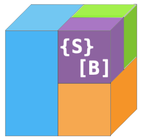


{S}[B]

The GA4GH SchemaBlocks Initiative

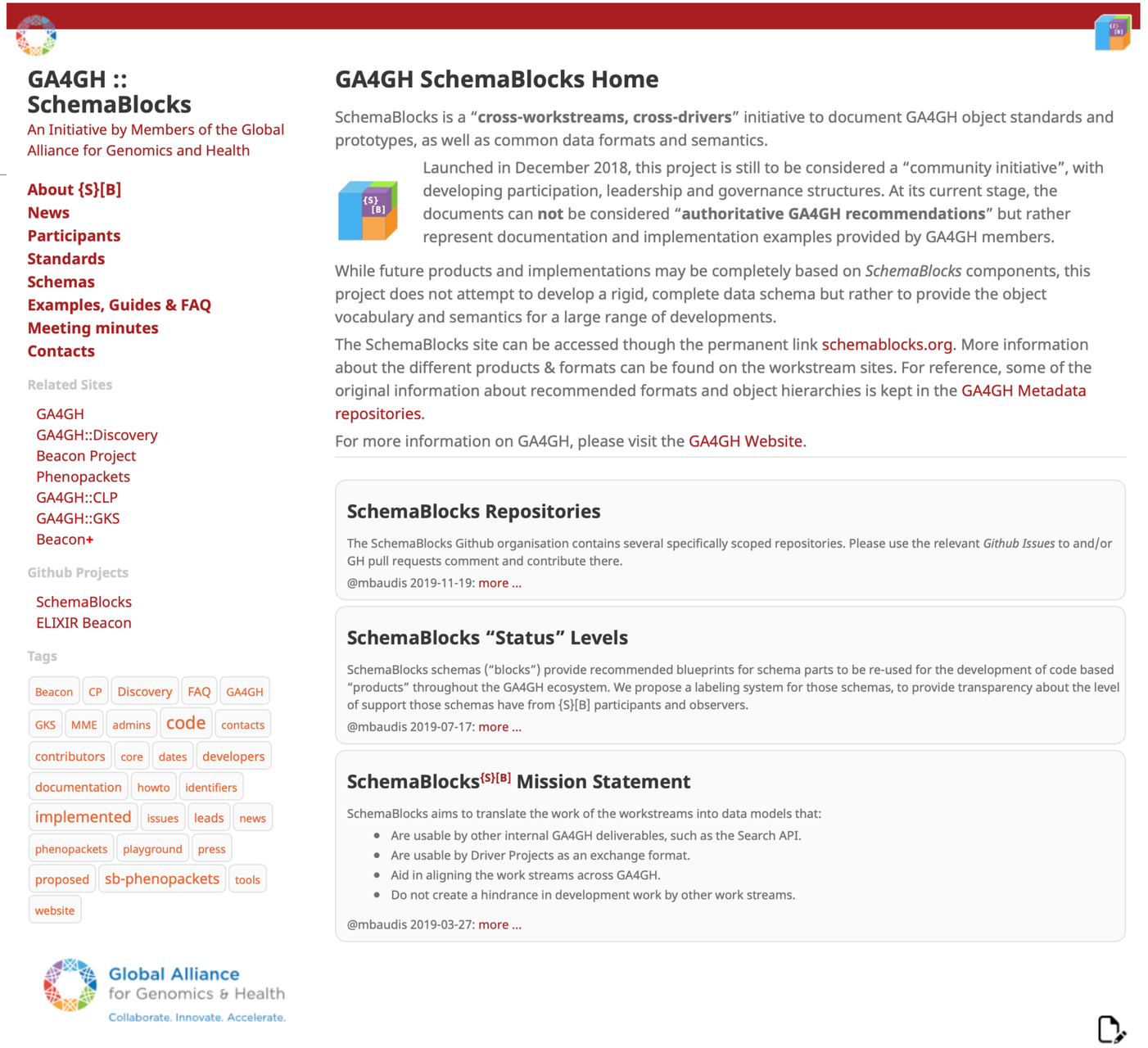
Michael Baudis | UZH & SIB | 2020-04-20

GA4GH {S}[B] *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



GA4GH :: SchemaBlocks

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

About {S}[B]
News
Participants
Standards
Schemas
Examples, Guides & FAQ
Meeting minutes
Contacts

Related Sites

GA4GH
GA4GH::Discovery
Beacon Project
Phenopackets
GA4GH::CLP
GA4GH::GKS
Beacon+

Github Projects

SchemaBlocks
ELIXIR Beacon

Tags

Beacon CP Discovery FAQ GA4GH GKS MME admins code contacts contributors core dates developers documentation howto identifiers implemented issues leads news phenopackets playground press proposed sb-phenopackets tools website

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

GA4GH 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 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.

