Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

SchemaBlocks

Michael Baudis

SchemaBlocks - Perceived Need

- “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?

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
- This SC meeting: Feedback & visibility will shape future directions

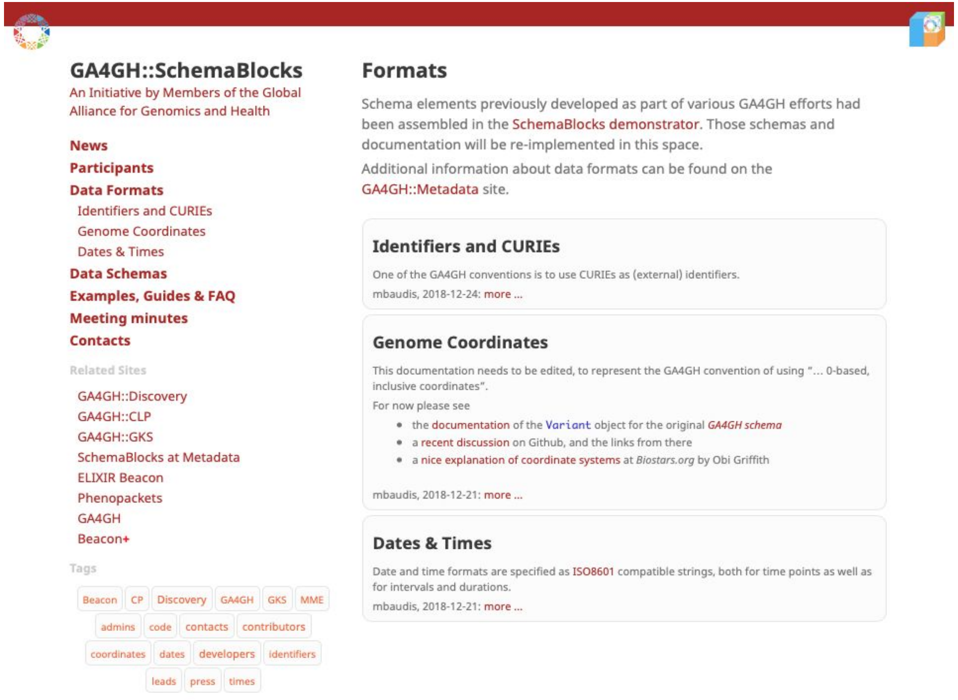
SchemaBlocks - Emerging Principles

- Machine readable ”blocks”, with lightweight structure
  - e.g. JSON schema as YAML
  - precedence of *documentation* over implementation
- 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

SchemaBlocks - Standards and Code

Global Alliance for Genomics & Health



GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

- **News**
- **Participants**
- **Data Formats**
  - Identifiers and CURIEs
  - Genome Coordinates
  - Dates & Times
- **Data Schemas**
- **Examples, Guides & FAQ**
- **Meeting minutes**
- **Contacts**

Related Sites

- GA4GH::Discovery
- GA4GH::CLP
- GA4GH::GKS
- SchemaBlocks at Metadata
- ELIXIR Beacon
- Phenopackets
- GA4GH
- Beacon+

Tags

Beacon CP Discovery GA4GH GKS MME admins code contacts contributors coordinates dates developers identifiers leads press times

Formats

Schema elements previously developed as part of various GA4GH efforts had been assembled in the SchemaBlocks demonstrator. Those schemas and documentation will be re-implemented in this space.

Additional information about data formats can be found on the GA4GH::Metadata site.

Identifiers and CURIEs

One of the GA4GH conventions is to use CURIEs as (external) identifiers.

mbaudis, 2018-12-24: more ...

Genome Coordinates

This documentation needs to be edited, to represent the GA4GH convention of using "… 0-based, inclusive coordinates".

