This version:

https://docs.google.com/document/d/1ExjXV5Y42SwH18cV91C9NbWkHooNjPpwL960lbT6eZ4/edit#

Editors:

Chris Mungall

Marcin Joachimiak

Status: **very early draft/outline**

Abstract

This document is intended as a starting point for standardizing and sharing design patterns around the representation of samples and their associated metadata and characteristics. The scope is all kinds of physical samples, including:

- Biomedical, such as tissue or blood samples from humans
- Environmental, including samples analyzes for metagenomics/metaomics
- Geological samples
- Samples collected from field sites as well as laboratories/hospitals, as well as mesoscale experimental setups

We propose to use the W3C RDF data model as the underlying representation, as this allows for interoperation and extensibility, and has a convenient serialization in JSON-LD. This also allows us to re-use URIs from standard ontologies, e.g. OBO ontologies such as OBI, ENVO and BCO, as well as adopted W3C ontologies such as SSNO.

The overall goals are:

- Support a wide variety of use cases for tracking, data integration, search, analysis, comparison, discovery
- Easy queryability in SPARQL (e.g avoiding blank nodes or artificial intermediate IRIs where possible) or graph databases
- Easy to index JSON-LD structures using document stores or indexes such as Elastic Search
- Extensibility to support domain-specific use cases through hierarchies of ‘packages’ for different domains, e.g. using MIxS packages. This allows for a combination of fine-grained specification and extensibility.
- Leverage the work done in multiple communities that may not have been aware of each other
- Reuse ontologies, especially OBOs, for fine-grained description of sample and source characteristics

Introduction

The gathering and analysis of material samples is key to many areas of science. In biomedicine, samples are collected from humans in biobanks and are analyzed and characterized, e.g. genome, transcriptome, metabolome. Understanding the microbiome requires gathering samples from either the external environment or organismal hosts and characterizing e.g. chemically and with -omics. In biodiversity science, material samples can take the form of whole organisms, parts of organisms, or portions of environmental materials (i.e., soil, water, air) that contain several organisms. These samples can be subjected to a wide array of analyses to measure physical, chemical, and genetic properties, some of which are destructive. The metadata that accompanies the collecting event as well as information about the storage and preservation of the sample, and the analytical method are very important for attribution, provenance, and reproducibility. TODO. Paleontology TODO. Earth science and geology TODO.

Different communities have arrived at different schemas, metadata templates, ontologies and information systems for recording sample descriptions. These communities may not always be aware of each other. Coming up with a single overarching schema is hard, due to the variety of use cases, the different kinds of things sampled (rocks, seawater, tissue), and different communities focus of interest (e.g. for some, representation of the process of collecting is paramount, for others, the characteristics of the sample itself are the only important thing; furthermore a soil scientist may care about a very specific set of parameters such as porosity that may not make sense or be as important for other kinds of sample).

We propose the use of an RDF-based datamodel in order to provide a flexible framework for these different use cases, arranged around a common backbone. This allows different

communities to extend for their own purposes, and for the different schemas to be mixed and matched.

The RDF datamodel consists of *triples* organized into a *named graph*. Each triple consists of a subject, predict, and object, which can be read as a sentence. For example, *sample1 type 'lung tissue sample', sample1 derivedFrom sample2*. Each position in the triple can be either (a) a URI (essentially a URL) that denotes some entity or (b) a literal, e.g. a string. [this is a simplification that works for our purposes here]. For convenience, URIs can be written in short "CURIE" form, e.g. ENVO:00012345.(McMurry et al. 2017)

For example:

- mydatabase:Sample1 derivedFrom mydatabase:Organism1 .
- mydatabase:Sample1 rdfs:label "my sample" .
- mydatabase:Sample1 rdf:type obi:Sample .

Each URI in the triple may represent a particular entity such as an actual physical entity, or a descriptor from an ontology. Ontologies can be expressed using the OWL language.

An example of a use of RDF is in the PubChem resource
https://pubchemdocs.ncbi.nlm.nih.gov/rdf$_1-1

Many other communities have adopted RDF as the core datamodel for representing sample data, e.g. the semantic sensor net community.

One advantage of RDF is that it simplifies data integration. Sets of triples can be merged together. The two datasets can partially or fully agree on an ontology, making integration even easier.

RDF datasets can easily be queried via the SPARQL language. There are a number of triplestore databases that can be used for managing RDF, with provision of SPARQL endpoints that can be queried over http. Many of these allow federated queries - e.g. a sample triplestore could be federated with a database of geolocation metadata, allowing queries for "any samples collected in areas of high rainfall". RDF can also be easily managed and queried as simple files.

JSON-LD allows RDF graphs to be serialized in JSON form, with an unambiguous automated transform to RDF. By using JSON-LD context files, the JSON can be made more 'developer-friendly'

Survey of Existing Work

In progress...

Schema	Scope	Schema Spec Language	Ontologies directly used	Structure	Serializations
EBI-BioSample RDF	Biosamples (organism tissue and metagenome)	RDFS		flat	
Biosamples		RDFS-like		flat	JSON-LD
Cancer GDC	Cancer samples	Custom		graph	
SSNO		RDFS/OWL	SSNO, PROV,	graph	RDF
TURBO		OWL	OBI	graph	RDF/OWL
GenSC		Excel	ENVO, GAZ, DO, ...	flat	Property-Value Tuples
DATS		JSON Schema	OBI	graph	JSON-LD
IGSN				flat?	
ENCODE (https://www.encodeproject.org/help/data-organization/)					