SchemaBlocks Repositories

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

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

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 {S}[B] participants and observers.

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

SchemaBlocks{S}[B] 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.

@mbaudis 2019-03-27: more ...

{S}[B] SchemaBlocks ***JSON Schema*** document format

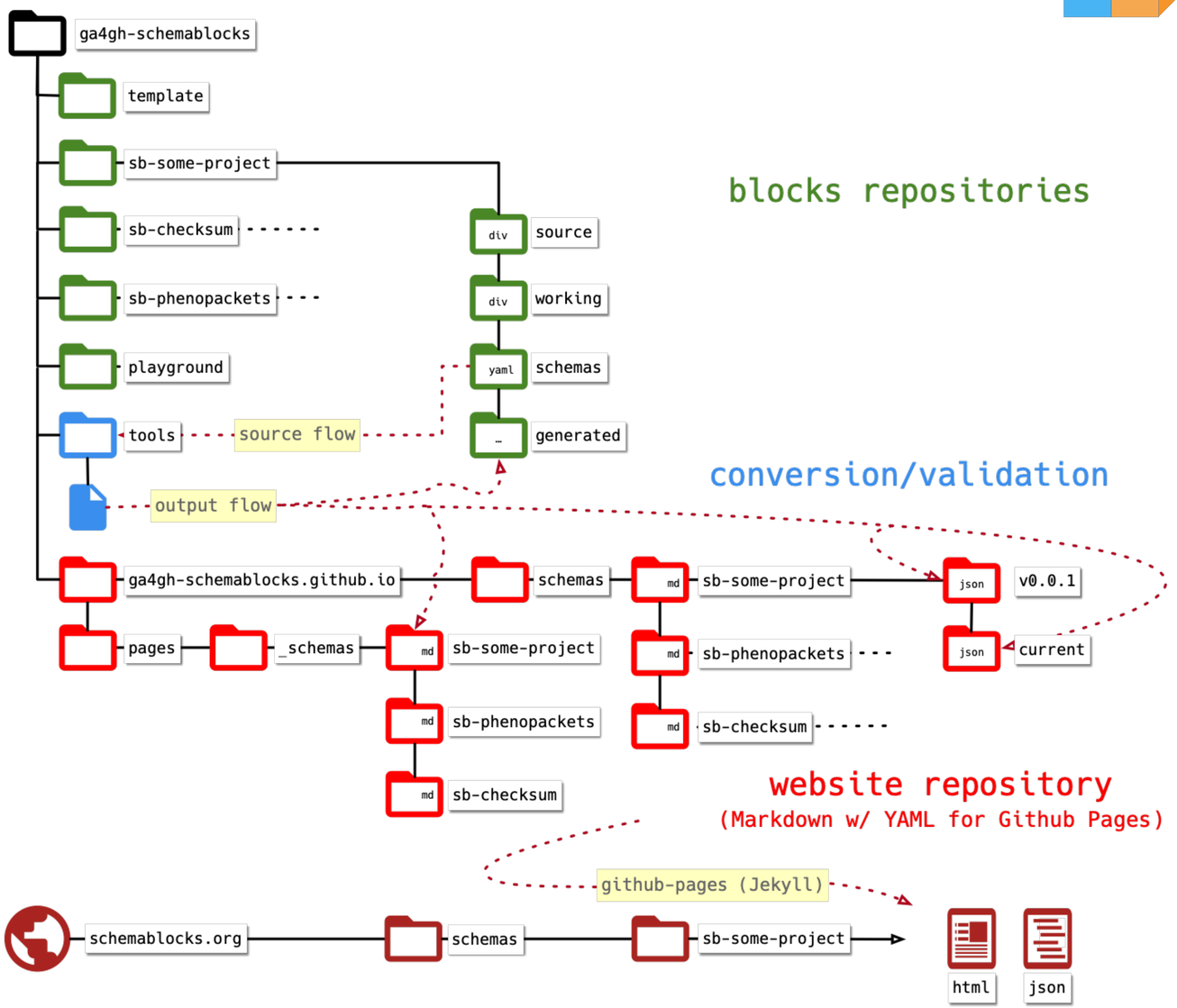
- {S}[B] "blocks" are written in the YAML version of a JSON Schema document format
  - convenience choice - flexibility, readability, tooling ...
  - ***not*** implying specific semantics beyond some format conventions - extensible for use-case driven requirements
- the meta part (itself defined as a 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*

```
"$schema": http://json-schema.org/draft-07/schema#
"$id": https://schemablocks.org/schemas/ga4gh/AgeRange/v0.0.1
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:0000-0002-9903-4248"
    - 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:
  anyof:
    - start
    - end
examples:
  - start:
      age: 'P12Y'
      ageClass:
        id: 'HsapDv:0000086'
        label: 'adolescent stage'
    end:
      age: 'P18Y'
```

