Welcome to the technical documentation for the GA4GH Pedigree Standard!

Release 0.1

Jun 27, 2022

TABLE OF CONTENTS

Note: This project is under active development.

CHAPTER

ONE

INTRODUCTION TO THE GA4GH PEDIGREE STANDARD

1.1 What is the GA4GH Pedigree Standard

The GA4GH Pedigree Standard allows for the computable exchange of family health history as well as representation of larger, more complex families. The collection of specific clinical or genetic data is outside the scope of this deliverable, and would instead be handled by other formats and references to individuals within the pedigree representation.

1.2 How did the Pedigree Standard Come About

The need for high quality, unambiguous, computable pedigree and family information is critical for scaling genomic analysis to larger, complex families. Pedigree data is currently represented in heterogeneous formats that frequently result in the use of lowest-common-denominator formats (e.g., PED) or custom JSON formats for data transfer. The HL7 FHIR standard core data models do not support pedigrees, but there is a draft extension to support genomic pedigrees that should be evaluated and potentially extended by the GA4GH. Standardizing the way systems represent family structure will allow patients to share this information more easily between healthcare systems and help software tools to use this information to improve genome analysis and diagnosis.

We asked our stakeholders our about their use of family health history and pedigree data - How are you using it? How is it stored? What do you wish you could do with your data that you currently can't? The results of the survey can be found here. A significant percentage of respondents were using a non-computable or non-interoperable format, and there was no common tool or format with which they inteded to import or export data. Importantly, 57% of respondents were experiencing challenges with standardization, including lack of computability and integration with analysis tools, and inability to represent complex families and share data easily.

1.3 Why PED is Not Enough

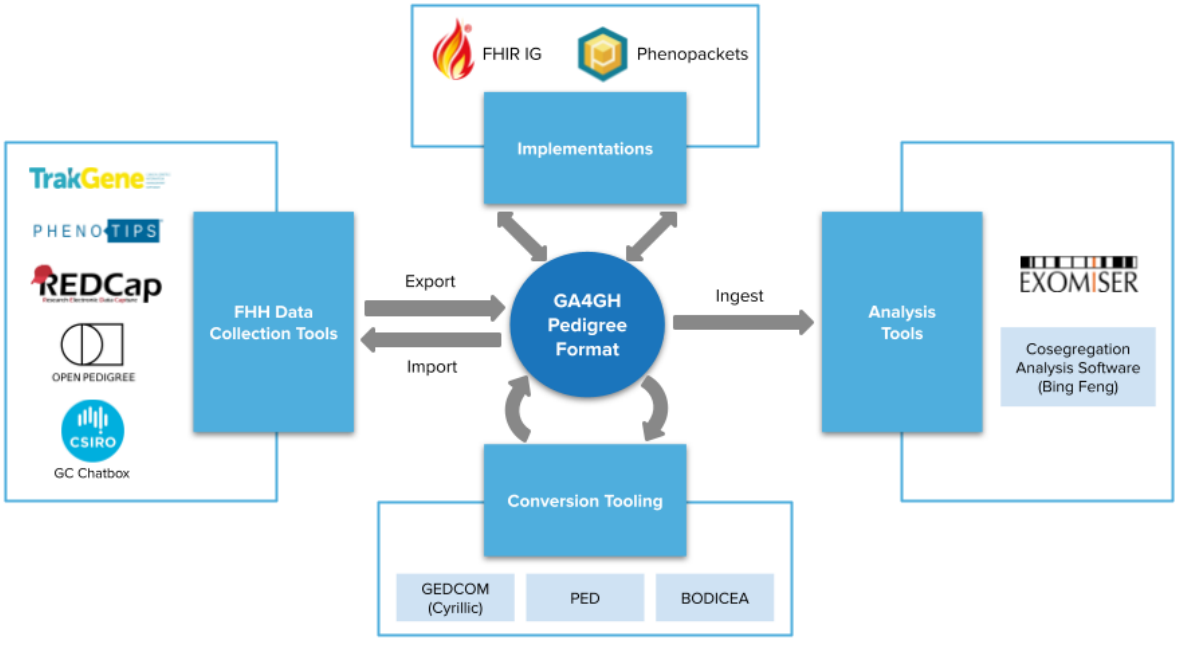
The PED format is a simple text file with 6 columns - IDs, a binary sex field, the phenotype (singular) and SNP genotypes. You can represent a basic parent-child trio, and that may cover a lot of use cases. However, you can't represent twins, things like adoption or donors, pregnancy, vital status, multiple phenotypes and data provenance. All of this type of data is important for genetic counseling and risk assessments where richer representations of relationships are valuable.