For now please see

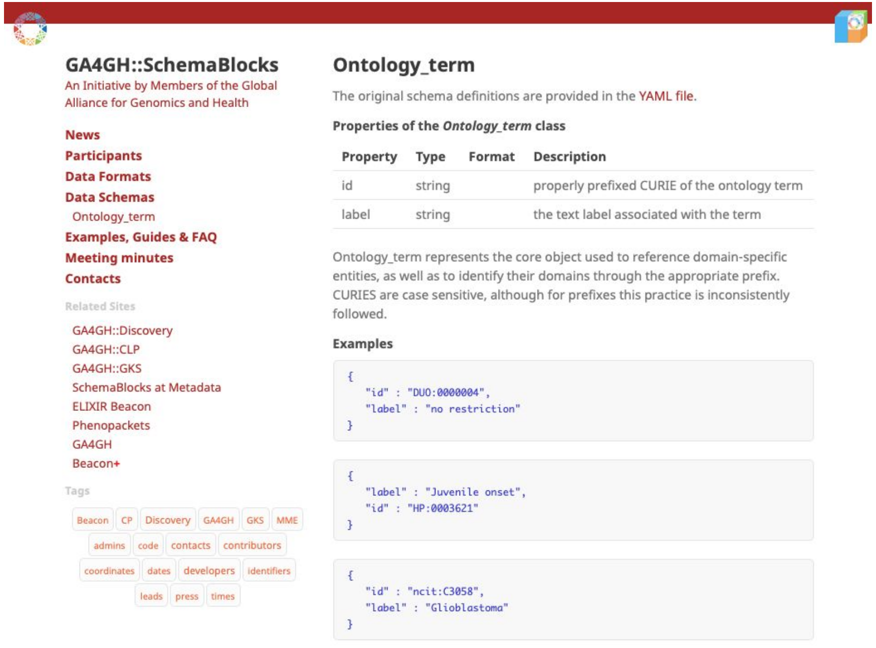
- the documentation of the `Variant` object for the original *GA4GH schema*
- a recent discussion on Github, and the links from there
- a nice explanation of coordinate systems at *Biostars.org* by Obi Griffith

mbaudis, 2018-12-21: more ...

Dates & Times

Date and time formats are specified as ISO8601 compatible strings, both for time points as well as for intervals and durations.

mbaudis, 2018-12-21: more ...



GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

- **News**
- **Participants**
- **Data Formats**
- **Data Schemas**
  - Ontology_term
- **Examples, Guides & FAQ**
- **Meeting minutes**
- **Contacts**

Related Sites

- GA4GH::Discovery
- GA4GH::CLP
- GA4GH::GKS
- SchemaBlocks at Metadata
- ELIXIR Beacon
- Phenopackets
- GA4GH
- Beacon+

Tags

Beacon CP Discovery GA4GH GKS MME admins code contacts contributors coordinates dates developers identifiers leads press times

Ontology_term

The original schema definitions are provided in the YAML file.

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

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.

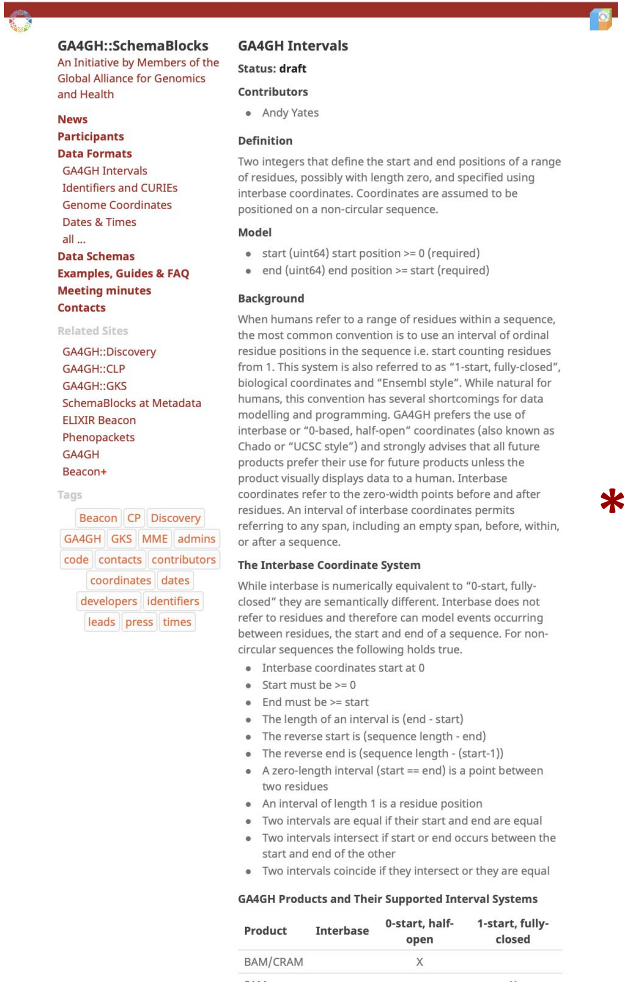
Examples

```
{
    "id" : "DUO:0000004",
    "label" : "no restriction"
}
```

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

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

SchemaBlocks - A GA4GH Community Initiative



GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

GA4GH Intervals

Status: draft

Contributors

- Andy Yates

Definition

Two integers that define the start and end positions of a range of residues, possibly with length zero, and specified using interbase coordinates. Coordinates are assumed to be positioned on a non-circular sequence.

Model

- start (uint64) start position >= 0 (required)
- end (uint64) end position >= start (required)

Background

When humans refer to a range of residues within a sequence, the most common convention is to use an interval of ordinal residue positions in the sequence i.e. start counting residues from 1. This system is also referred to as “1-start, fully-closed”, biological coordinates and “Ensembl style”. While natural for humans, this convention has several shortcomings for data modelling and programming. GA4GH prefers the use of interbase or “0-based, half-open” coordinates (also known as Chado or “UCSC style”) and strongly advises that all future products prefer their use for future products unless the product visually displays data to a human. Interbase coordinates refer to the zero-width points before and after residues. An interval of interbase coordinates permits referring to any span, including an empty span, before, within, or after a sequence.