Representation of samples in BioSchemas

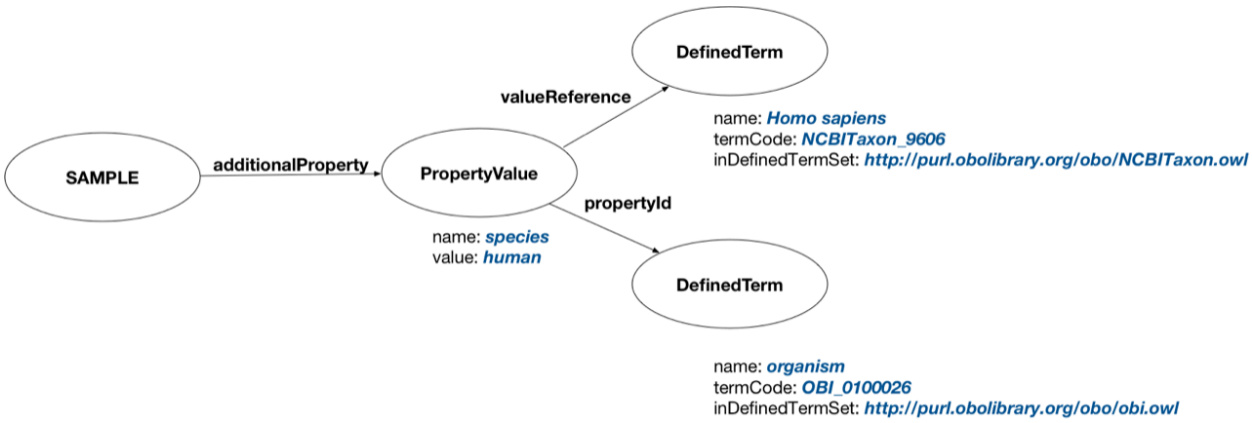
Use cases/scope:
https://bioschemas.org/useCases/Samples/
The listed use cases are mainly **biobanking**; metagenomic or environmental samples are not explicitly listed, but not explicitly ruled out. EBI biosample people are involved so presumably would cover anything here.

Specification:
https://bioschemas.org/specifications/Sample/
https://bioschemas.org/groups/Samples/

The specification is fairly minimal. The core fields are id, type, url, description [text]. Samples are described using an extension of the generic schema.org property-value model. For example, to record the species a sample was derived from a property may be “organism” and a value may be “Homo sapiens”, with ontologies used for both.

Notes:
Bioschema.org itself does not dictate what properties should be used or what ontologies should be used. This permits a lot of freedom and extensibility. However, it could potentially be an interoperability problem, as different groups will use different vocabularies or represent the same thing in different ways.

Example:



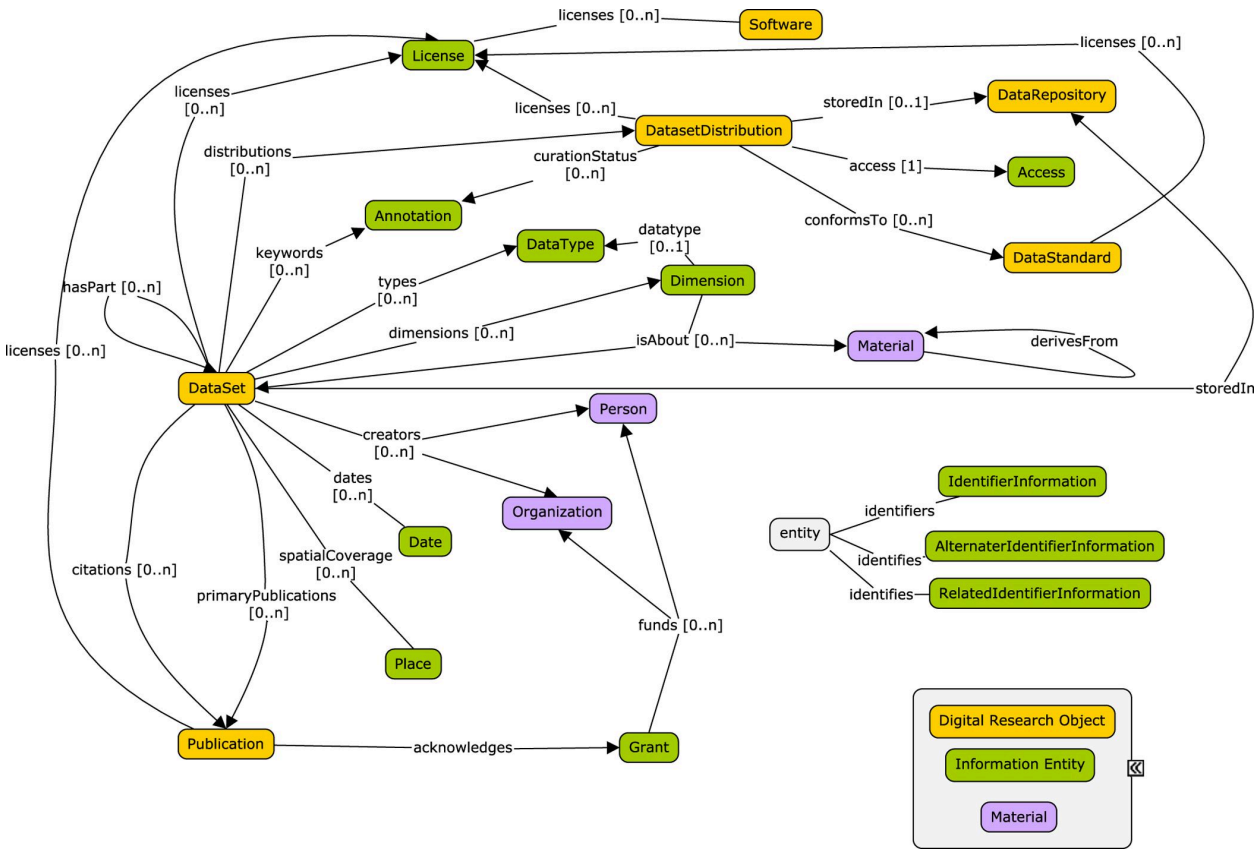
Representation of samples in DATS

Use cases/scope:
DATS is used for indexing of data to enhance search.