The GA4GH Pedigree Standard will natively incorporate PED to enable interoperability with legacy tools.

1.4 Example Use Cases

A full listing of the use cases that informed development can be reviewed here.

- Diagnostic genomic testing across a range of rare disease groups, with de-identified data from unsolved patients progressing into discovery research (data from solved cases also stored in research environment). Majority of testing undertaken as singletons, <5% as trios, other family configurations extremely rare (parent/child duo, sib pair, half-sib pair, quad). Pedigree information is required to inform clinical and research genomic data analysis.
- Seamlessly share collected clinical and family health history information with bioinformatics systems and research environments (or other services) to unambiguously document relationships between sequenced individuals to support joint calling of variants and filtering of variants based on segregation, as well as describing wider family history (re: non-sequenced individuals).
- Using family health history, genotype, and phenotype results of a patient or relative, to determine if the patient needs further testing or sequence analysis, and/or if a relative needs the same
- It is difficult to predict all of the secondary uses of this information and so having it in a programmatic standard that people can consume across a number of resources in both a format for analysis as well as for building algorithms and tools over would be of high utility.
- Link multiple individuals within same pedigree.
- Describe multiple phenotypic/diagnostic/genetic features per individual.
- Robustly represent relationships necessary for counseling (e.g., adoption), risk assessment (e.g., infertility, miscarriage, health history), and assisted reproduction (e.g., IVF, MRT).



The GA4GH Pedigree Ecosystem

1.5 The Common Datset for FHH

The collection and use of family health histories span medical activities from genetic research to heritable risk assessment in patient care. For all the stakeholders in this process, the goal must be data that is accurate and coded for effective analysis, and transferable between systems. To achieve this, a globally accepted and universally implemented family health history (FHH) data set should be established as a benchmark. The purpose of the common dataset document is to create an updated recommended data set that can be used not only in both research and clinical settings, but to eliminate the gap between the two disciplines. This recommendation should also guide the development of research, clinical, and patient-facing FHH data and information collection tools, applications, and data repositories. This document should only be used as informative.

Common Dataset Document

This work was inspired by the efforts of the Personalized Health Care Workgroup of the American Health Information Community, which first released its recommendation on a core family health history (FHH) minimum data set on October 25, 2007. A peer-reviewed paper was published in December 2008.

1.6 Requirement Levels

The Pedigree model uses two requirement levels.

1.6.1 Required

If a field is required, its presence is an absolute requirement of the specification, failing which the entire model is regarded as malformed. This corresponds to the key words `MUST`, `REQUIRED`, and `SHALL` in RFC2119.

1.6.2 Optional

A field is truly optional. This category can be applied to fields that are only useful for a certain type of data. For instance, the Proband ID and Type field is only required when the pedigree is used to focus on heritable risk for a specific person in the pedigree. For other use cases such as research, a Proband type may be needed.

1.7 Brief Explainer of Protobuf and HL7 FHIR

Depending on how you choose to work with the GA4GH Pedigree model, you may be working with different formats.

If using the Pedigree model in the context of Phenopackets: Phenopackets schema uses protobuf, an exchange format developed in 2008 by Google. It is recommened to review to the Wikipedia page on Protobuf and to Google's documentation for details. This page intends to get curious readers who are unfamiliar with protobuf up to speed with the main aspects of this technology, but it is not necessary to understand protobuf to use the phenopacket or pedigree schemas. Learn more about the Phenopackets here, and the draft Phenopackets implementation of Pedigree here.

If using the Pedigree model in the context of HL7 FHIR: Fast Health Interoperability Resources (FHIR) is a loosely defined base model describing things in healthcare (e.g. Patient, Specimen) and how they relate to each other, developed by Health Level 7 (HL7). The FHIR specification is completely technology agnostic. Thus, it does not depend on programming languages or include things like relational database schemas. It is up to the implementers to decide how to implement the data model (i.e. relational database, nosql database, etc) and RESTful API. To learn more about FHIR, we recommend you check out the following resources: HL7.org, FHIR Basics, and this excellent FHIR 101 Jupyter Notebook developed by NIH Cloud-based Platform Interoperability (NCPI) Working Groups. Learn more about the Pedigree FHIR Implementation Guide here.