{S}[B] 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 ↗

{S}[B] Status [i]	implemented
Provenance	○ Beacon API
Used by	○ Beacon ○ Progenetix database schema (Beacon+ backend)
Contributors	○ Marc Fiume ○ Michael Baudis ○ Sabela de la Torre Pernas ○ Jordi Rambla ○ Beacon developers...
Source (v1.1.0)	○ raw source [JSON] ○ Github

Attributes

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

Properties

Property	Type
alternateBases	string
assemblyId	string
datasetIds	array of string
end	integer
endMax	integer
endMin	integer
mateName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
referenceBases	string
referenceName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
start	integer (int64)
startMax	integer
startMin	integer
variantType	string

alternateBases

- type: string

The bases that appear instead of the reference bases. Accepted values: [ACGTN]*. N 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 [ACGT], or the bases are unknown: ANNT the Ns can take take any form of [ACGT], which makes both ACCT and ATGT (or any other combination) viable sequences.

Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be represented in `variantType`.

Optional: either `alternateBases` or `variantType` is required.

`alternateBases` **Value Example**

assemblyId

- type: string

Assembly identifier (GRC notation, e.g. GRCh37).

`assemblyId` **Value Example**

Curie sb-vr-spec ↗

{S}[B] Status [i]	implemented
Provenance	○ vr-spec
Used by	○ vr-spec
Contributors	○ Reece Hart ○ Michael Baudis
Source (v1.0)	○ raw source [JSON] ○ Github

Attributes

Type: string
Pattern: ^\w[^:]+:.+$
Description: A string that refers to an object uniquely. T sender.

VR does not impose any contraints on strings used as ids data, the VR Specification RECOMMENDS that implement

String CURIEs are represented as `prefix:reference` (W `namespace:accession` or `namespace:local id` colloqu

The VR specification also RECOMMENDS that `prefix` be

The `reference` component is an unconstrained string.

A CURIE is a URI. URIs may *locate* objects (i.e., specify wh VR uses CURIEs primarily as a naming mechanism.

Implementations MAY provide CURIE resolution mechan

Using internal ids in public messages is strongly discoura

`Curie` **Value Examples**

```
"ga4gh:GA.01234abcde"
```

```
"DUO:0000004"
```

```
"orcid:0000-0003-3463-0775"
```

```
"PMID:15254584"
```

Biosample sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	○ Phenopackets
Used by	○ Phenopackets
Contributors	○ GA4GH Data Working Group ○ Jules Jacobsen ○ Peter Robinson ○ Michael Baudis ○ Melanie Courtot ○ Isuru Liyanage
Source (v1.0.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: A Biosample refers to a unit of biological material from which the substrate molec genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hybridisation, mas spectrometry) are extracted.

Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencii 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
ageOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Age.json [SRC] [HTML]
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [SRC] [HTML]
description	string
diagnosticMarkers	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
htsFiles	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/HtsFile.json [SRC] [HTML]
id	string
individualId	string
isControlSample	boolean
phenotypicFeature	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/PhenotypicFeature.json [SRC] [HTML]
procedure	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Procedure.json [SRC] [HTML]
sampledTissue	https://schemablocks.org/schemas/sb-

Checksum sb-checksum ↗

{S}[B] Status [i]	proposed
Provenance	○ GA4GH DRS (`develop` branch)
Used by	○ GA4GH DRS ○ GA4GH TRS
Contributors	○ Susheel Varma
Source (v0.0.1)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: Checksum

Properties

Property	Type
checksum	string
type	string

checksum

- type: string

The hexadecimal encoded (Base16) checksum for the data

`checksum` **Value Example**

```
"77af4d6b9913e693e8d0b4b294fa62ade6054e6b2f1ffb617ac955dd63fb0182"
```

type

- type: string

The digest method used to create the checksum. The value (e.g. `sha-256`) SHOULD be listed as `Hash Name String` in the GA4GH Hash Algorithm Registry. Other values MAY be used, as long as implementors are aware of the issues discussed in RFC6920.

GA4GH may provide more explicit guidance for use of non-IANA-registered algorithms in the future.

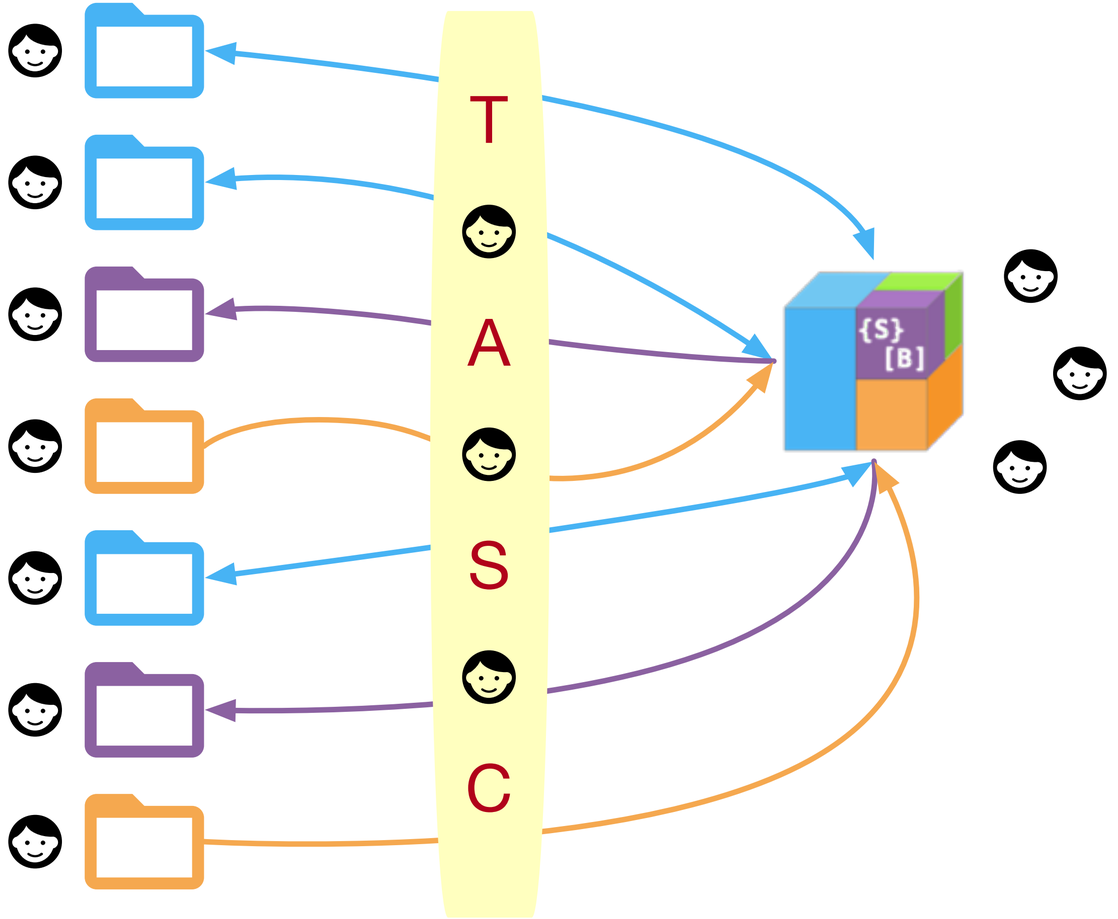
`type` **Value Example**

```
"sha-256"
```