Primarily biomedical?

Publication:
https://www.nature.com/articles/sdata201759
Main docs: https://datatagsuite.github.io/docs/html/dats.html

Specification:
These are called ‘material’ here:
https://github.com/datatagsuite/schema/blob/master/material_schema.json
Materials can be connected via the derivesFrom relation. Materials can be described using a generic data Dimension vector. Schemas can be combined to describe other aspects.



The schema for ‘material’ permits identifier, description altIdentifiers and taxon (e.g human). It also allows generic extensible descriptions via roles and characteristics, akin to bioschemas.

Ontologies used:
DATS allows for two different JSON-LD context “profiles”

- Schema.org vocabulary
- OBO vocabulary

Example:

The following JSON-LD shows DNA material extracted from a blood sample/specimen which comes from a patient:

https://github.com/datatagsuite/examples/blob/9654f4a7351ee79ad128d28ceaf371fec4de45a5/datacommons/topmed.json#L138-L396

Notes:
Note the high level of granularity permitted. Rather than collapsing everything to table of values centered around a sample, each distinct entity (DNA, blood, patient) is tracked.

Representation of samples in Schemablocks

Use cases/scope:
?biomedical only?

Specification:
https://schemablocks.org/schemas/sb-phenopackets/Biosample.html

Notes:

The biosample schema is designed to be used in conjunction with the other ga4gh schemas, e.g. it has foreign keys to other ga4gh schemas in schemablocks, e.g project_id, individual_id.

The geo_provenance field states:

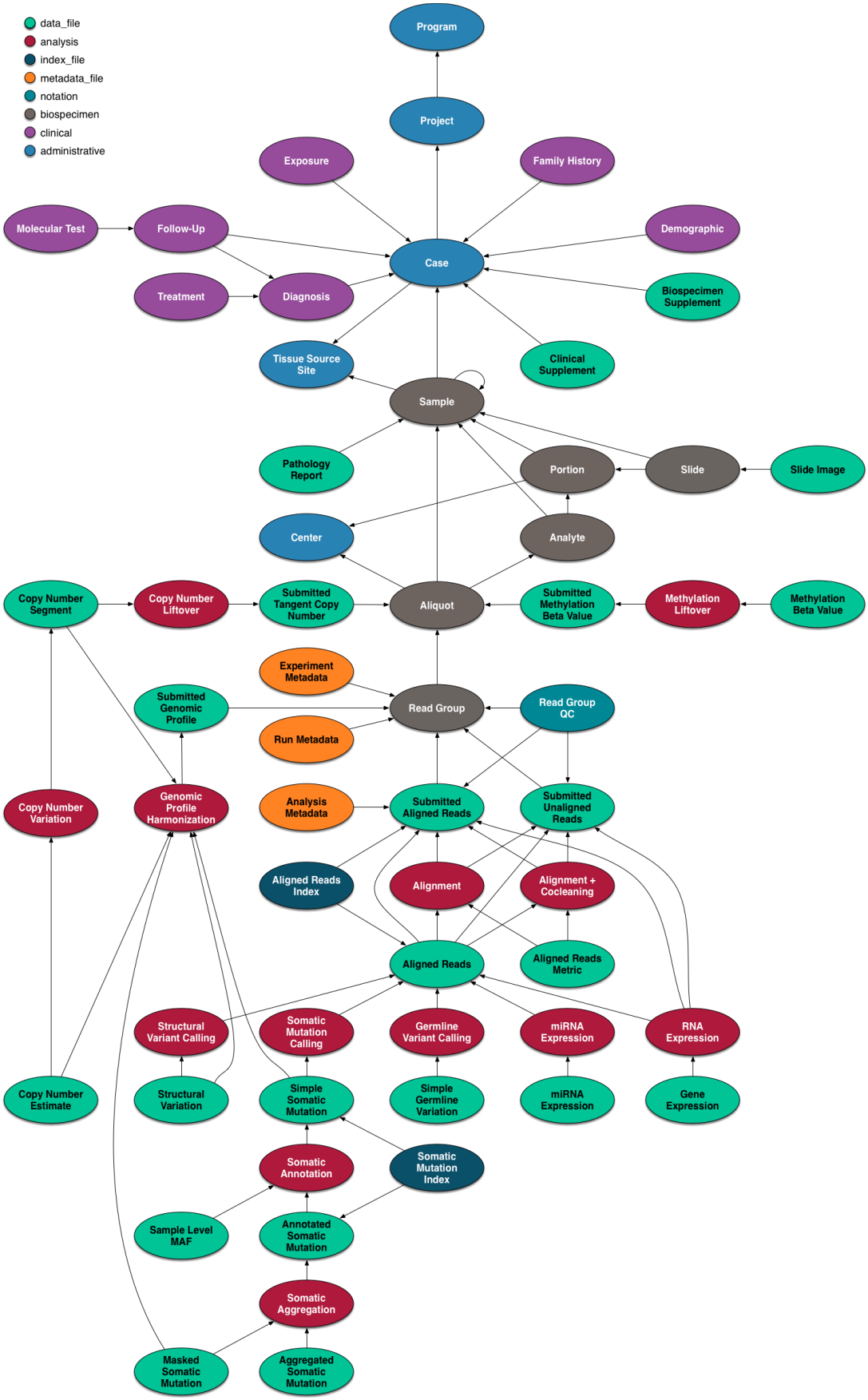
> *This geo_class attribute ideally describes the geographic location of where the sample was extracted. Frequently, this value may reflect either the place of the laboratory where the analysis was performed, or correspond to the corresponding author's institution*

