Global Alliance
for Genomics & Health

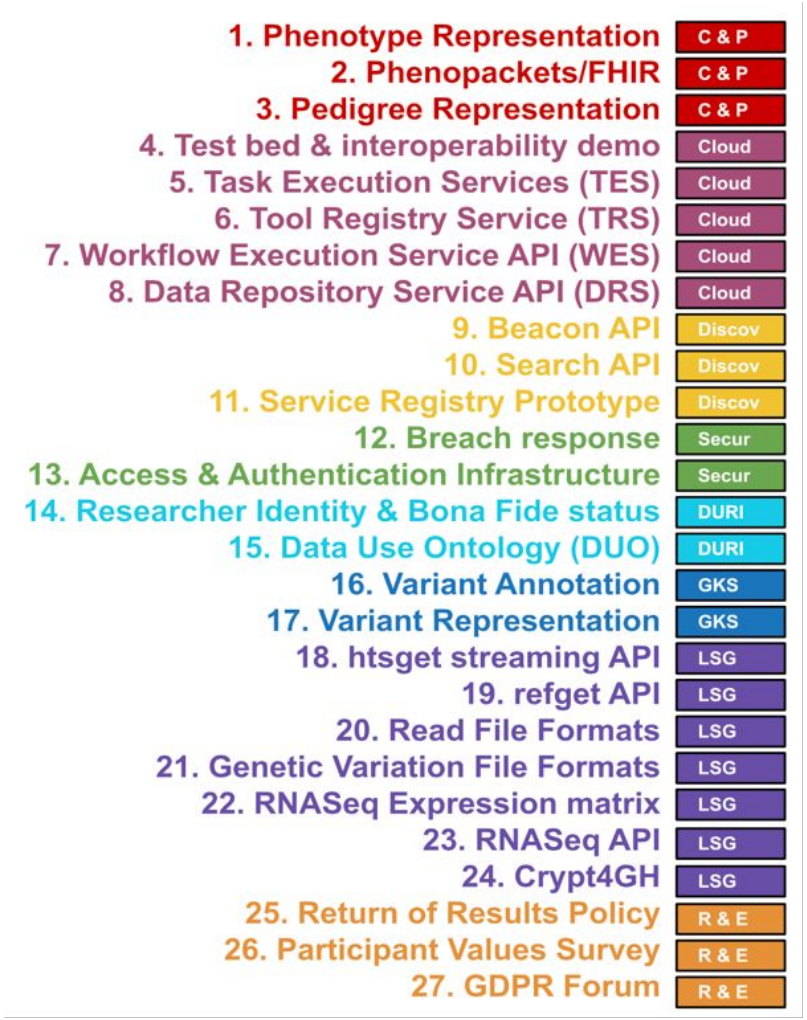
Collaborate. Innovate. Accelerate.

SchemaBlocks

Michael Baudis, Melanie Courtot

Global Alliance for Genomics & Health

- **Why SchemaBlocks? What is it?**
  - Intro & use cases
- **Goals of the session**
  - Acceptance of the path to product
  - Clear planning for workstream engagement and targeting of which need to be engaged next and by whom
  - Which driver projects to work with in the *short term*
    - Establishing of working principles
    - Solving of immediate unmet needs (e.g. Beacon variants …)
  - Alignment with driver projects, e.g. HCA



Global Alliance
for Genomics & Health

SchemaBlocks - Perceived Need

Global Alliance for Genomics & Health

- "GA4GH schemas" by the DWG provided object model and documentation
- rigid, top-down managed development model was abandoned => WS + DP
- now no place - outside individual WS & DP - in GA4GH ecosystem to provide
  - Data models
  - Standard recommendations
  - Object prototypes
- lack of shared objects & documentation leads to duplicate development efforts and lack of citable references - examples:
  - Use of genome coordinates in GA4GH products?
  - Variant formats (placeholders, future ...) e.g. for Beacon, Search ...?
  - Dataset specific parameters related to consent code (DURI)?
  - Object hierarchies & relations (e.g. dataset | subject | sample | callset | variant ...)?
  - How to use external reference systems (e.g. ontologies) in queries and data delivery?

1. Phenotype Representation	C & P
2. Phenopackets/FHIR	C & P
3. Pedigree Representation	C & P
4. Test bed & interoperability demo	Cloud
5. Task Execution Services (TES)	Cloud
6. Tool Registry Service (TRS)	Cloud
7. Workflow Execution Service API (WES)	Cloud
8. Data Repository Service API (DRS)	Cloud
9. Beacon API	Discov
10. Search API	Discov
11. Service Registry Prototype	Discov
12. Breach response	Secur
13. Access & Authentication Infrastructure	Secur
14. Researcher Identity & Bona Fide status	DURI
15. Data Use Ontology (DUO)	DURI
16. Variant Annotation	GKS
17. Variant Representation	GKS
18. htsget streaming API	LSG
19. refget API	LSG
20. Read File Formats	LSG
21. Genetic Variation File Formats	LSG
22. RNASeq Expression matrix	LSG
23. RNASeq API	LSG
24. Crypt4GH	LSG
25. Return of Results Policy	R & E
26. Participant Values Survey	R & E
27. GDPR Forum	R & E

Global Alliance
for Genomics & Health

Common data structures, formats that could be aligned?