CHAPTER

TWO

PEDIGREE MODEL

2.1 Overview

To support the interoperability of family health history data within and between existing standards (such as HL7 FHIR and Phenopackets), the GA4GH Clinical and PhenoTips Data Capture Workstream developed the Pedigree Conceptual Model.

The Pedigree Conceptual Model defines core classes and their properties, and is based on the A Recommendation for The Common Data Set for Family Health History.

2.2 Key Concepts

The model defines three core classes:

2.2.1 Individual

A person or entity.

- Individual id - required
- Sex at birth - required
- Gender
- Name
- DOB
- Age / Age Range / Estimated Age / Gestational Age
- REA - Concept - suggested list of concepts from HANCESTRO
- Deceased
- Disease/condition: code, onset, contributed to death
- Affected: Y/N/? - for backwards-compatibility with PED
- Other risk-relevant observations

2.2.2 Relationship

A relationship that one individual has with another relative.

- individual - required
- relationship - required, coded using KIN terms
- relative - required

2.2.3 Pedigree

A collection of information about related individuals and relationships between them.

- ID - Required
- Index patients (proband, consultand, first person tested positive for a particular condition/variant) - Individual - Type enum: Proband, Consultand, First Person Tested Positive
- Completion status
- Language
- Narrative
- Date collected/updated

2.3 Direction of Relationships

A Relationship defines a relationship between one individual and another, such as *isBiologicalMotherOf* or *isTwinOf*. Only one of the two directions needs to be specified, and it does not matter which.

Symmetric relationships are those where both individuals share the same relationship with one another. These include: *isTwinOf* and *isPartnerOf*.

Non-symmetric relationships are those where the relationship that individual X has to individual Y is not the same as the relationship that individual Y has to individual X. For example, if individual X has relationship *isBiologicalParentOf* to individual Y, then individual Y has relationship *isBiologicalChildOf* individual X.

Because of this inherent flexibility in the way that relationships can be described, we define the notion of a **minimum standard form** for describing a pedigree. A pedigree in minimum standard form: 1. Has explicit parent-child relationships between all parents and their offspring, and they are directed downwards, with the parent as the individual and the child as the relative. 2. Has sibling relationships only when this is not implied by having shared parents, and in the event of multiple siblings, all sibling relationships are defined relative to the same individual 3. Defines all twin relationships relative to the same individual 4. Has partnership relationships only when this is not implied by having shared children 5. Has extended relative relationships only when this is not implied by the previously-defined relationships, and they are directed downwards, with the ancestor as the individual and the descendant as the relative.

2.4 Compatible standards

Compatible standards provide an implementation guide for capturing and representing pedigree data in a manner that is compatible with this model.

The representation of each core data element within each standard is summarized in *Classes*.

The current list of compatible standards are:

Phenopackets
: A Phenopacket implementation guide is currently underway. At the moment, an aligned implementation is defined on a branch of the Phenopacket repository: https://github.com/phenopackets/phenopacket-schema/blob/pedigree/src/main/proto/ga4gh/pedigree/v1/pedigree.proto

HL7 FHIR
: The FHIR Implementation Guide is here: https://github.com/GA4GH-Pedigree-Standard/pedigree-fhir-ig

2.5 Examples

The following examples demonstrate the way in which pedigrees of various complexity can be represented using the pedigree model.

The precise representation within the context of one of the standards, such as FHIR or Phenopacket.

