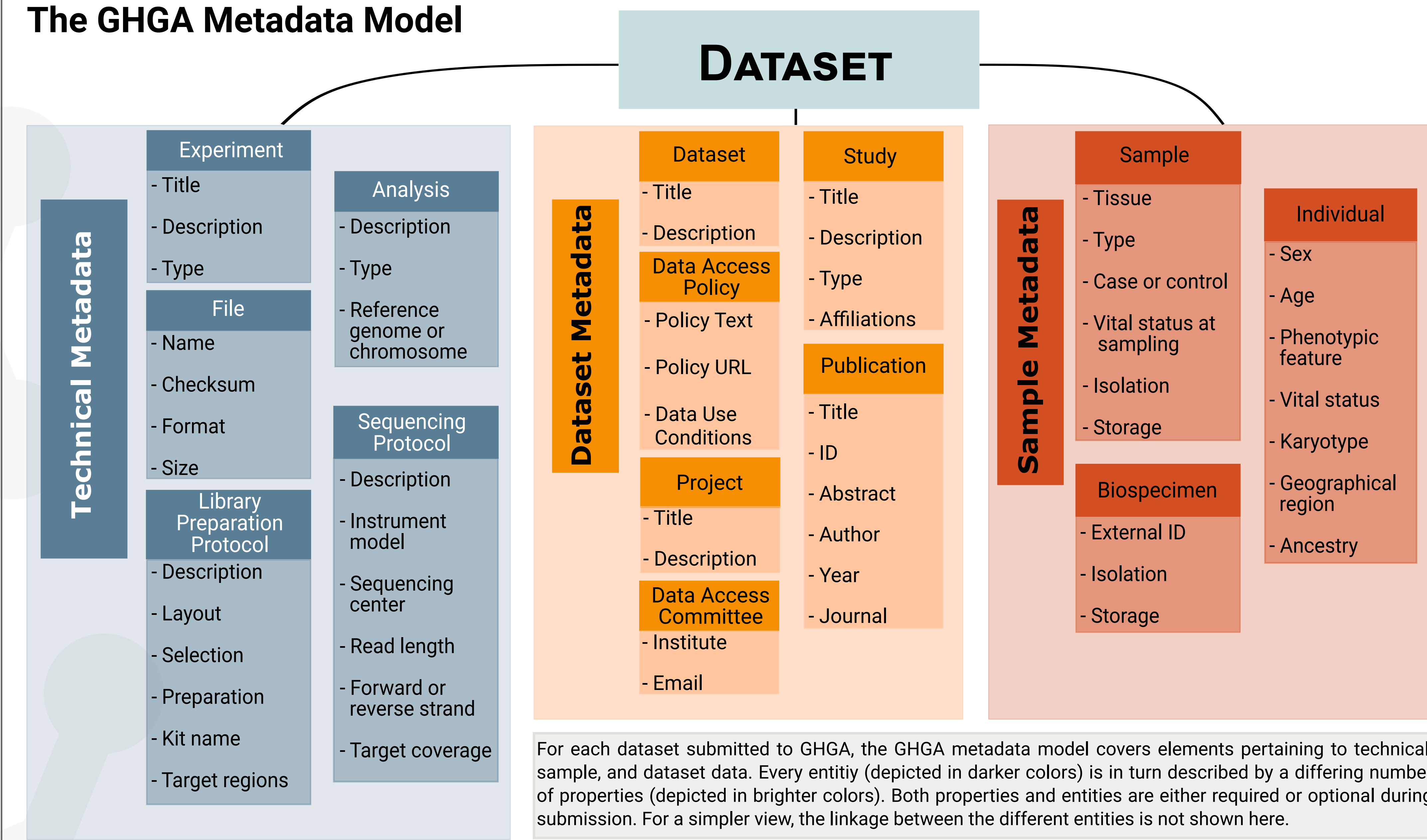
The German Human Genome-Phenome Archive (GHGA) was established as a consortium in 2020 with the aim to build a national federated infrastructure to store and share human omics data. GHGA is funded as part of the National Research Data Infrastructure (NFDI) initiative. With 46 participants from 21 institutions, GHGA is organized in six data hubs across Germany combining leading institutions in genomic medicine, major omics data producers and high-performance computing centers committed to provide scalable infrastructure.

GHGA's Mission and Interaction

GHGA will receive data from major sequencing centers and other institutions in Germany, while exchanging (meta)data with national and international initiatives such as genomDE and the European Genome-Phenome Archive (EGA). Technical harmonization of data will ensure quality and comparability of data for large-scale analyses. Data access will be regulated via Data Access Committees with a solid ethico-legal framework set up by GHGA providing a secure space for data storage and access. GHGA will build on existing infrastructures for high-performance computing to establish a cloud based analysis platform.



The GHGA Metadata Model



Community Standards in Use

- FAIR Data Principles
- Global Alliance for Genomic Health
- Research Data Alliance
- Genomics Standards Consortium
- Genomic Data Commons
- Dublin Core
- Minimum Information Standards

- Data Use Ontology
- Human Phenotype Ontology
- Experimental Factor Ontology
- SNOMED CT
- ICD-11
- NCT
- HANCESTRO
- other controlled vocabularies

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, the Metadata Workstream at the GHGA Consortium has developed a 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 principles¹.

GHGA Metadata Schema

The GHGA Metadata Schema is built to support domain specific requirements for representing information about data generated by various research communities starting with the Cancer and Rare Disease communities. 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 itself is inspired by the metadata schema from EGA. This is primarily because a lot of the data in the initial phase of GHGA will be similar to the kind of information that is submitted to EGA. Furthermore, GHGA's commitment to be a national node in the Federated EGA network requires the schema to be easily translated to EGA with minimum loss of information.

Modeling Framework

The GHGA Metadata Schema is implemented using LinkML, a modeling language and a framework that can be used to build a schema along with semantics. The schema exists as a Yet another Model Language (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.