SchemaBlocks - History & Status

Global Alliance for Genomics & Health

- Started by members of C/P & GKS, as ***continuation*** of former DWG Metadata work & other parts from GA4GH Schemas
  - core data model, objects
  - documentation
- Integration and exchange with *Phenopackets*, *Beacon* developments
- Maintained updated documentation and models in the Metadata repository
- December 2018:
  - first call with participants of different WS (GKS, C/P, Discovery)
  - launch of Github organisation “ga4gh-schemablocks”
  - New website @ schemablocks.org, with some initial documentation

SchemaBlocks - Emerging Principles

Global Alliance for Genomics & Health

- Machine readable ”blocks”, with lightweight structure
  - Feedback from Jules & Ben, interest in aligning representation with documentation
  - Common schemas that can be validated against
- Human readable documentation
  - representing block descriptions & examples, also standards & conventions
- Competing standards and alternative objects entirely possible
  - e.g. different variant standards & coordinate systems - VCF | VMC | Beacon
  - external references to non-GA4GH standards, e.g. ISO, IEEE
- Cross-cutting initiative: Not “part of” a single WS
  - **C/P** & **GKS** (+ others, drivers...) for **standards**; requirements ... by Discovery
- Aligns with GA4GH standard setting mission

Not an attempt to build a “one size fits all”, monolithic schema

Use cases: GA4GH products and SchemaBlocks

SchemaBlocks - Phenopacket use case

Global Alliance for Genomics & Health

- A product of the ClinPheno group, nearing product approval
- Extensive documentation
- Distributed in Protobuf
- Includes Phenopacket building blocks, built from pre existing GA4GH MTT elements and collaborative discussions

Phenopacket building blocks

The phenopacket standard consists of several protobuf messages each of which contains information about a certain topic such as phenotype, variant, pedigree, and so on. One message can contain other messages, which allows a rich representation of data. For instance, the Phenopacket message contains messages of type Individual, Phenotype, Biosample, and so on. Individual messages can therefore be regarded as building blocks that are combined to create larger structures. It would also be straightforward to include the Phenopackets schema into larger schema for particular use cases, which we will cover in the Discussion. Follow the links to read more information about individual building blocks.

- Age
- AgeRange
- Disease
- Evidence element
- External Reference element
- File and HtsFile
- Gene
- Geolocation
- Individual
- karyotypic sex
- Metadata
- Ontology Class
- Pedigree
- Phenotype
- Procedure
- Resource
- Sex
- Variant element

Ontology representation - an example

Global Alliance for Genomics & Health

Ontology Class

This element is used to represent classes (terms) from ontologies, and is used in many places throughout the Phenopacket standard. It is a simple, two element message that represents the identifier and the label of an ontology class.

```
message OntologyClass {
   string id = 1;
   string label = 2;
}
```

The ID should be a CURIE-style identifier e.g. HP:0100024, MP:0001284, UBERON:0001690, i.e., the primary key for the ontology class. The label should be the corresponding class name. The Phenopacket standard requires that the id and the label match in the original ontology. We note that occasionally, ontology maintainers change the primary label of a term. The id used in a Phenopacket should match the label of the version of the ontology indicated in the metadata element.

Phenopacket requirements for the `OntologyClass` *element*

Field	Example	Status
id	HP:0001875	required
label	Neutropenia	required

The FHIR mapping is a CodeableConcept. See also Coding.

Ontology_term

Status: proposed

Properties of the *Ontology_term* class

Property	Type	Format	Description
id	string		properly prefixed CURIE of the ontology term
label	string		the text label associated with the term

Description

Ontology_term represents the core object used to reference domain-specific entities, as well as to identify their domains through the appropriate prefix. CURIES are case sensitive, although for prefixes this practice is inconsistently followed.

