{S}[B] SchemaBlocks

GA4GH Standards Documentation and Alignment Initiative
Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.
Geneticists push for global data-sharing
Nature - Erika Check Hayden
June 5, 2013

DNA data to be shared worldwide in medical research project
The Guardian - Ian Sample
June 5, 2013

New alliance aims to create international system for sharing genomic data
The Globe and Mail - By André Picard
June 5, 2013

Scientists Seek Order to Potential Confusion of Gene Data
Bloomberg - Drew Armstrong & Robert Langreth
June 5, 2013

Q&A: David Altshuler on How to Share Millions of Human Genomes
Science - Jocelyn Kaiser
June 7, 2013

Accord Aims to Create Global Trove of Genetic Data
The New York Times - Gina Kolata
June 5, 2013

Geneticists push for global data-sharing
Nature - Erika Check Hayden
June 5, 2013

Our Genes, Their Secrets
The New York Times
June 18, 2013

Global alliance to create framework for sharing genomic data
The Boston Globe - Carolyn Y. Johnson
June 5, 2013

Poking Holes in Genetic Privacy
The New York Times - Gina Kolata
June 16, 2013

Une alliance pour partager les données génomiques cliniques
Le Monde - Sandrine Cabut
June 14, 2013

White House Open Science 'Champions' Highlights Genomic Data Pioneers
GenomeWeb
June 19, 2013
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.

The Global Alliance for Genomics and Health®
Organizational Structure - Work Streams & Driver Projects

GA4GH Driver Projects are real-world genomic data initiatives that help guide our development efforts and pilot our tools. Stakeholders around the globe advocate, mandate, implement, and use our frameworks and standards in their local contexts.

GA4GH Foundational and Technical Work Streams develop standards and tools that are designed to overcome technical and regulatory hurdles to international genomic data-sharing.

The GA4GH Partner Engagement initiative facilitates two-way dialogue with the international community, including national initiatives, major health care centres, and patient advocacy groups.
GA4GH :: Discovery

A Work Stream of The Global Alliance for Genomics and Health

We build standards for federated, secured networks of data and services, forming an “Internet of Genomics”, and asking meaningful questions across it.

- Marc Fiume
- Discovery Networks
- Search API / Data Discovery
- Michael Baudis
- Beacon
- SchemaBlocks (S)B

GA4GH Discovery Work Stream

Welcome to the homepage for the GA4GH Discovery Work Stream. We build standards for federated, secured networks of data and services, forming an “Internet of Genomics”, and asking meaningful questions across it.

The Discovery Work Stream is lead by Marc Fiume and Michael Baudis. For details on how this Work Stream operates please read the Discovery Work Stream Organizational Structure & Vision document.

This group meets at a high-level monthly. Meeting minutes are available to view here. In addition, the sub-groups listed below meet on their own schedules. Participation in these groups require participants to adhere to the GA4GH Standards for Professional Conduct.

For more information on GA4GH, please visit the GA4GH Website.

Products

Product development in GA4GH follows a process outlined in a GA4GH Product Approval Process Guide, in draft. Products developed by the work stream undergo an initial investigation phase, followed by a formal Proposed Product Phase, in which most of the work is done, followed by a formal Approval Phase during which the products gain GA4GH Approval. The formal steps require the approval of the Work Stream leads.

The following products are currently under development for this Work Stream.

Beacon API

A Beacon is a federated, web-accessible service that can be queried for information about a specific genomic variant, e.g. a single nucleotide polymorphism (SNP/SNV) or a copy number variation (CNV), and reports about its existence in the queried resources. Future versions of the Beacon protocol will support different usage scenarios and offer the opportunity to link to the matched data using e.g. a handover scenario.

The Beacon API specification is now coordinated through the ELDOR Beacon project and accessible there or directly through its repository.

Discovery Search API

The Discovery Search API aims at developing a component based approach towards the implementation of interfaces for genomic data and related information, for instance for global, federated data sharing through the querying, and subsequent optional processing of the results in a cloud environment. The in-development specification for the Search API can be accessed here.

Discovery Networks API

The Beacon Network was the first successful implementation of an open, federated API for world-wide querying of genome resources. Current and future developments target especially the integration of user authentication for different access levels, extensions to the query language as provided through the emerging Beacon API and the evaluation of different topologies, especially with respect to security concerns.
GA4GH {S}[B]

SchemaBlocks

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, data formats and semantics
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- no attempt to develop a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
- currently not “authoritative GA4GH recommendations”
- recognized in GA4GH roadmap as element in "TASC" effort

schemablocks.org

GA4GH SchemaBlocks

SchemaBlocks is a “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics.