2.5.1 Basic Trio

A basic family trio consists of one male parent, one female parent, and a child. This would be represented as a Pedigree with three Individuals and two parent-child Relationships:

```
individuals:
  -
    id: MOTHER
    sex: FEMALE
  -
    id: FATHER
    sex: MALE
  -
    id: CHILD
    sex: UNKNOWN
relationships:
  -
    individual: MOTHER
    relationship: isBiologicalMotherOf
    relative: CHILD
  -
    individual: FATHER
    relationship: isBiologicalFatherOf
    relative: CHILD
```

2.5.2 Twins

The relationship between twins (TWIN1 and TWIN2) can be represented by adding another Individual, parent-child relationships and a twin Relationship to the Pedigree:

```
individuals:
  -
    id: MOTHER
    sex: FEMALE
  -
    id: FATHER
    sex: MALE
  -
    id: TWIN1
    sex: UNKNOWN
  -
    id: TWIN2
    sex: UNKNOWN
relationships:
  -
    individual: MOTHER
    relationship: isBiologicalMotherOf
    relative: CHILD
  -
    individual: FATHER
    relationship: isBiologicalFatherOf
    relative: CHILD
  -
    individual: TWIN1
    relationship: isMonozygoticTwinOf
    relative: TWIN2
```

The parent-child relationships for TWIN2 are not strictly necessary. Because the *isMonozygoticTwinOf* relationship is symmetric, it would be equally valid to have said that TWIN2 isMonozygoticTwinOf TWIN1.

2.5.3 Adoption

```
individuals:
  -
    id: MOTHER
    sex: FEMALE
  -
    id: BIOLOGICAL_MOTHER
    sex: FEMALE
  -
    id: FATHER
    sex: MALE
  -
    id: CHILD
    sex: UNKNOWN
relationships:
  -
```

(continues on next page)

(continued from previous page)

```
    individual: MOTHER
    relationship: isAdoptiveParentOf
    relative: CHILD
  -
    individual: BIOLOGICAL_MOTHER
    relationship: isBiologicalMotherOf
    relative: CHILD
  -
    individual: FATHER
    relationship: isBiologicalFatherOf
    relative: CHILD
```

2.5.4 IVF

```
individuals:
  -
    id: MOTHER
    sex: FEMALE
  -
    id: SURROGATE
    sex: FEMALE
  -
    id: FATHER
    sex: MALE
  -
    id: CHILD
    sex: UNKNOWN
relationships:
  -
    individual: MOTHER
    relationship: isOvumDonorOf
    relative: CHILD
  -
    individual: SURROGATE
    relationship: isGestationalCarrierOf
    relative: CHILD
  -
    individual: FATHER
    relationship: isBiologicalFatherOf
    relative: CHILD
```