Note this level of ambiguity should not be permitted for environmental samples as it will be impossible to automatically disentangle the important environmental source from the less important location of lab/storage.,

Representation of samples in Cancer Genomic Data Commons

Use cases/scope:
Cancer samples

Specification:
https://gdc.cancer.gov/developers/gdc-data-model/gdc-data-model-components


data_file
analysis
index_file
metadata_file
notation
biospecimen
clinical
administrative
Program
Project
Exposure
Family History
Molecular Test
Follow-Up
Demographic
Case
Treatment
Diagnosis
Biospecimen Supplement
Tissue Source Site
Clinical Supplement
Sample
Pathology Report
Portion
Slide
Slide Image
Center
Analyte
Copy Number Segment
Copy Number Liftover
Submitted Tangent Copy Number
Aliquot
Submitted Methylation Beta Value
Methylation Liftover
Methylation Beta Value
Experiment Metadata
Submitted Genomic Profile
Read Group
Read Group QC
Run Metadata
Copy Number Variation
Genomic Profile Harmonization
Analysis Metadata
Submitted Aligned Reads
Submitted Unaligned Reads
Aligned Reads Index
Alignment
Alignment + Cocleaning
Aligned Reads
Aligned Reads Metric
Structural Variant Calling
Somatic Mutation Calling
Germline Variant Calling
miRNA Expression
RNA Expression
Copy Number Estimate
Structural Variation
Simple Somatic Mutation
Simple Germline Variation
miRNA Expression
Gene Expression
Somatic Annotation
Somatic Mutation Index
Sample Level MAF
Annotated Somatic Mutation
Somatic Aggregation
Masked Somatic Mutation
Aggregated Somatic Mutation

https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-definition-view&id=sample
Properties of a sample:

- Type; e.g. benign neoplasm
- Tissue
- Anatomic site (e.g lungs)
- Laterality (e.g left)
- Composition (e.g. 3D organoid)
- Weight
- Days to collection
- Distance to tumor
- Tumor code
- [various others]

Examples:
…

Representation of biosamples in EBI RDF triplestore

Note the docs here pertain to the existing EBI triplestore. At the 2017 Biohackathon there were plans to align EBI and DDBJ around a bioschemas-like representation, this is documented here:

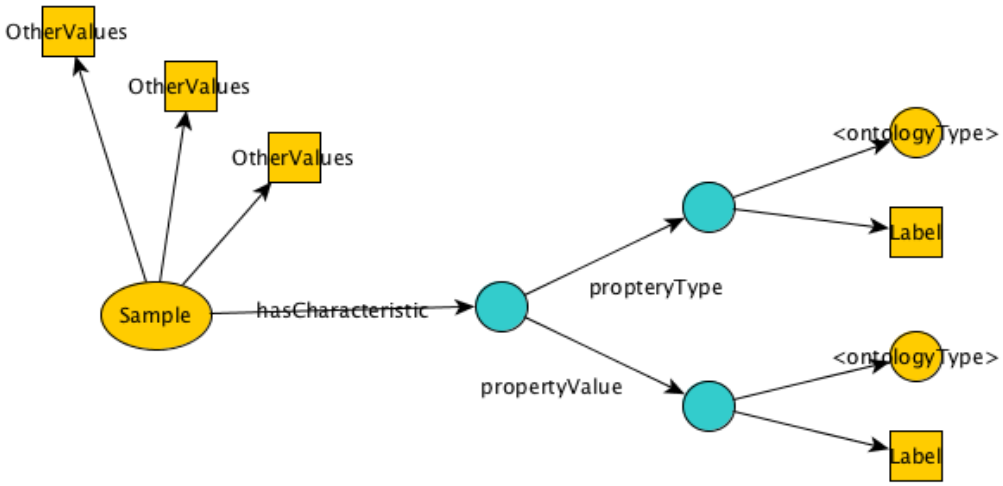
- *https://docs.google.com/document/d/15NN1I2bl9Zs_wcd0hgUFse3KMACPaBgeOhaBe8Gqe4s/edit*
- https://www.slideshare.net/mbrandizi/biosd-tutorial-2014-edititon
- https://github.com/EBIBioSamples/biosd2rdf/blob/master/src/main/assembly/resources/rdf/biosd_terms.ttl

At this time this new schema has not been implemented?
Use cases/scope:
Same as EBI BioSamples

Specification:
RDF represented here: https://www.ebi.ac.uk/rdf/documentation/biosamples/

Some screenshots of the model can be found in https://prezi.com/vxox0pgra6d7/biosd-linked-data-lessons-learned/ but these are quite hard to read.

I found another piece of documentation here:

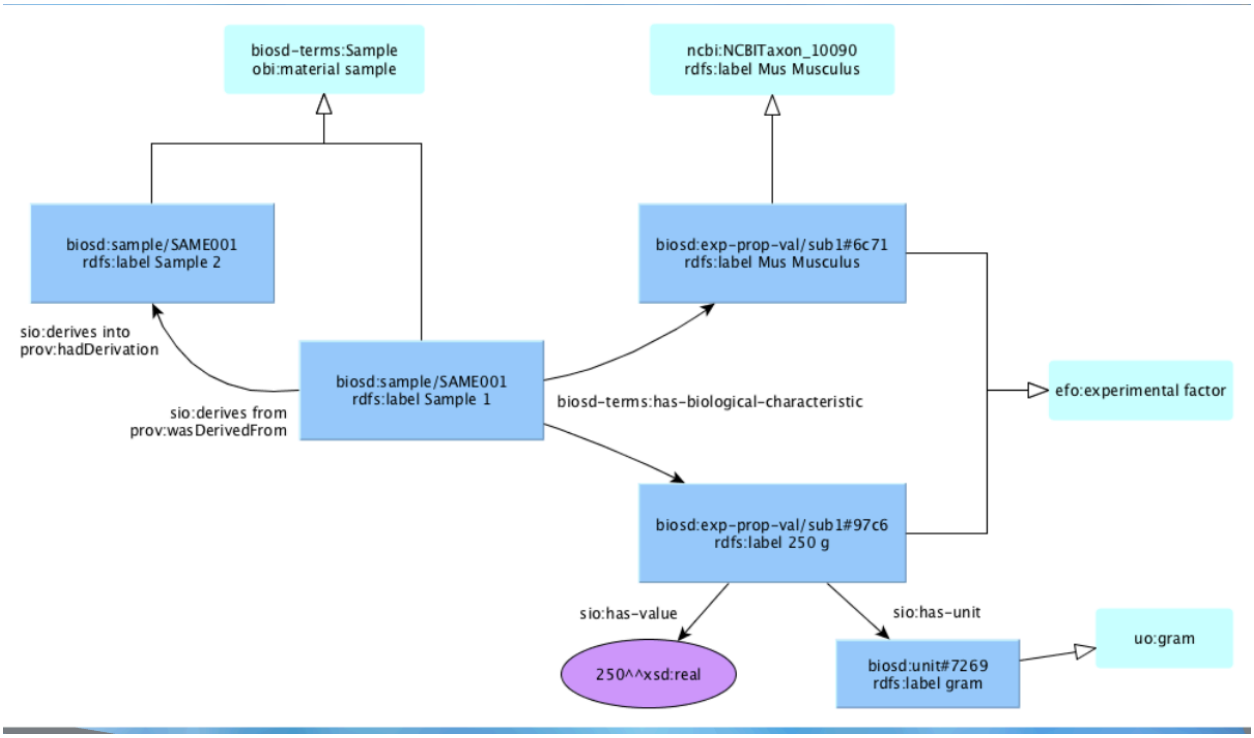


Note this is essentially the same as the bioschemas model

Relevant Ontologies:

- Definitions specific to Biosamples RDF dataset
- Experimental Factor Ontology (EFO)
- Biomedical Investigation Ontology (OBI)
- Information Artifact Ontology (IAO)
- Semantic Science Ontology (SIO)
- NCBI Taxonomy
- Chemical Entities of Biological Interest (ChEBI)

Example:
Doesn't seem to be quite the same schema as above



Representation of samples in OBI

Scope/Use Cases:
OBI is an OBO ontology for dealing with investigations. Despite its name, many classes can be reused outside biomedicine.

Note that OBI is an ontology rather than a schema. It provides open-world constraints over how an RDF instance graph of samples should be organized. E.g. processes are linked to material entities via has-specified-input and has-specified-output

Paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154556

Specification:
See the OWL file

Ontologies used:
Imports other OBO ontologies. For example, to represent a sample of lung tissue you would have an instance graph with an instance of an uberon lung and an instance of obi sample

Examples:
See examples of use in TURBO and BCO below

Representation of samples in TURBO

Scope/Use Cases:

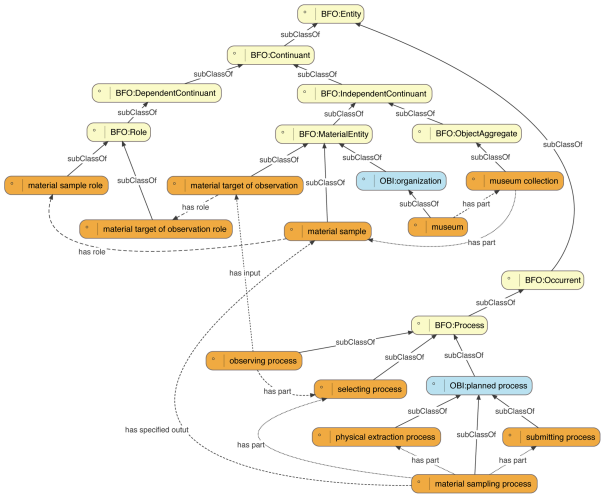
Specification:

Ontologies used:

Examples:

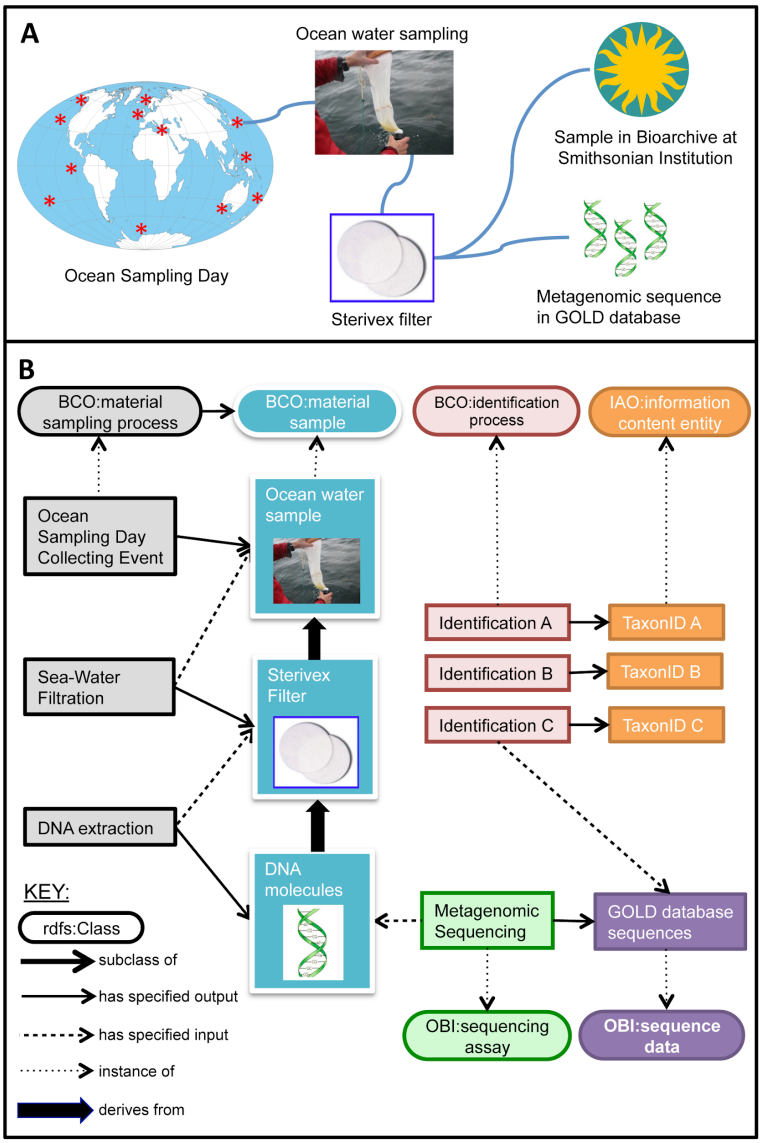
Representation of samples in the Biological Collections Ontology (BCO)