Launched in December 2018, this project is still to be considered a “community initiative”, with developing participation, leadership and governance structures. At its current stage, the documents can not be considered “authoritative GA4GH recommendations” but rather represent documentation and implementation examples provided by GA4GH members.

While future products and implementations may be completely based on SchemaBlocks components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

The SchemaBlocks site can be accessed though the permanent link schemablocks.org. More information about the different products & formats can be found on the workstream sites. For reference, some of the original information about recommended formats and object hierarchies is kept in the GA4GH Metadata Repositories.

For more information on GA4GH, please visit the GA4GH Website.

GA4GH SchemaBlocks Home

schemablocks.org

GA4GH SchemaBlocks Repositories

The SchemaBlocks Github organisation contains several specifically scoped repositories. Please use the relevant Github issues and/or GitHub pull requests to comment and contribute there.

@rebaudo 2019-11-19: more ...

GA4GH SchemaBlocks “Status” Levels

SchemaBlocks schemas (“blocks”) provide recommended blueprints for schema parts to be re-used for the development of code based “products” throughout the GA4GH ecosystem. We propose a labeling system for these schemas, to provide transparency about the level of support those schemas have from [S][B] participants and observers.

@rebaudo 2019-07-17: more ...


SchemaBlocks aims to translate the work of the workstreams into data models that:
- Are usable by other internal GA4GH deliverables, such as the Search API.
- Are usable by Driver Projects as an exchange format.
- Aid in aligning the work streams across GA4GH.
- Do not create a hindrance in development work by other work streams.

@rebaudo 2019-03-27: more ...
**SchemaBlocks Home**

SchemaBlocks is a "cross-workstreams, cross-drivers" initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics.

Launched in December 2018, this project is still to be considered a "community initiative", with developing participation, leadership and governance structures. At its current stage, the documents can not be considered "authoritative GA4GH recommendations" but rather represent documentation and implementation examples provided by GA4GH members.

While future products and implementations may be completely based on SchemaBlocks components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

The SchemaBlocks site can be accessed through the permanent link schemablocks.org. More information about the different products & formats can be found on the workstream sites. For reference, some of the original information about recommended formats and object hierarchies is kept in the GA4GH Metadata repositories.

For more information on GA4GH, please visit the GA4GH Website.

**SchemaBlocks “Status” Levels**

SchemaBlocks schemas ("blocks") provide recommended blueprints for schema parts to be re-used for the development of code based "products" throughout the GA4GH ecosystem. We propose a labeling system for those schemas, to provide transparency about the level of support those schemas have from [5][8] participants and observers.

**Proposed [5][8] Status Levels**
The current status level of those recommendations is "proposed".

- **playground**
  - early development or import stage, of any quality
  - no recommendation; existence does not mean any current or future [5][8] support
- **proposed**
  - at least some [5][8] contributors are in favour of such a block
  - the code may undergo considerable maturation
  - not recommended for integration into products w/o close tracking
  - contributions and discussions are encouraged
- **implemented**
  - mature block which is implemented in one or more [5][8] aligned schemas
  - may be extended from a core block or be too specific for general ("core") usability

**core**
- a schema block with recommended use
- stable through minor version changes
- has to be used in at least 2 standards/products approved by the GA4GH Steering Committee

---

**SchemaBlocks [6][8] Mission Statement**

SchemaBlocks aims to translate the work of the workstreams into data models that:

- Are usable by other internal GA4GH deliverables, such as the Search API.
- Are usable by Driver Projects as an exchange format.
- Aid in aligning the work streams across GA4GH.
- Do not create a hindrance in development work by other work streams.

After discussions with stakeholders from GA4GH work streams and drivers projects who create data models (such as Phenopackets, Search API) or who would use SchemaBlocks for the development of their APIs and data exchange formats (Beacon, EGA, GeL), the SchemaBlocks team has come up with the following principles for this initiative:

**Work Stream Interactions**

Work streams will continue to create standards proposals and their own coherent project implementations, but will work with the SchemaBlocks group to write the Blocks that will come from their own work and are considered of overarching use. Generally, primary work stream and driver project outputs will live in their own spaces outside of SchemaBlocks, with shareable, mature elements - code, documentation, implementation snapshots - being represented in [5][8].
{S}[B] SchemaBlocks Github Repository Structure

blocks repositories

coloration/validation tools

website repository
(Markdown w/ YAML for Github Pages)
Use Case

Transforming Phenopackets objects (here "Age") into JSON Schema documents with (proposed) stable id and address as well as "human readable" documentation & examples.

