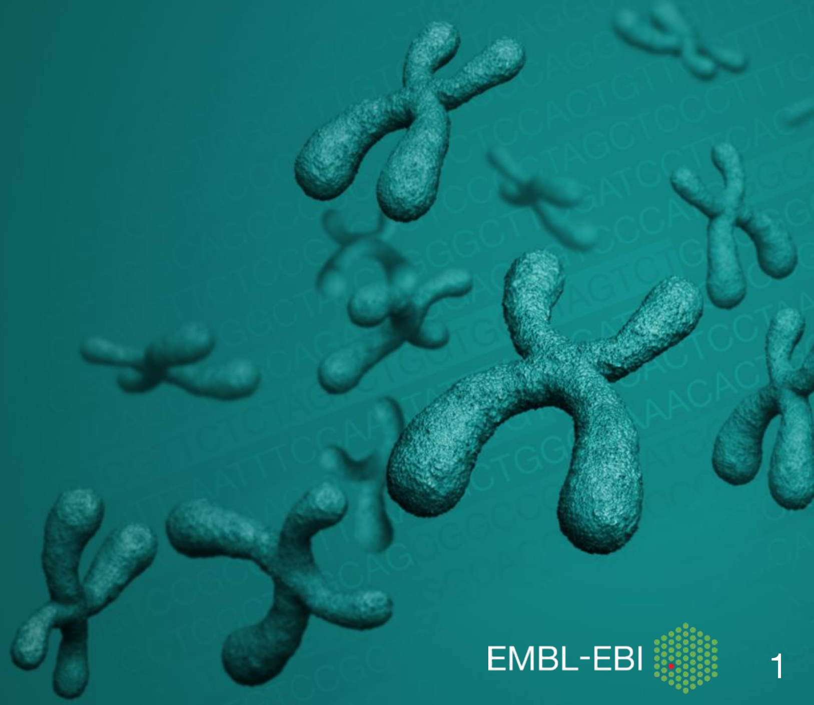
SchemaBlocks Use Case

Phenopackets



Isuru Liyanage
isuru@ebi.ac.uk

EMBL-EBI

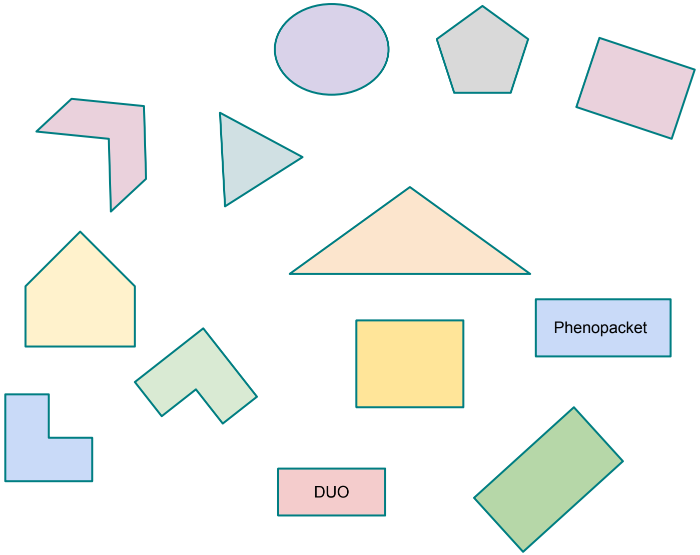
Importance of GA4GH Common Data Models

GA4GH needs a platform to disseminate, expose, increase visibility and enable shared development

Place in GA4GH ecosystem to provide

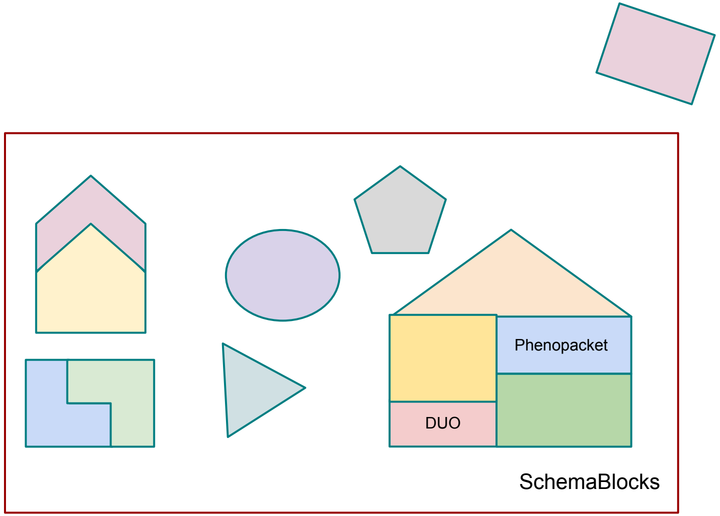
- Data models
- Standard recommendations