2.5.5 Complete cancer family

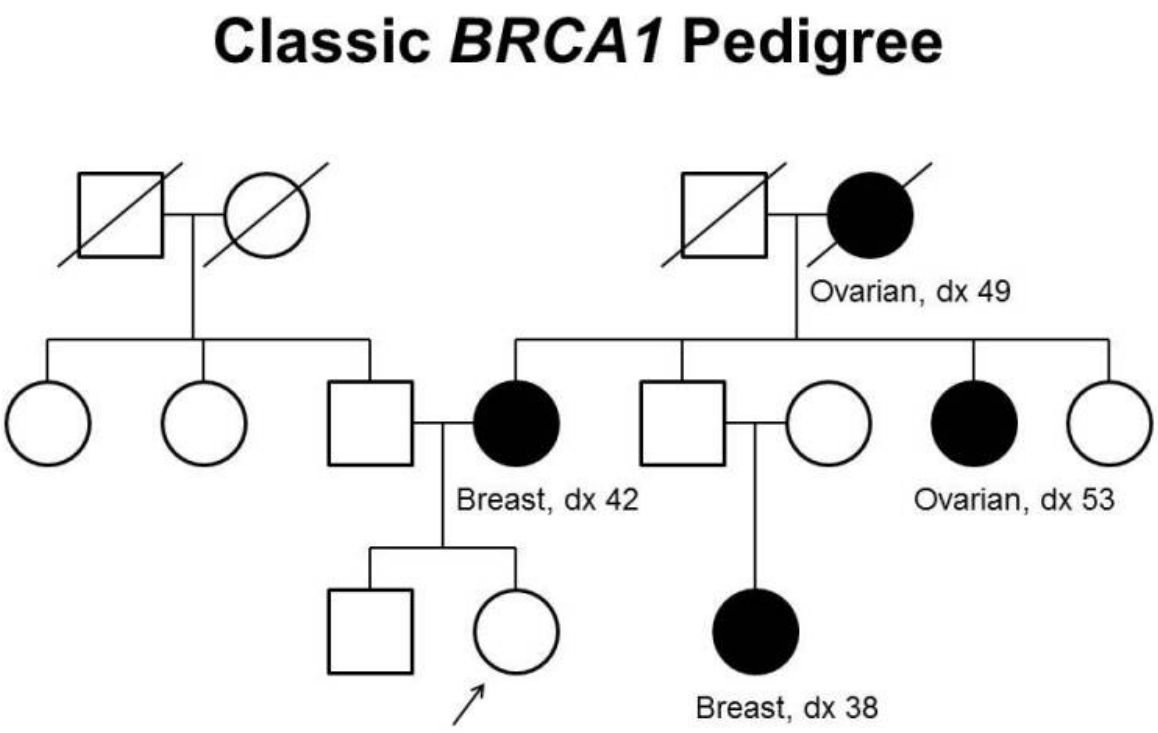


Fig. 1: Example BRCA1 pedigree. Source: https://visualsonline.cancer.gov/details.cfm?imageid=10436

```
index:
  -
    id: 14
    type: proband
individuals:
  -
    id: 1
    sex: MALE
    deceased: true
  -
    id: 2
    sex: FEMALE
    deceased: true
  -
    id: 3
    sex: MALE
    deceased: true
  -
```

(continues on next page)

(continued from previous page)

```
  id: 4
  sex: FEMALE
  deceased: true
  attributes:
    -
      term: Ovarian cancer
      ageAtDiagnosis: 49 yrs
-
  id: 5
  sex: FEMALE
-
  id: 6
  sex: FEMALE
-
  id: 7
  sex: MALE
-
  id: 8
  sex: FEMALE
  attributes:
    -
      term: Breast cancer
      ageAtDiagnosis: 42 yrs
-
  id: 9
  sex: MALE
-
  id: 10
  sex: FEMALE
-
  id: 11
  sex: FEMALE
  attributes:
    -
      term: Ovarian cancer
      ageAtDiagnosis: 53 yrs
-
  id: 12
  sex: FEMALE
-
  id: 13
  sex: MALE
-
  id: 14
  sex: FEMALE
-
  id: 15
  sex: FEMALE
  attributes:
    -
      term: Breast cancer
      ageAtDiagnosis: 38 yrs
```

(continues on next page)

(continued from previous page)

```
relationships:
  -
    individual: 1
    relationship: isBiologicalFatherOf
    relative: 5
  -
    individual: 2
    relationship: isBiologicalMotherOf
    relative: 5
  -
    individual: 1
    relationship: isBiologicalFatherOf
    relative: 6
  -
    individual: 2
    relationship: isBiologicalMotherOf
    relative: 6
  -
    individual: 1
    relationship: isBiologicalFatherOf
    relative: 7
  -
    individual: 2
    relationship: isBiologicalMotherOf
    relative: 7
  -
    individual: 3
    relationship: isBiologicalFatherOf
    relative: 8
  -
    individual: 4
    relationship: isBiologicalMotherOf
    relative: 8
  -
    individual: 3
    relationship: isBiologicalFatherOf
    relative: 9
  -
    individual: 4
    relationship: isBiologicalMotherOf
    relative: 9
  -
    individual: 3
    relationship: isBiologicalFatherOf
    relative: 11
  -
    individual: 4
    relationship: isBiologicalMotherOf
    relative: 11
  -
    individual: 3
    relationship: isBiologicalFatherOf
```

(continues on next page)

(continued from previous page)

```
    relative: 12
  -
    individual: 4
    relationship: isBiologicalMotherOf
    relative: 12
  -
    individual: 7
    relationship: isBiologicalFatherOf
    relative: 13
  -
    individual: 8
    relationship: isBiologicalMotherOf
    relative: 13
  -
    individual: 9
    relationship: isBiologicalFatherOf
    relative: 14
  -
    individual: 10
    relationship: isBiologicalMotherOf
    relative: 14
  -
    individual: 7
    relationship: isBiologicalFatherOf
    relative: 15
  -
    individual: 8
    relationship: isBiologicalMotherOf
    relative: 15
```

2.6 Design motivations

Design motivation:

- avoid overlap with other standards (fhir, phenopacket)
- focus on relationship
- graphical model, bringing relationships as top-level entities
- allow for the synthesizing of patient-reported family history data, such as comes out of family history questionnaires and EHR records (and can be represented with the FamilyMemberHistoryResource), and support this information through to risk models
- provide a standard interface for validation
- facilitate conversion among existing standards for pedigree data

Relationships between individuals are standardized using concepts from the newly developed Kinship Ontology. To allow existing workflows and tools to gracefully add interoperability with this standard, we developed an open-source pedigree data interoperability library, pedigree-tools.