- Excerpt from Phenopackets v1.0 Schema
- written in Protocol Buffers (Google's data serializing format)
- separate documentation rendered in "ReadTheDocs"
Dissection & Transformation

- Separate [S][B] repository for parental project
- Here “sb-phenopackets”
- Individual schema documents for each original object
- (Currently) manual re-write into JSON Schema documents (YAML version), including metadata header (id, provenance ...)
- Versioned
Dissection & Transformation

- Schema documents are programmatically converted into different outputs
- A versioned JSON document serves as a canonical reference for integration into other products/schemas
Dissection & Transformation

- schema documents are programmatically converted into different outputs
- a Markdown document with "Jekyll" header is auto-converted by Github into a complete website document, including inline code examples
SchemaBlocks **JSON Schema** document format

- {S}[B] "blocks" are written in the YAML version of a JSON Schema document format
  - convenience choice - flexibility, readability, tooling ...
  - **not** implying specific semantics beyond some format conventions - extensible for use-case driven requirements
- the **meta** part (itself defined as a schema "block") contains housekeeping information
  - reference address & version
  - provenance & use cases
  - sb_status about "blessing level"
- the properties part defines the attributes including their description and usage examples
  - descriptions & examples provide the core documentation which is deparsed to the website documents

```json
{
  "title": "AgeRange",
  "description": "Age range",
  "type": "object",
  "meta": {
    "contributors": [
      {
        "description": "Jules Jacobsen",
        "id": "orcid:0000-0002-3265-15918"
      },
      {
        "description": "Peter Robinson",
        "id": "orcid:0000-0002-0736-91998"
      },
      {
        "description": "Michael Baudis",
        "id": "orcid:0000-0002-9993-4248"
      },
      {
        "description": "Isuru Liyanage",
        "id": "orcid:0000-0002-4839-5158"
      }
    ],
    "provenance": {
      "description": "Phenopackets",
      "id": "https://github.com/phenopackets/phenopacket-schema/blob/master/docs/age.rst"
    },
    "used_by": {
      "description": "Phenopackets",
      "id": "https://github.com/phenopackets/phenopacket-schema/blob/master/docs/age.rst"
    },
    "sb_status": "implemented"
  },
  "properties": {
    "start": {
      "allOf": {
        "$ref": "https://schemablocks.org/schemas/ga4gh/v0.0.1/Age.json",
        "description": "Age as ISO8601 string or OntologyClass",
        "examples": [
          {"age": "P12Y"}
        ]
      }
    },
    "end": {
      "allOf": {
        "$ref": "https://schemablocks.org/schemas/ga4gh/v0.0.1/Age.json",
        "description": "Age as ISO8601 string or OntologyClass",
        "examples": [
          {"ageClass": {"id": "HsapDv:0000086", "label": "adolescent stage"}, "age": "P16Y6M"}
        ]
      }
    }
  },
  "required": [
    "start",
    "end"
  ],
  "examples": [
    {"start": {"age": "P12Y"}, "end": {"ageClass": {"id": "HsapDv:0000086", "label": "adolescent stage"}}}
  ]
}
```
BeaconAlleleRequest beacon

[D3 status] Implemented

Provenance
- Beacon-AMT

Used by
- Beacon
- Progenetics database schemas [Bacon+ backend]

Contributors
- Marc Turner
- Michael Buxhues
- Tabela de La Torre Perreas
- jordi zarzosa
- bacon-developers

Source (v1.0.8)
- raw source [GSN]
- Github

Attributes
Type: object
Description: Allele request as interpreted by the beacon.

Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternateNames</td>
<td>string</td>
</tr>
<tr>
<td>assembleBy</td>
<td>string</td>
</tr>
<tr>
<td>datasets</td>
<td>array of string</td>
</tr>
<tr>
<td>end</td>
<td>integer</td>
</tr>
<tr>
<td>endLimit</td>
<td>integer</td>
</tr>
<tr>
<td>externalName</td>
<td>string</td>
</tr>
<tr>
<td>metainfo</td>
<td>string</td>
</tr>
<tr>
<td>references</td>
<td>array of string</td>
</tr>
<tr>
<td>referenceName</td>
<td>string</td>
</tr>
<tr>
<td>start</td>
<td>integer</td>
</tr>
<tr>
<td>startLimit</td>
<td>integer</td>
</tr>
<tr>
<td>variantType</td>
<td>string</td>
</tr>
</tbody>
</table>

