BioSchemas for Samples v0.2

# Summary

This document summarises the discussion around the next version of the BioSamples schema. This follows previous discussions about the last version v0.1 [documented here](https://docs.google.com/document/d/1JS7ofrdO3UzkAI12C702AdniJ5fr-CRGIC84bQC-_VI/edit#heading=h.tifzibl1jmv). We intend to publish this schema on bioschema.org (on 2.2.18) while remaining open to comments from the community which can be fed into the next version.

Generally debate has centred around which properties should be recycled from either schema.org or bioschemas.org and which additional features we would require to meet our use cases. Due to the broad scope of biological samples the associated descriptive metadata is equally broad. Therefore, our schema requires an unrestricted additionalProperty field. As this is already available in schema.org we do not need to define any new properties to describe samples.

Due to the wide-range of potential attributes we require an equally flexible code annotation to link values to existing ontologies. **CategoryCode** is a new proposed type in bioschemas that is able to act as a standalone ontology object. The Ontology Lookup Service will expose their ontologies as CategoryCodes. To mirror this we propose that CategoryCode should be a new StructuredValue thus allowing us to use CategoryCode within PropertyValue:valueReference. We envisage that this type would be useful to other bioschemas domains and when used in conjunction with our samples schema, would meet our use cases.

After community discussion it became clear that the identifier would need to be more flexible than a BioSamples ID however this ID would need to be properly prefixed and ideally be resolvable in identifiers.org (A valid BioSamples ID would be ‘biosample:SAME12345’. Although we are currently engaging with identifers.org to change our prefix to an optional [s]?.). To enable RDF validation it was suggested that at least one type to be of minimum marginality. Despite appreciation that some internal identifiers should not be exposed to avoid privacy violation (as is often the case with biobanks), a public identifier should be assigned (potentially a BioSamples ID) to avoid duplication and enable RDF validation. This should ideally be a properly prefixed ID.

# 

# Version Differences

N.B. Motivation and implementation plans (including objectives, milestones and deliverables) are still relevant and are discussed in the previous [documented here](https://docs.google.com/document/d/1JS7ofrdO3UzkAI12C702AdniJ5fr-CRGIC84bQC-_VI/edit#heading=h.tifzibl1jmv).

To address comments on v0.1 of the schema we have:

1. Added marginality and cardinality
2. Proposed CategoryCode (within PropertyValue:valueReference)
3. Added more detail to the BioSchemas descriptions
4. Changed how we link to data. Now using http://schema.org/Dataset

Some preliminary suggestions for v0.3 (ADD NEW SUGGESTIONS HERE):

1. @type : [“BioChemEntity”, “Sample”] could be extended to [“BioChemEntity”, “Sample”, “Record’]. Other groups are adopting this method to describe their entities as records which may better assert potential for linking the record to datasets.
2. To annotate the attribute with an ontology term we could introduce a new PV type nameReference that allows us to append another CategoryCode. This would suit the [MIABIS use case](https://github.com/MIABIS/miabis/wiki/Data-describing-Sample-Collection) where they have specific codes for their attributes which specify value validation.
3. Use CategoryCode as value for unitCode inside the PV

# 

# Proposed Sample Schema

@type" : ["BioChemEntity","Sample"]

Note that ‘Sample’ is not linked to a specific ontology URI.

## Sample

[*http://bioschemas.org/Sample*](http://bioschemas.org/Sample)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Property | Schema.org | BioSchemas Description | Type | Cardinality | Marginality |
| identifier | <http://schema.org/identifier> | Unique sample IDs. Where possible this should be an identifiers.org compliant prefixed id e.g. a BioSamples ID biosample:SAME12345 | Text | MANY | Minimum |
| name | <http://schema.org/name> | A human readable name for the sample. This should not be an additional identifier. Additional identifiers should be added to the identifier field. | [Text](http://schema.org/Text) | MANY | Optional |
| description | <http://schema.org/description> | A description of the sample in free text. This should not contain information that could be better expressed as key/value pairs. These should be expressed using additionalProperty | [Text](http://schema.org/Text) | ONE | Optional |
| url | <http://schema.org/url> | An access URL for this sample, either in BioSamples or in a Biobank or elsewhere. | [URL](http://schema.org/URL) | ONE | Recommended |
| dataset | [http://](http://schema.org/url)schema.org/Dataset | An access URL that provides a link to a dataset that contains data about this sample record. | [URL](http://schema.org/URL) | MANY | Optional |
| additionalProperty | <http://schema.org/additionalProperty> | A property-value pair representing an additional characteristics of the entity, e.g. “Organism: Homo sapiens” or “tissue type: leaf” | [PropertyValue](http://schema.org/PropertyValue) | MANY | Optional |

