{S][B]
The GA4GH SchemaBlocks Initiative
**SchemaBlocks**

Standardized formats and data schemas for developing an "Internet of Genomics"

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- not a rigid, complete data schema
- object **vocabulary** and **semantics** for a large range of developments
- recognized in **GA4GH roadmap** as possible element in "TASC" effort

schemablocks.org
**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
- Schema documents (.json) can be referenced in other schemas through their $id

---

```json
{
    "title": "AgeRange",
    "description": "Age range",
    "type": "object",

    "meta": {
        "contributors": [
            {
                "description": "Jules Jacobsen",
                "id": "orcid:0000-0002-3265-15918"
            },
            {
                "description": "Peter Robinson",
                "id": "orcid:0000-0002-8736-91998"
            },
            {
                "description": "Michael Baudis",
                "id": "orcid:0000-0002-9803-424B"
            },
            {
                "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": {
                "age": "P18Y"
            }
        }
    ]
}
```
Repositories
From Source to Web

- donor project repositories
  - versioned sources
  - working documents
  - formatted schema "blocks" JSON Schema
  - generated .json, .md

- conversion parser
  - parses the schema documents and extracts JSON, MarkDown documentation
  - current Perl implementation distributes files across local document tree w/ canonical URIs for JSON & HTML
  - per-repository Github synchronisation
  - project for new parser w/ GitHub integration in planning stage at EBI (GSOC proposal)
BeaconAlleleRequest beacon

<table>
<thead>
<tr>
<th>[URI] Status</th>
<th>Implemented</th>
</tr>
</thead>
</table>

Provenance
- Beacon
- Progenetics database schema (Beacon backbone)

Contributors
- Marc Turner
- Michael Buchs
- Nabila de la Torre Pernas
- Joni Karmi
- Beacon-developers...

Source (v1.8.0)
- raw source [GSON]
- Github

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

Properties
- alternateNames: string
- assemblyId: string
- dataStates: array of string
- end: integer
- endIx: integer
- endLin: integer
- metaName: https://schemasألمbra.org/schemas/beacon/v1.1.0/Chrom [HTML]
- referenceDatabases: string
- referenceName: https://schemasألمbra.org/schemas/beacon/v1.1.0/Chrom [HTML]
- start: integer (inclusive)
- startLin: integer
- variantType: string

Curie Value Examples

```
<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>sb-locus</td>
<td>string</td>
</tr>
<tr>
<td>sb-locus</td>
<td>string</td>
</tr>
</tbody>
</table>
```

The loci that appear instead of the reference bases. Accepted values: [ACS]T[ACS]T. A is a wildcard that denotes the position of any base, and can be used as a standalone base of any type or within a partially known sequence. For example a sequence where the first and last bases are known, but the middle portion can exhibit countless variations of [ACS]T, or the bases are unknown. ANNI the loci can take any form of [ACS]T, which makes both ACS and ACS (or any other combinations) viable sequences. Symbols: ACS alleles (DEL, INS, DUP, INV, CNV, DUS/UTM, DELME, INSME) will be represented in variantType. Optional: either alternateDatabases or variantType is required.

Biosample sb-phenopackets

<table>
<thead>
<tr>
<th>[URI] Status</th>
<th>Implemented</th>
</tr>
</thead>
</table>

Provenance
- Phenopackets

Contributors
- GA4GH Data Working Group
- Jules Jacobson
- Peter Robinson
- Michael Baudis
- Melanie Courtot
- Travis Upgarage

Source (v1.8.0)
- raw source [GSON]
- Github

Attributes
- Type: object
- Description: A Biosample refers to a unit of biological material from which the substrate molecule (genomic DNA, RNA, proteins) for molecular analyses (i.e., sequencing, array hybridisation, metagenomics) are extracted. Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencing, or a 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.

FHIR mapping: Specimen

Properties
- Property     | Type     |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ageOfIndividualCollection</td>
<td>string</td>
</tr>
</tbody>
</table>

Curie Value Examples

```
<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>sb-phenopacket</td>
<td>string</td>
</tr>
</tbody>
</table>
```

CheckSum sb-checksum

<table>
<thead>
<tr>
<th>[URI] Status</th>
<th>Implemented</th>
</tr>
</thead>
</table>

Provenance
- GaGaNG CMS (develop branch)

Contributors
- GA4GH CMS
- GA4GH TRG

Source (v1.8.1)
- raw source [GSON]

Attributes
- Type: object
- Description: Checksum

Properties
- Property     | Type     |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>checksum</td>
<td>string</td>
</tr>
</tbody>
</table>

Curie Value Examples

```
<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>sb-checksum</td>
<td>string</td>
</tr>
</tbody>
</table>
```

checksum

```
<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD5</td>
<td>string</td>
</tr>
</tbody>
</table>
```

The digest method used to create the checksum. The value (e.g., "md5-256") SHOULD be listed as Hash. The md5-256 in the GA4GH Hash Algorithm Registry. Other values may be used, as long as Implementors are aware of the issues discussed in RFC3194.

GA4GH may provide even more explicit guidance for use of non-256-bit algorithms in the future.

Curie Value Examples

```
<table>
<thead>
<tr>
<th>Term</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;md-256&quot;</td>
<td>string</td>
</tr>
</tbody>
</table>
```
{S}[B] & TASC
Managing Project Alignment

• donor project repositories
  ‣ different structures, technologies
  ‣ donors, recipients or both
  ‣ to do: dedicated technical contact

• TASC
  ‣ encouraging project exchange
  ‣ using product review process to propose, request schema donations, alignment
  ‣ reviewing documentation

• {S}[B] members
  ‣ maintaining repository structure
  ‣ tool development
{S}[B] and TASC
Technical Alignment through Documentation & Distribution

- SchemaBlocks is well suited for driving the **exchange** of standards, code, procedures, data schemas in the heterogeneous GA4GH ecosystem.
- There is a large amount of forward projecting "this will be represented as/in SchemaBlocks" throughout GA4GH workstreams and projects (Beacon, Discovery Search, DUO...).
- While the initiative is driven by the need for an alignment of general standards and principles favoured by GA4GH participants, so far it consists of voluntary contributions w/o embedding in GA4GH administrative procedures, or dedicated project support (exceptions: SPHN, EBI).
- A **lightweight managed process** through TASC (e.g. encouraging, requesting exchange through {S}[B] in product review, driver projects) would have a high impact on the cohesion and common recognition of "GA4GH standards".
- Such a process can **co-exist** with tightly controlled schema developments for subsets of the GA4GH ecosystem, if intended.
Leads
  • Melanie Courtot
  • Michael Baudis

Coordination
  • Melissa Konopko
  • Rishi Nag

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

Meeting minutes
  • schemablocks.org/categories/minutes.html