Scope/Use Cases:
The BCO was designed to model the collection of biological entities and their interactions (Walls et al. 2014). It was created to fill a gap in biodiversity informatics that prevented integration between data from environmental samples and observations with or without a physical specimen (Fig). This barrier was a problem because observations of organisms have limited usefulness outside the context of the environment from which they were collected. For example, biodiversity observations have been used to answer important questions about climate change only because their environmental context had been added using BCO (Li et al. 2019).
Specification:



From Walls et al. 2014 - Orange is BCO core

Ontologies used:
ENVO, PCO, OBI, IAO
Examples:



From Walls et al. 2014

MIxS Sample packages in GenSC

Scope/Use Cases:
Minimal information about a sequence (including metadata about the sample the sequence was derived from). Focus on metagenomics.

Specification:
The genomic standards consortium (https://press3.mcs.anl.gov/gensc/) provide minimum information checklists (https://press3.mcs.anl.gov/gensc/mixs/) (see https://www.nature.com/articles/nbt.1823) in the form of Excel files with fields that should be filled in when submitting sample data. There are different packages for different sample types.

For example, a built environment sample has generic fields (e.g collection_date) as well as specific fields such as absolute air humidity.

Note that in contrast to a graph representation like DATS, MIxS-described samples are “flat” sets of property-value pairs, similar to bioschemas.

Examples:
TODO
Can see in ncbi or ebi biosamples

Representation of samples in IGSN

Scope/Use Cases:
Geological samples

Specification:

See https://github.com/IGSN/metadata/wiki; uses XSD

Metadata is split between ‘registration’ and ‘descriptive’

Descriptive metadata has an RDF-like model whereby samples can be inter-related via typed edges. Standard fields like geolocation.

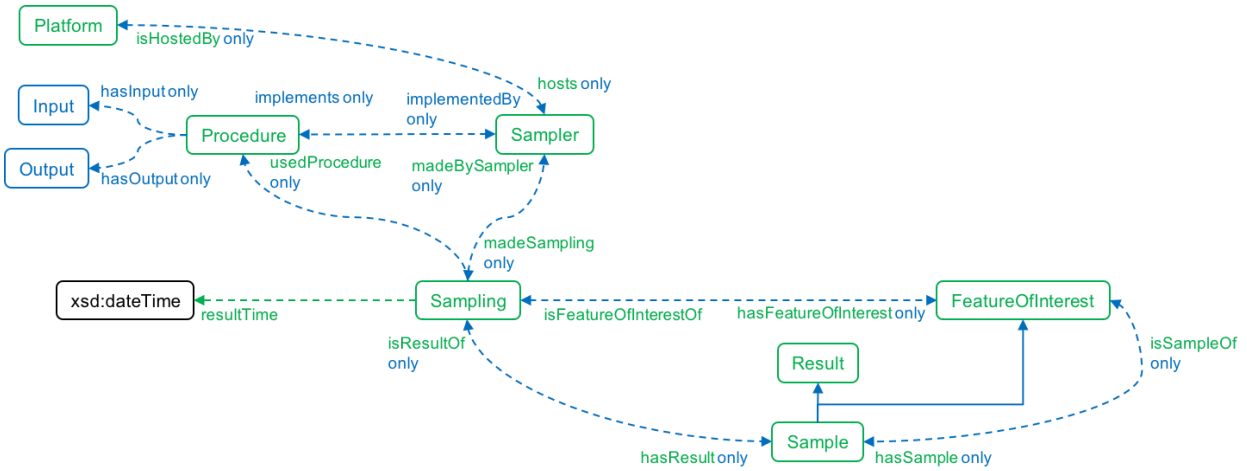
Has a field ‘material’ *Categorize the material that composes the sample, e.g. water, granite, tissue. Idea is to create a high-level cross-domain vocabulary. (1..N, nillable). ‘lot’ type samples (dredge haul, drill core) may have multiple materials included. Material may be categorized under different schemes. Implementation should be a ‘scoped’ name (vocabulary URI, concept/term URI, label for display).*