Examples

```
{
   "id" : "HP:0003621",
   "label" : "Juvenile onset"
}
```

```
{
   "id" : "ncit:C3058",
   "label" : "Glioblastoma"
}
```

Phenopackets - SchemaBlocks desiderata

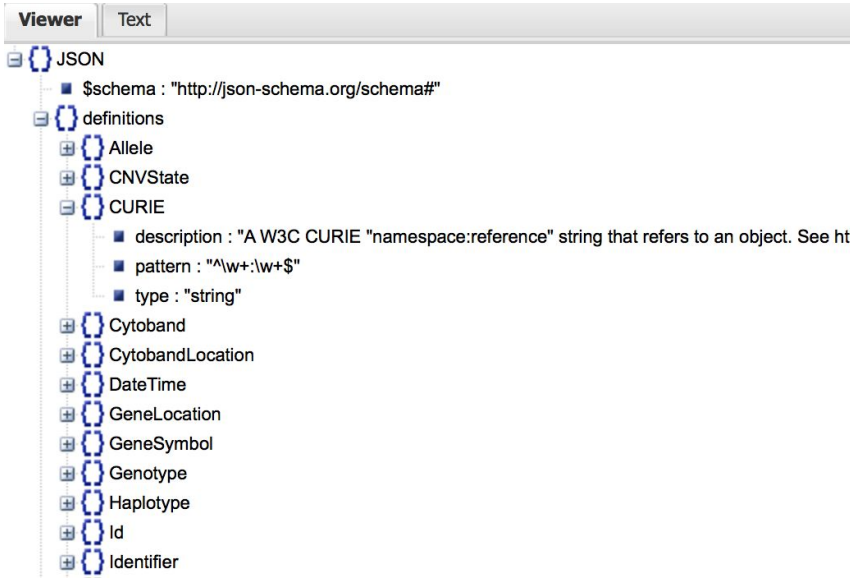
Global Alliance for Genomics & Health

- It would be good for ga4gh to converge on a single phenotype standard that would be implemented in phenopackets and reflected in {S}[B]
- Beacon wants phenotype queries, we need to know the correct ontology terms and this should not be something that one project does on its own.
- We have to be cautious that we don't over align, need to allow for divergent ways to work. *SchemaBlocks* not RoadBlocks!
- There has to be a balance. People working on the standards need to listen to the community, but it will degrade the quality if we make it too hard to align. GA4GH Should say "this is the standard and if you use this, you can interact with the entire network"

SchemaBlocks - VMC use case

Global Alliance for Genomics & Health

- A product of the GKS group
- Has a JSON schema already
- Still under very active development
- Open to aligning with general GA4GH blocks



SchemaBlocks - DUO use case

Global Alliance for Genomics & Health

- A GA4GH product of the DURI group
- Distributed as an OWL file
- No consistent way at the moment to use DUO codes for implementers - Broad, Optum, BioSamples/EGA/Sanger
- Desiderata from DU to work with SB on this to have consistency with other GA4GH products

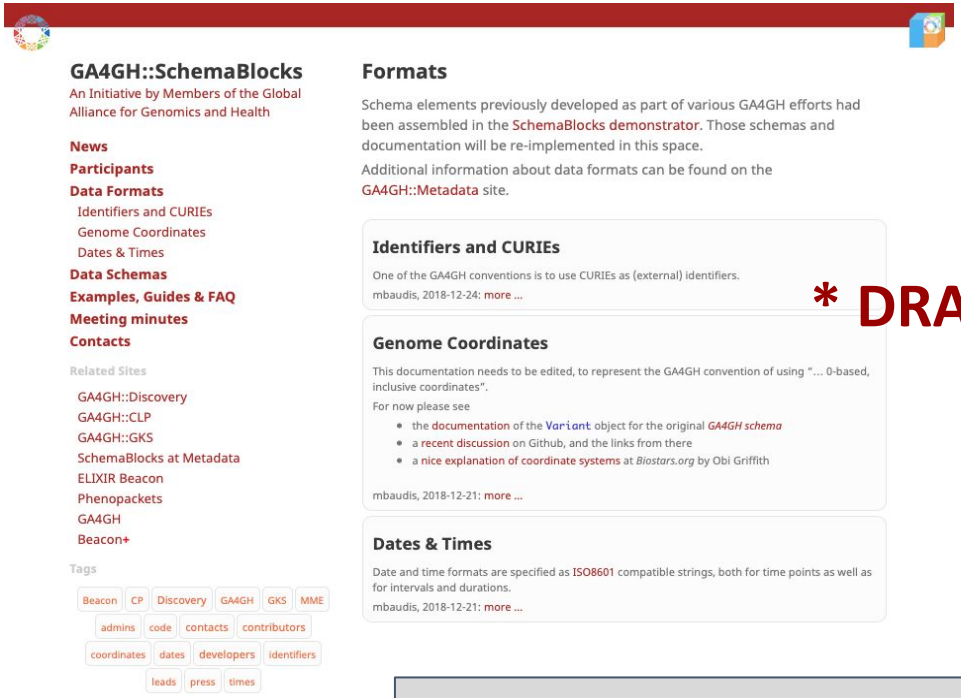
```
{
    "consent_codes": [{
        "id": "<subclass of DUO:0000001>",
        "label": "<label>"
    }],
    "restrictions": {
        "use_until": "<datetime> for time restriction",
        "publication_memorandom": "<datetime> for time restriction",
        "users": [{
            "id": "<user id>"
        }],
        "institutions": [{
            "id": "<uri> i.e http://optum.com"
        }],
        "geographic_locations": [{
            "id": "<<ISO 3166-1> USA",
            "label":"i.e USA"
        }],
        "ethics_approvals": [{
            "id": "<uri> i.e nih.gov",
            "label": "NIH"
        }],
        "commercial_use": "<bool>",
        "not_for_profit_only": "<bool>",
        "return_to_database": "<bool>",
        "publication_required": "<bool>"

    }
}
```