The Interbase Coordinate System

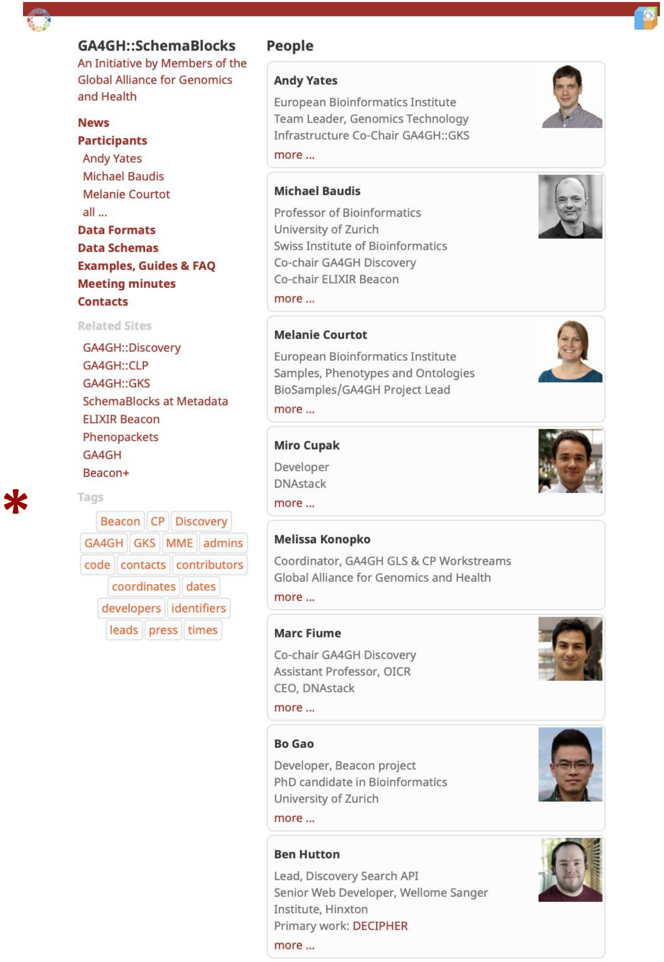
While interbase is numerically equivalent to “0-start, fully-closed” they are semantically different. Interbase does not refer to residues and therefore can model events occurring between residues, the start and end of a sequence. For non-circular sequences the following holds true.

- Interbase coordinates start at 0
- Start must be >= 0
- End must be >= start
- The length of an interval is (end - start)
- The reverse start is (sequence length - end)
- The reverse end is (sequence length - (start-1))
- A zero-length interval (start == end) is a point between two residues
- An interval of length 1 is a residue position
- Two intervals are equal if their start and end are equal
- Two intervals intersect if start or end occurs between the start and end of the other
- Two intervals coincide if they intersect or they are equal

GA4GH Products and Their Supported Interval Systems

Product	Interbase	0-start, half-open	1-start, fully-closed
BAM/CRAM		X	

*** DRAFT ***



GA4GH::SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

People

Andy Yates
European Bioinformatics Institute
Team Leader, Genomics Technology Infrastructure Co-Chair GA4GH::GKS
more ...

Michael Baudis
Professor of Bioinformatics
University of Zurich
Swiss Institute of Bioinformatics
Co-chair GA4GH Discovery
Co-chair ELIXIR Beacon
more ...

Melanie Courtot
European Bioinformatics Institute
Samples, Phenotypes and Ontologies
BioSamples/GA4GH Project Lead
more ...

Miro Cupak
Developer
DNAstack
more ...

Melissa Konopko
Coordinator, GA4GH GLS & CP Workstreams
Global Alliance for Genomics and Health
more ...

Marc Fiume
Co-chair GA4GH Discovery
Assistant Professor, OICR
CEO, DNAstack
more ...

Bo Gao
Developer, Beacon project
PhD candidate in Bioinformatics
University of Zurich
more ...

Ben Hutton
Lead, Discovery Search API
Senior Web Developer, Wellome Sanger Institute, Hinxton
Primary work: DECIPHER
more ...

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 & governance model
- 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 - Feedback?

- How do we formalise this in the GA4GH structure?
  - Currently “An initiative by members of the GA4GH”, linked from Discovery...
  - GA4GH staff support (since need for regular calls, minutes)
- Depending on that - Structure, leadership?
  - “Self-assembly” (w/ direction from WS leads) or formal set-up with dedicated WS interaction?
- Future place in product development & approval processes?
  - Early for decision - but suggestions about direction?