Representation of samples in SSNO

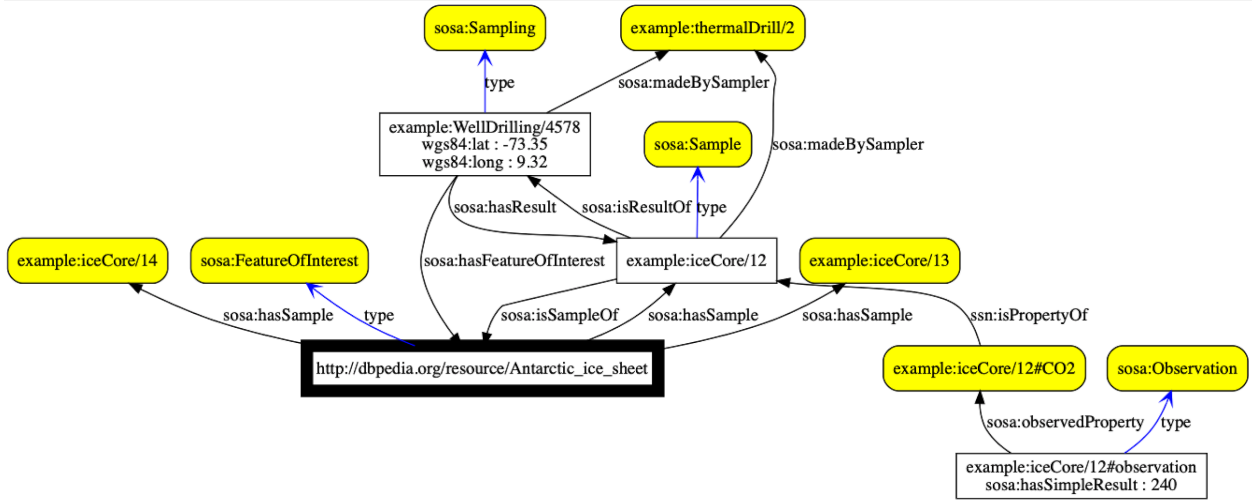
Scope/Use Cases:
Sensor networks

Specification:

https://www.w3.org/TR/vocab-ssn/#Sampling
https://www.w3.org/TR/vocab-ssn/#SOSASample



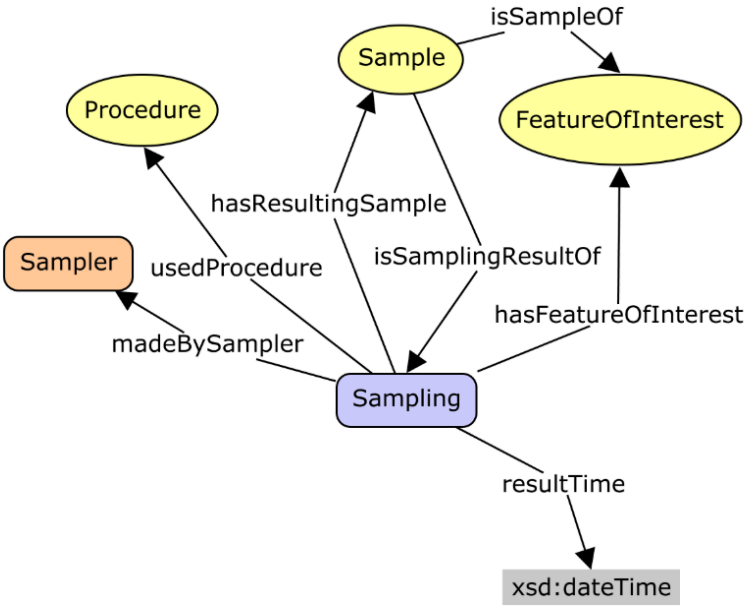
Visualization of https://www.w3.org/TR/vocab-ssn/integrated/examples/ice-core.ttl



Representation of samples in RDA

The Research Data Alliance have a working group:
https://www.rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig

From Simon Cox's presentation to RDA:
https://docs.google.com/presentation/d/1j2uRZ8aAQImgmcJ6MkNKtrzAi7QSY-mJd5jWWGjjR98/edit#slide=id.g265a3b0a90_0_12



In addition to the representation of physical samples and their metadata, RDA has also endorsed a working group to develop standards for attribution metadata for the curation and maintentance of research collections (Thessen et al. 2019). This standard links people to samples via the curatorial actions they perform, can be represented in RDF, and used the Contributor Role Ontology, VIVO, and PROV.

Use Cases for a unified RDF/JSON-LD model

- Database integration: combine sample data from different databases (e.g. biogeochemical and metagenomic)
    - Comparing across datasets
    - Discovery / bioinformatics analyses
    - Cohort building
- Web Search/SEO
- Advanced search, e.g. in faceted browser
- Resolving IDs to useful information for scientists

TODO: competency questions

Preliminary Material

RDF Datamodel

RDF is...

JSON-LD

JSON-LD is a set of conventions in a JSON document that allows unambiguous mapping to an RDF model...

OBO

OBO is a collection of ontologies intended to interoperate together and collectively fulfil a variety of different use cases for describing data. Example ontologies in OBO include ENVO (environments), OBI (investigations), GO (gene ontology), RO (relations).

Document Conventions and IRI prefixes

Within this document, the following namespace prefix bindings are used:

Prefix	Namespace
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs:	http://www.w3.org/2000/01/rdf-schema#
shex:	
xsd:	http://www.w3.org/2001/XMLSchema#
owl:	
sio:	
wgs84	http://www.w3.org/2003/01/geo/wgs84_pos#
OBI:	http://purl.obolibrary.org/obo/OBI_
sample:	http://purl.obolibrary.org/obo/OBI_0000747
ebi_bs:	http://rdf.ebi.ac.uk/resource/biosamples/sample/

derived_from:	USE RO (EBI uses http://purl.org/pav/2.0/derivedFrom)
sampled_from:	

Note that for convenience we declare prefixes for OWL classes. This means we can write the more human readable “sample:” which expands to the complete IRI for http://purl.obolibrary.org/obo/OBI_0000747 rather than the less transparent “OBI:0000747”.

Note in all JSON-LD examples in this document we assume the presence of a JSON-LD context that pre-declares all of these.