Curie sb-vr-spec

[D3 status] Implemented

Provenance
- sb-vr-spec

Used by
- sb-vr-spec

Contributors
- Recce Hart
- Michael Buxhues

Source (v1.0.8)
- raw source [GSN]
- Github

Attributes
Type: object
Description: A Beacon refers to a unit of biological material from which the substrate (e.g. genomic DNA, RNA, protein) for molecular analyses (e.g. sequencing, array hybridisation, mass spectrometry) are extracted. Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencing fraction from a gradient centrifugation. Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well experiments) may refer to the same Biosample.

Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ageOfIndividulCollection</td>
<td>string</td>
</tr>
<tr>
<td>ageRangeOfIndividualCollection</td>
<td>string</td>
</tr>
<tr>
<td>diagnosticMarkers</td>
<td>array of string</td>
</tr>
<tr>
<td>histologicalDiagnosis</td>
<td>string</td>
</tr>
<tr>
<td>isoFiles</td>
<td>array of string</td>
</tr>
<tr>
<td>jtl</td>
<td>string</td>
</tr>
<tr>
<td>jtlIndividul</td>
<td>string</td>
</tr>
<tr>
<td>jtlControlSample</td>
<td>boolean</td>
</tr>
<tr>
<td>nutype</td>
<td>string</td>
</tr>
<tr>
<td>phenotypeFeature</td>
<td>array of string</td>
</tr>
<tr>
<td>procedure</td>
<td>string</td>
</tr>
<tr>
<td>sampledTissue</td>
<td>string</td>
</tr>
</tbody>
</table>

Biosample sb-phenopackets

[D3 status] Implemented

Provenance
- Phenopackets

Used by
- Phenopackets

Contributors
- GA4GH Data Working Group
- Jules Jacobson
- Peter Robinson
- Michael Buxhues
- Mikelare Courtot
- Tony Updage

Source (v1.0.8)
- raw source [GSN]
- Github

Attributes
Type: object
Description: Phenopacket is a unit which describes a phenotype along with any genotype data. The phenotype may include qualitative, quantitative, and time-dependent variables. A Phenopacket may also include a set of diagnostic markers (e.g. SNPs) that are informative of the phenotype, in which case it is a Phenopacket of type Diagnostic. Phenopackets in the Phenopacket Registry may be linked to curated resources or incorporated as evidence in an mdmPublication. Phenopacket metadata is stored in Phenopacket JSON format. Phenopackets may contain associations with other Phenopackets, allowing the creation of phenotype networks. Phenopackets may be linked to other Phenopackets using Phenopacket links. Phenopackets may be linked to other Phenopackets using Phenopacket links.

Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>checkSum</td>
<td>string</td>
</tr>
<tr>
<td>checksum</td>
<td>string</td>
</tr>
</tbody>
</table>

Checksum sb-checksum

[D3 status] implemented

Provenance
- Ga4GH CRG (develop branch)

Used by
- Ga4GH CRG
- Ga4GH TRG

Contributors
- Suhas Varma

Source (v1.0.8)
- raw source [GSN]
- Github

Attributes
Type: object
Description: Checksum

Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>checkValue</td>
<td>string</td>
</tr>
</tbody>
</table>

CheckValue Example

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>checkValue</td>
<td>&quot;sha-256&quot;</td>
</tr>
</tbody>
</table>
SchemaBlocks {S}[B] - Directions & Contributions

- Recognized need of having a set of recommended standards for integrating into product development
  - no need to work through complex standards/projects like FHIR, Phenopackets ...
  - simplification of development

- SchemaBlocks {S}[B] to assume strategic position in GA4GH *TASC system
  - Inclusion into product approval processes?
  - Management/Support?

- Wish for participation of (GA4GH affiliated) groups & individuals, to **expose** their standards & products

- Most important role is the **community aspect**, the interactive exchange of concepts, ideas, code, knowledge, resources ...

- Technical to-dos:
  - Lifecycle: Versioning and representation of donor schemas?
  - Development of conversion workflows for updated source products?
  - Alternative/conflicting blocks...: Graded recommendations? Name spacing?
Leads

- Melanie Courtot
- Michael Baudis

Coordination

- Melissa Konopko

Websites

- schemablocks.org
- github.com/ga4gh-schemablocks/

Meeting minutes

- schemablocks.org/categories/minutes.html