CHAPTER

THREE

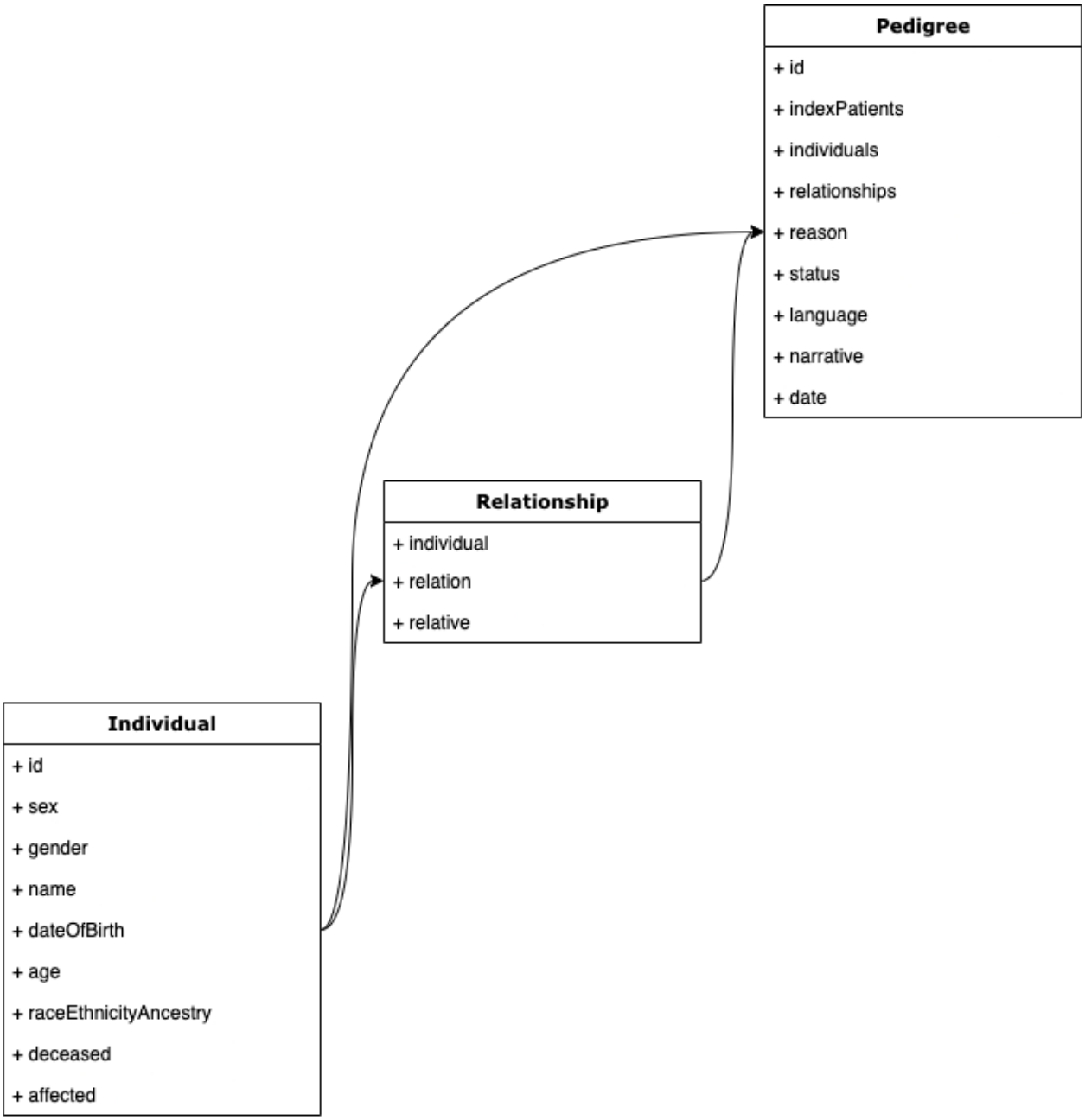
CLASSES

The diagram below shows an overview of the pedigree classes. Lines between classes indicate composition.