Thus we can write:

```
{
 “@id” : “ SAMPLE:123”,
 “@type” : “sample”
}
```

And it is interpreted as rdf:type http://purl.obolibrary.org/obo/OBI_0000747

Identifiers

In RDF/JSON-LD all identifiers are IRIs. These can be represented in compact form as CURIEs. CURIEs can be independently resolved using resolvers such as n2t.net and identifiers.org

The schema presented here is independent of any IRI schema. However, we recommend:

- Choosing IRIs that are likely to be stable
- Allowing resolution to either computable data (e.g. JSON-LD or RDF) or human-friendly web pages
- Registering a standard prefix for the IRI space with n2tnet and identifiers.org
- See McMurry et al(McMurry et al. 2017)

RDF/JSON-LD Representation of Samples

We provide examples of how to create RDF graphs describing samples and related properties in a way that is conformant with OBO ontologies. We later provide ShEx profiles for particular applications.

Basic instantiation

Example RDF (turtle):

```
ebi_bs:SAMN02847463 a sample: ;
        rdfs:label “Environmental/Metagenome sample from hot springs
metagenome” .
```

Here we have two triples with the same subject URI (the sample URI). The first tells us what the type of the URI is, the second provides the name/label for the sample. [note the label is taken directly from the EBI record, it’s maybe not the best way of phrasing it, e.g. a sample is not taken from a metagenome…]

Equivalent JSON-LD

```
“samples” : [
  {“id” : “ebi_bs:SAMN02847463”,
   “label” : “Environmental/Metagenome sample from hot springs metagenome” } ]
```

(assumes context induces typing on “samples” list)

Representing derivation

It is important not to confuse the features of the sample with the source it was sampled from. Also, samples may be derived from samples.

We recommend a ‘sampled from’ triple to indicate the relationship between a sample and the original source material (e.g. human, mountain, ocean layer), and ‘derived from’ triples for connecting a processed sample with the original sample. An OWL property chain provides the rule that propagates sampled-from over derived-from.

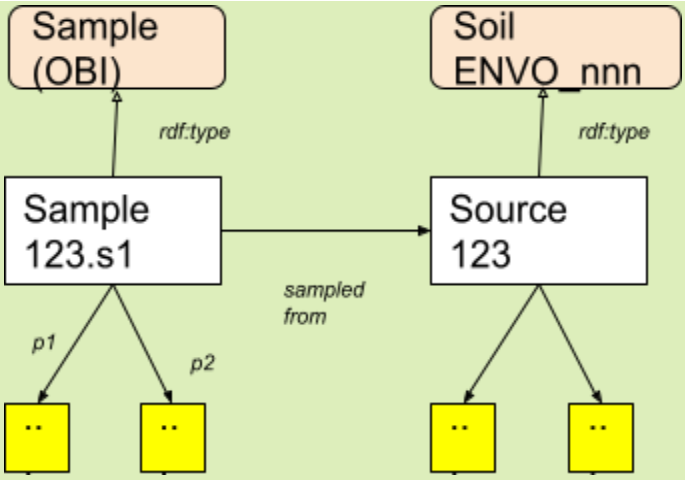
Note that is possible to use a blank node to represent the sample source, e.g

```
ebi_bs:SAMN02847463 a sample: ;
```

```
        rdfs:label “Environmental/Metagenome sample from hot springs
metagenome” ;
        sampled_from: [
            A ENVO:....
        ]
```

However, we recommend minting a IRI for this purpose

```
ebi_bs:SAMN02847463 a sample: ;
        rdfs:label “Environmental/Metagenome sample from hot springs
metagenome” ;
        sampled_from: ebi_bs:SAMN02847463.src .
ebi_bs:SAMN02847463.src a ENVO: … ;
```



Graphical depiction of the above named graph. Two instances (white boxes) of sample and source entity, each is an instance of an OBI class and an ENVO class respectively.

Geo-location

Note: we must be careful to distinguish where sample was derived (usually the more interesting scientifically) from where the sample is stored or processed. In “flattened” sample representations there is typically a single geolocation field for the sample, and this is *implicitly* the geolocation of the source. However, this is not explicit and there are examples where people have used the location of the sample itself. For some use cases (e.g. museum collections) both the location of the specimen and where it came from are important.

TODO - use W3C standards here, no need to reinvent

```
ebi_bs:SAMN02847463 a sample: ;
      wgs84:lat ….
```

Sample preparation

Sample registration metadata

Sample bio/chemical/geological characteristics

TODO: Description of the PATO model for attaching qualities to material entities; use of different ontologies for qualitative characteristics; examples for soil.

Packages and profiles

GenSC/MIxS pioneered the concept of a “package”, or a bundle of properties applicable in a certain domain. For example, the properties of relevance to a soil sample have some overlap but are distinct from properties of relevance from a human gut sample in a healthcare or health research context.

Currently MIxS packages are unformalized and exist as excel templates.

Here we provide mappings between MIxS properties and our RDF schema, as well as formalize the relationship between these properties and their values

TODO...

Biomedical and biological characteristics

```
ebi_bs:SAMN02847463 a sample: ;
       rdfs:label “Environmental/Metagenome sample from hot springs
metagenome” ;
       derived_from: [
           A ENVO:....
       ]
```

Chemical characteristics

Geological characteristics

Tools

Validation

We can use OWL for open-world validation. TODO Provide details.

For closed-world validation we can use ShEx. We can have separate ShEx shapes for different profiles, e.g. wastewater, soil, tumor.

Flattening/Unflattening

Most biosamples (i.e those in ebi/ncbi biosamples) use a “flat” property-value representation where all properties are attached to the sample; there is not a separate record ID for the source or intermediates.

It will be useful to have standard tools that flatten/unflatten. It’s not clear if the latter can be done deterministically in all cases.

References

McMurry, Julie A., Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot, et al. 2017. “Identifiers for the 21st Century: How to Design, Provision, and Reuse Persistent Identifiers to Maximize Utility and Impact of Life Science Data.” *PLoS Biology* 15 (6): e2001414.