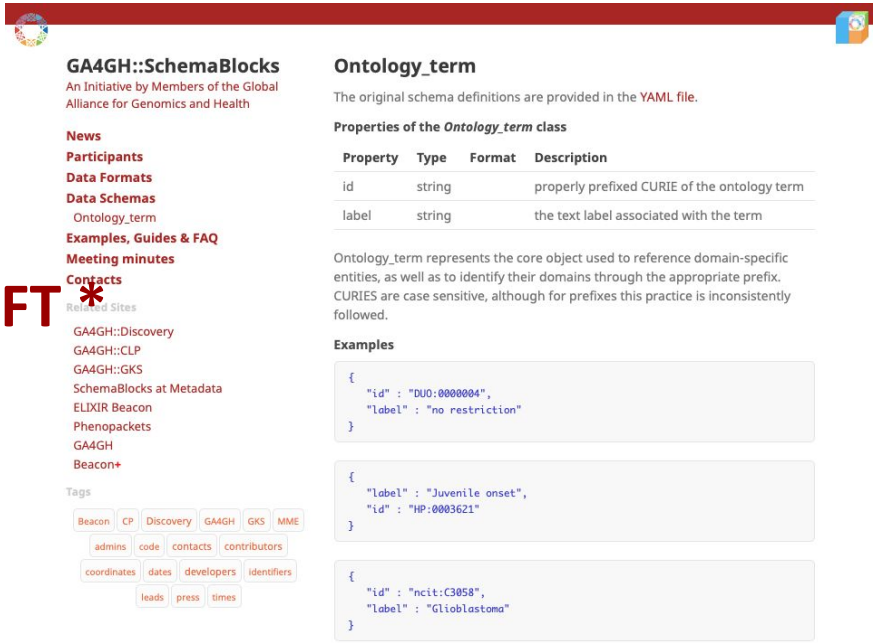
https://github.com/ga4gh-schemablocks/blocks/issues/5

SchemaBlocks - Current status

Global Alliance for Genomics & Health



* DRAFT *



Join Ben Hutton later for hands on work on JSON schema representation!

SchemaBlocks - Future Directions

Global Alliance for Genomics & Health

- Receive continuous contributions from WS in form of “blocks” and documentation through interaction w/ different development teams
  - Variant annotation types and models from **GKS**
  - Ontology, phenotype format & recommendations from **C/P** (*phenopackets…*)
  - Search components from **Discovery** & Beacon, use conditions (**DURI**)…
- Formalise approval levels & rules of engagements, leveraging existing GA4GH processes
- Become part of GA4GH product approval process
  - products document awareness of SchemaBlocks through
    - Contribution of code or documentation
    - Use of existing code or formats
    - (Or Statement about lack of applicability…)

SchemaBlocks - Blocks creation & approval

Global Alliance for Genomics & Health

- **GA4GH Governance process well established**: work streams and driver projects develop products, which undergo GA4GH review & approval process
- {S}[B] schemas & documentation reflects “GA4GH approved” concepts and conventions?
  - Interact w/ WS & DP to have “blocks” represented in a way approved by the respective projects
  - Document provenance & product use, including versioned snap-shots of implementations (?)
  - Include “under development” concepts, with clear labeling & designation

This concept needs designated “executive level” volunteers from DP & WS

SchemaBlocks - Practical next steps

Global Alliance for Genomics & Health

- How do we formalise this in the GA4GH structure?
  - Currently “An initiative by members of the GA4GH”, linked from Discovery...
  - scheduling regular calls, minutes (Melissa Konopko)
- Structure | leadership | subgroups
  - formal set-up with dedicated WS interaction, e.g. WS leads representation w/ dedicated “sign-off” votes for documentation?
- Future place in product development & approval processes?
  - Ensuring interoperability through requesting documented {S}[B] interaction during product review...
  - Early for decision - but suggestions about direction?