3.1 Individual

The subject of a **Pedigree** is represented by an **Individual** class. This class intends to represent an individual person or patient who is a member of the pedigree being investigated.

Field	Multiplicity	Description
id	1..1	External identifier for the individual
sex	1..1	Sex assigned at birth
gender	0..1	Presumed or reported gender identity
name	0..1	Name of the individual
dateOfBirth	0..1	Birth date of the individual, can be just birth year in most cases
age	0..1	Age of the individual, can be either Age, Estimated Age (or Ontology Class), Age Range, and/or Gestational Age; See also Phenopackets' TimeElement.
raceEthnicityAncestry	0..*	Race, Ethnicity, or Ancestry of the individual; terms from the Human Ancestry Ontology (HANCESTRO) are recommended.
deceased	0..1	The presumed/accepted life status of the individual as of the pedigree collection date
affected	0..1	Whether or not the individual is affected by the condition being investigated in this pedigree (`Pedigree.reason`)


Pedigree
+ id
+ indexPatients
+ individuals
+ relationships
+ reason
+ status
+ language
+ narrative
+ date
Relationship
+ individual
+ relation
+ relative
Individual
+ id
+ sex
+ gender
+ name
+ dateOfBirth
+ age
+ raceEthnicityAncestry
+ deceased
+ affected

3.2 Relationship

The *Relationship* class defines the family relations in a pedigree.

Field	Multiplicity	Description
individual	1..1	Identifier of the subject `Individual`; equivalent to the Biolink "subject"
relation	1..1	The relationship the `individual` has to the `relative` (*e.g.*, if the `individual` is the `relative`'s biological mother, then relation could be `isBiologicalMotherOf [KIN:027]`); terms should come from the KIN Ontology.
relative	1..1	Identifier of the relative `Individual`; equivalent to the Biolink "object"

3.3 Pedigree

A clinical **Pedigree** is curated selection of information about a family, including the individuals, relationships between them, and relevant health conditions.

Field	Multiplicity	Description
id	1..1	External identifier for the family being investigated
indexPatients	0..*	Identified `Individual` in the family of a health condition of focus being investigated: `Proband`, `Consultand`, `First Person Tested Positive`
individuals	0..*	Collection of `Individual` who are the members of this pedigree
relationships	0..*	Collection of `Relationship` between the `individuals` who are the members of this pedigree
reason	0..1	The reason for pedigree collection, a health condition of focus being investigated in the family; if any `Individual` has the `affected` property defined, it refers to this condition.
status	0..1	Status of the pedigree resource collection
narrative	0..1	Summary of the pedigree resource for human interpretation
date	0..1	The date the pedigree was collected or last updated, as ISO full or partial date, *i.e.* `YYYY`, `YYYY-MM`, or `YYYY-MM-DD`

CHAPTER

FOUR

WORKING WITH THE PEDIGREE MODEL

4.1 Kinship Ontology (KIN)

The Kinship Ontology (KIN) is a family relations ontology developed as part of the Global Alliance for Genomics and Health Pedigree Standard project. It allows using an OWL reasoner to automatically validate a family history graph and infer new relations.

The latest version of the ontology can be found at: http://purl.org/ga4gh/kin.owl.

The Ontology is open-source and managed in this GitHub repo: https://github.com/GA4GH-Pedigree-Standard/family_history_terminology

4.2 Pedigree Tools

Pedigree-tools is a library for supporting the conversion of pedigree data between various file formats.

It can currently support importing from the following formats:

- GA4GH Pedigree
- PED/Linkage
- GEDCOM (Cyrillic)
- BOADICEA

In can currently export into the following formats:

- GA4GH Pedigree
- PED/Linkage

This tool is available at the following GitHub repository: https://github.com/GA4GH-Pedigree-Standard/pedigree-tools

4.3 Pedigree Validator

This is a simple command line application that shows how validation of a FHIR pedigree file can be implemented using the HAPI FHIR libraries and the artifacts produced by the FHIR implementation guide.