And while doing so we need to make sure it does not slow the development process



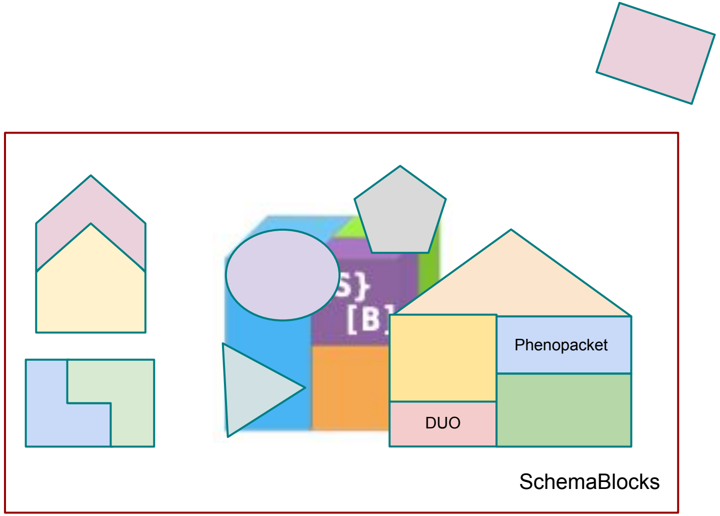
SchemaBlocks

- Cross-workstreams, cross-drivers initiative
- Document GA4GH object standards and prototypes
  - common data formats and semantics
- Catalog of models
  - a place to search for
- Plug and play modules
- Recommendation in product approval
- Transparency, exposure and visibility
  - people from all WS involved



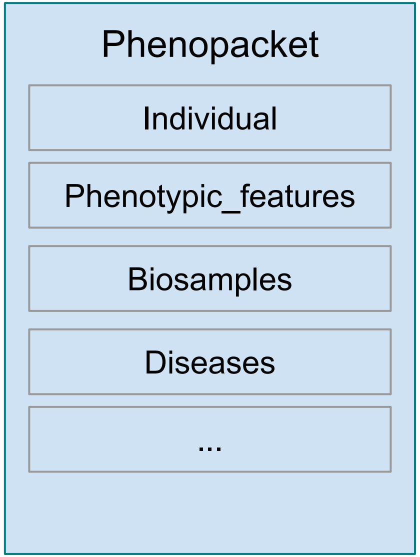
Inside SchemaBlocks

- Expressed in JSON schema
  - Expressiveness
  - Extensibility
  - Validation using standard tooling
- Development
  - Source in YAML format
    - Facilitates human read/edit
  - Generate JSON and documentation



Use case: Phenopackets

- Open standard for sharing disease and phenotype information
- GA4GH (almost-)approved product
- Modular - consists of several messages
- Implemented using protobuf
  - Generate code for many languages, fast
  - Once defined, easy to use
- Use generated library
  - Function to generate JSON output

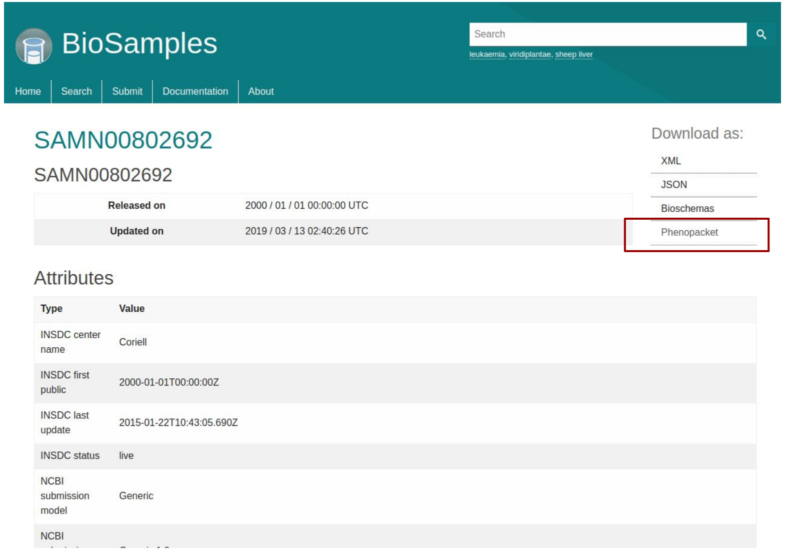


EMBL-EBI BioSamples and Phenopackets

- Export EMBL-EBI BioSamples data to phenopacket
  - Download from web
  - Define content-type to direct download
- Phenopacket version 1.0.0-RC2

```
Content-type: "application/phenopacket+json"
```

```
https://www.ebi.ac.uk/biosamples/samples/SAMN00802692.pxf
```