{S}[B] & TASC

Managing Project Alignment

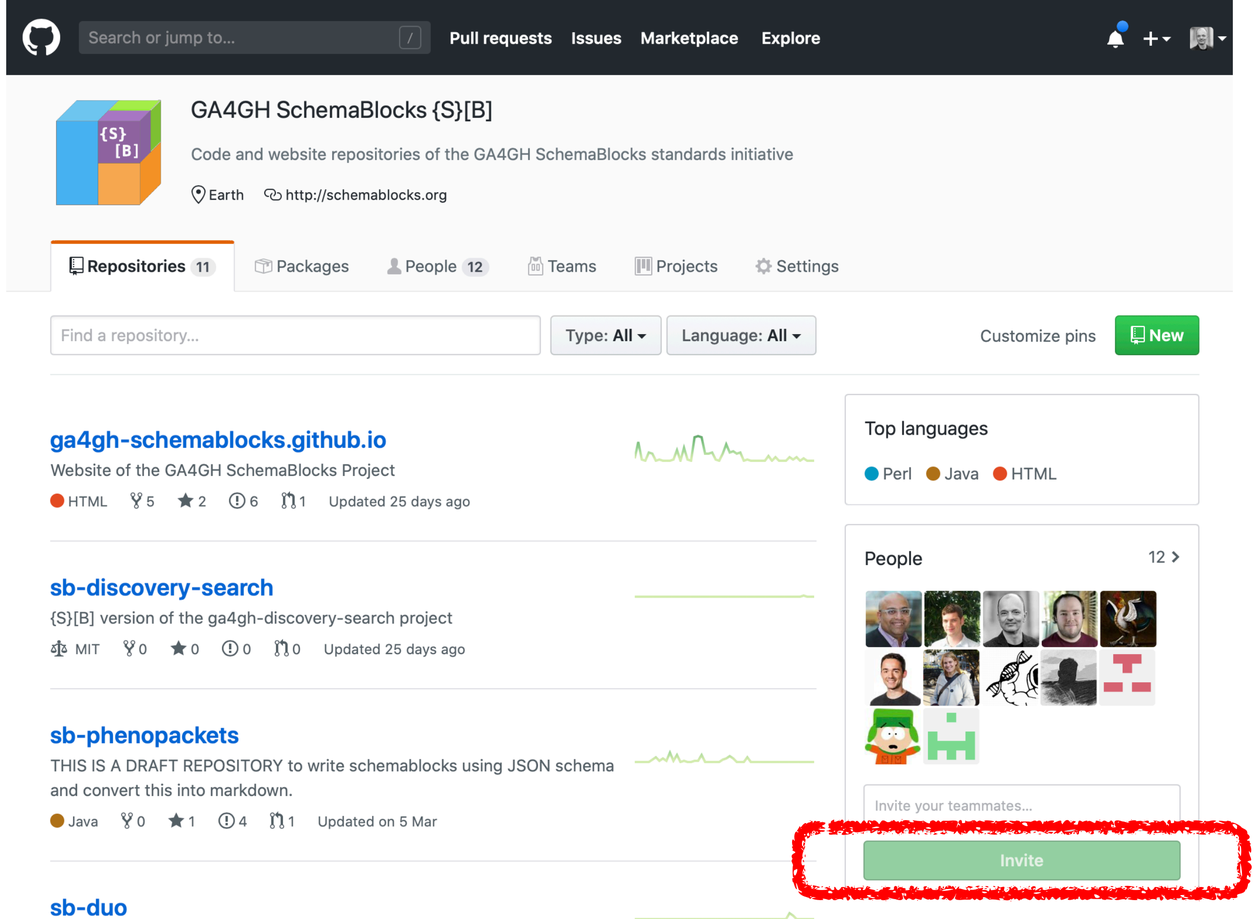
- donor project repositories
  - different structures, technologies
  - donors, recipients or both
  - to do: dedicated technical contact
- TASC
  - encouraging project exchange
  - using product review process to propose, request schema donations, alignment
  - reviewing documentation
- {S}[B] members
  - 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.



{S}[B] Info

Leads

- Melanie Courtot [↗]
- Michael Baudis [↗]

Coordination

- Melissa Konopko
- Rishi Nag

Websites

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

Meeting minutes

- schemablocks.org/categories/minutes.html