It also shows how an OWL reasoner can be used to implement additional validation based on the KIN ontology.

The application is available at the following GitHub repository: https://github.com/GA4GH-Pedigree-Standard/pedigree-validator

4.4 Example Implementations

The following systems have implemented the GA4GH Pedigree Standard:

FHIR implementations:

- CSIRO Redcap Pedigree Plugin (Open Source)
- Open Pedigree (Open Source)

Phenopacket implementations:

- In progress…

CHAPTER

FIVE

ACKNOWLEDGEMENTS

This standard was developed by Clinical and Phenotpic Data Capture Work Stream of the GA4GH, and is the result of the collaborative work, comments, and input of many individual and organizational contributors. We thank all contributors for their time and expertise.

5.1 Pedigree Standard Contributors (in alphabetical order)

Louis Bergelson (Broad Institute)
Eva Bermejo (EJP RD)
Chris Bun (CancerIQ)
Orion Buske (PhenoTips)
Hannah Calkins (CHOP)
Chen Chen (GA4GH)
Melissa Cline (UCSC, BRCA Exchange)
Melissa Cook (NCI, CRDC)
Shahim Essaid (OHSU)
Alex Felmeister (Illumina)
Bingjian Feng (Univeristy of Utah, BRCA Exchange)
Dietmar Fernandez (EGA)
Michael Franklin (Centre for Population Genomics, Garvan Institute)
Sean Garin (NIH)
Lisa Glaspie (CancerIQ)
Melissa Haendel (University of Colorado, Monarch Initiative)
David Hansen (CSIRO, Australian Genomics)
Allison Heath (CHOP, Kids First DRC)
Tim Jackson (TrakGene)
Julius Jacobsen (QMUL, Monarch Initiative)
Katherine Johnston (H3Africa)
Meen Chul Kim (CHOP, Kids First DRC)
Guida Landouré (H3Africa)
Steven Laurie (CNAG-CRG)
Tara Lichtenberg (University of Chicago)
Mamana Mbiyavanga (University of Cape Town, H3Africa)
Alejandro Metke (CSIRO, Australian Genomics)
Moni Munoz-Torres (University of Colorado, Monarch Initiative)
Thanh-Phuong Nguyen (Megeno S.A.)

Soichi Ogishima (Tohoku University, GEM-Japan)
Kevin Power (Children's Mercy Hospital)
Peter Robinson (Jackson Laboratory, Monarch Initiative)
Richard Scott (Genomics England)
Natasha Singh (CHOP, D3B)
Neerjah Skantharajah (GA4GH)
Lindsay Smith (GA4GH)
Amanda Spurdle (QIMR Berghofer Medical Research Institute, BRCA Exchange)
Zornitza Stark (Australian Genomics)
Deanne Taylor (CHOP)
Alex Tsai (GA4GH)
Katheryn Van Diemen (TrakGene)
Grant Wood (MyGenomeTrust)
Teruhiko Yoshida (National Cancer Center, Japan)

5.2 Driver Project Survey Participants (if not listed above)

Luca Barcella (EJP RD)
Sergi Beltran (CRG, EJP RD)
Eva Brezinova (EJP RD)
Erwin Brosens (EJP RD)
Candice Feben (DDD-Africa)
Krisztian Gaspar
James Gyamfi (Genomics England)
Vesta Kucinskiene (EJP RD)
Audald Lloret-Villas (EGA)
Balzas Mayer (Semmelweis University)
Anna Need (Genomics England)
Oscar Nyangiri (H3Africa)
Olusola Omosaiye (Genomics England)
Suzanne Pasmans (ErasmusMC, EJP RD)
Alessia Pepe
Anita Rauch (Universität Zürich)
Elzbieta Radzikowska
Marina Vivarelli (EJP RD)
Daryl Waggott (Genome Canada)
Zhenyu Zhang (Univesity of Chicago, NCI GDC)

5.3 Special Thanks To

Robert Freimuth (Mayo Clinic)
Hoa Ngo (CSIRO)

5.4 Funding

J.J. would like to acknowledge National Institutes of Health (NIH) grants 1R24OD011883 and NIH, National Institute of Child Health and Human Development 1R01HD103805-01
A.B.S. was supported by an NHMRC Investigator Fellowship (APP177524)