Table 1: The ‘Sample’ Bioschemas concept and suggested properties

## Property Value

[*http://schema.org/PropertyValue*](http://schema.org/PropertyValue)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Property | Schema.org | BioSchemas Description | Type | Cardinality | Marginality |
| name | <http://schema.org/name> | The name of the additional property. Otherwise known as a key or attribute. | [Text](http://schema.org/Text) | ONE | Optional |
| unitCode | schema.org/unitCode | The unit of measurement given using the UN/CEFACT Common Code (3 characters) or a URL. Other codes than the UN/CEFACT Common Code may be used with a prefix followed by a colon. | Text or URL | ONE | Optional |
| unitText | schema.org/unitText | [A string or text indicating the unit of measurement. Useful if you cannot provide a standard unit code forunitCode.](http://schema.org/unitCode) | Text | ONE | Optional |
| value | schema.org/value | The value of the quantitative value or property value node. | Boolean or Number or Text | ONE | Optional |
| valueReference | schema.org/valueReference | Catagory code to enhance the value this could be a link to an ontology term in the Ontology Lookup Service. | **CatagoryCode** | MANY | Optional |

# 

# Category Code [pending]

https://pending.schema.org/CategoryCode

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Property | Bioschema.org | BioSchemas Description | Type | Cardinality | Marginality |
| codeValue | https://pending.schema.org/codeValue | A short textual code that uniquely identifies the value | Text | ONE | Optional |
| name | schema.org/name | The label | Text | ONE | Optional |
| url | schema.org/url | The URI for the vocabulary code | URL | ONE | Optional |

Including the Category Code as a valueReference is a new proposal. The current walkthrough from ‘thing’ is this:

EXISTS IN SCHEMA.ORG

https://pending.schema.org/Thing

https://pending.schema.org/subjectOf

https://pending.schema.org/CreativeWork

http://schema.org/CategoryCodeSet

http://schema.org/hasCategoryCode

https://pending.schema.org/CategoryCode

We propose that CategoryCode should be a StructuredValue so that it can be used to annotate terms throughout schemas.org

# 

# Mockup example of schema in use

{

"@context" : "http://schema.org",

"@type" : ["BioChemEntity","Sample"],

"identifier" : "biosample:SAMEA104383111",

"name" : "HPSI0617pf-tuji",

"description" : "Fibroblast cell line ...",

"url" : "<http://www.ebi.ac.uk/biosamples/samples/SAMEA104383111>",

"dataset": [...],

"additionalProperty" : [ {

"@type" : "PropertyValue",

"name" : "cellType",

"value" : "fibroblast from the user",

"valueReference": {

"@type": "CategoryCode",

"name": "fibroblast",

"url": "http://purl.obolibrary.org/obo/CL\_0000057",

"codeValue": "CL:0000057"

}

}]

}

### Crawler Usecases

1. Biobanks should be able to crawl the BioSamples database to identify all the published (and searchable) datasets derived from samples they have provided
2. Public archives should be able to crawl Biobank websites, in order to identify samples that are known to have public accessions in the BioSamples database AND that can be made publicly available, and thereby link public samples to a biobank (“where can I get more of this sample?”).
3. In case of privacy or consent considerations, only the biobank should know what are the specific samples connected to publicly available datasets
4. Public archives should be able to crawl Biobank websites, in order to identify ‘sanitised’ sample metadata descriptions (again, in case of confidentiality or consent considerations). Biobanks remain responsible for ensuring only authorised metadata is visible, and can control access to restricted samples.
5. BioSamples should be able to crawl MarRef and identify updated information (versus what is already stored in BioSamples). This is to enable BioSamples to build curation objects for these samples.
6. OLS should be able to crawl BioSamples and count how many times their markup is used.
7. BioSamples should be able to crawl OLS and look for enhancements to metadata for marked up terms e.g. synonyms.

### Assumptions

1. Each sample provided by a biobank has an opaque pseudonymous identifier that is assigned by the biobank to identify a specific sample (referred to hereafter as the “sample name”)
2. Each sample reported in a public archive or used to generate a public dataset has a public BioSamples database accession (hereafter called “sample identifier”).
3. In some cases, a biobank may issue different sample identifiers when providing the same sample to different projects as best practice to prevent exposing private information. This may result in duplicated sample accessions in the BioSamples database

Given these use cases and assumptions, we will use Bioschemas to describe sample links. The main challenge is therefore the identification of links between sample identifiers (within Biobanks) and sample accessions (from the BioSamples database). This is not always possible without considerable additional curation effort, but of the 5 million samples in the BioSamples database, over 4 million declare either a ‘synonym’, ‘sample source name’ or ‘source name’ attribute, frequently used to encode the original biobank sample name. Exposing these in a structured manner through the BioSamples database would allow Biobanks to crawl and analyse this content, marrying sample that are recognised with their own internal identifiers.

Once this mapping is done, Biobanks can then re-expose these links through structured content on their own websites, allowing public resources to reciprocate links from public records back to the sample provider.

### Ontology Lookup Service Usecases

1. OLS wants to crawl ontology markup in bioschema resources to count which ontologies are being used. E.g. crawling biosamples to count how many times we link to certain terms in OLS.
2. OLS wants to expose its own resource as bioschemas markup that can be referenced in other bioschemas domains. Thus allowing OLS to enrich markup terms when necessary. E.g. what are the synonyms for this term I have marked up in my sample?

OLS needs to expose a standalone curation object in bioschema.org markup. This type (aka VocabularyCode/OntologyCode/CatagoryCode) will be used to markup the ontologies in OLS. Other resources should be able to map/include these ontology types in their own schema.

Therefore CategoryCode meets the requirements. OLS can use this as a schema.org compliant standalone object and the samples schema can reflect this usage in its schema by providing CategoryCode as a valueReference.