BioSamples

SAMN00802692

SAMN00802692

Released on	2000 / 01 / 01 00:00:00 UTC
Updated on	2019 / 03 / 13 02:40:26 UTC

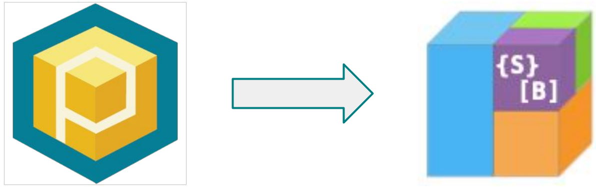
Download as:

Attributes

Type	Value
INSDC center name	Coriell
INSDC first public	2000-01-01T00:00:00Z
INSDC last update	2015-01-22T10:43:05.690Z
INSDC status	live
NCBI submission model	Generic
NCBI	

Phenopackets to SchemaBlocks

- Manual conversion from Phenopackets to SchemaBlocks
  - PXF uses Google's Protocol Buffers schema description format
    - Efficient for message serialization & good tooling
    - Limited expressibility and flexibility
  - Protobuf to JSON schema w/o dedicated tools
- Once product is stable easy to convert
  - If there are active changes hard to keep in sync
  - Tooling possible, but judgement call (repeated use?)

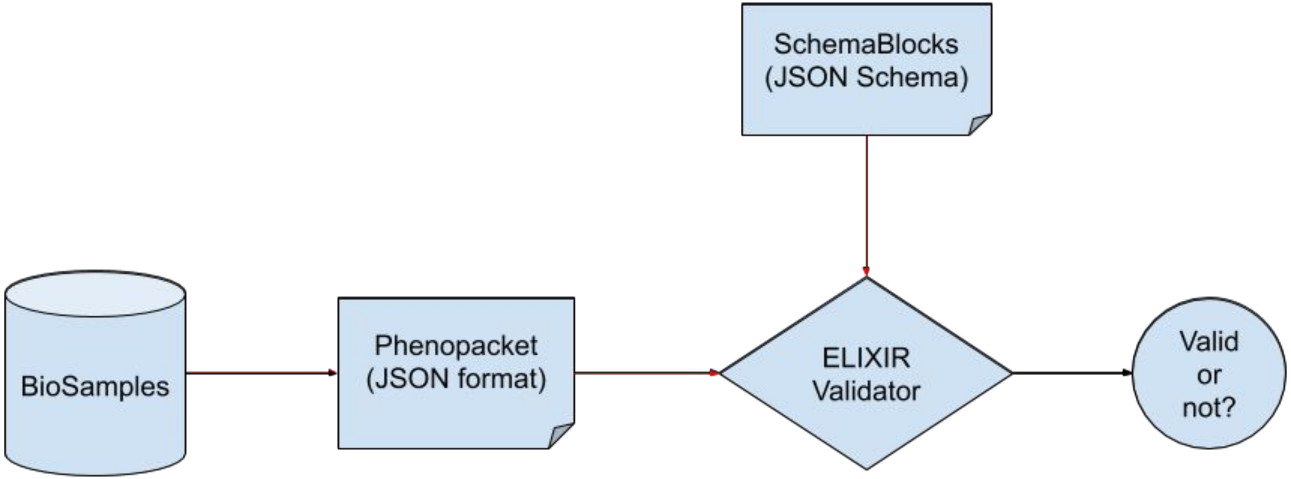


JSON Schema and Validation

- Validate JSON data using schema
- Many implementations of schema validators
- ELIXIR JSON schema validator
    - Strategic partner
    - Easy to run as a separate server
    - Custom extensions of life science data
    - Already used in driver projects eg. HCA

All Put Together

- Export samples into phenopacket (JSON format)
  - More than 11M samples in BioSamples
- Validate using ELIXIR validator
  - Against SchemaBlocks schema



Conclusion

- What next
  - Work with DURI and REWS
  - Adaptors - Beacon
  - Place in product development and approval process

Language independent consistent representation throughout GA4GH products

THANK YOU

Melanie Courtot
Michael Baudis
Ben Hutton
Jules Jacobsen

Phenopackets
ELIXIR
GSoC
HCA
WellcomeTrust-EBI grant 201535/Z/16Z

Links

https://schemablocks.org/

https://github.com/ga4gh-schemablocks/

https://github.com/ga4gh-schemablocks/sb-phenopackets

https://schemablocks.org/schemas/ga4gh/v0.0.1/Age.json

https://phenopackets-schema.readthedocs.io/en/latest/

https://github.com/phenopackets/phenopacket-schema

https://www.ebi.ac.uk/biosamples

https://www.ebi.ac.uk/biosamples/samples/SAMN00802692.pxf

https://github.com/elixir-europe/json-schema-validator