

JSON schema extension for semantic data validation

Simon Jupp

Ontology Project Lead

Samples, Phenotypes and Ontologies

EMBL-EBI

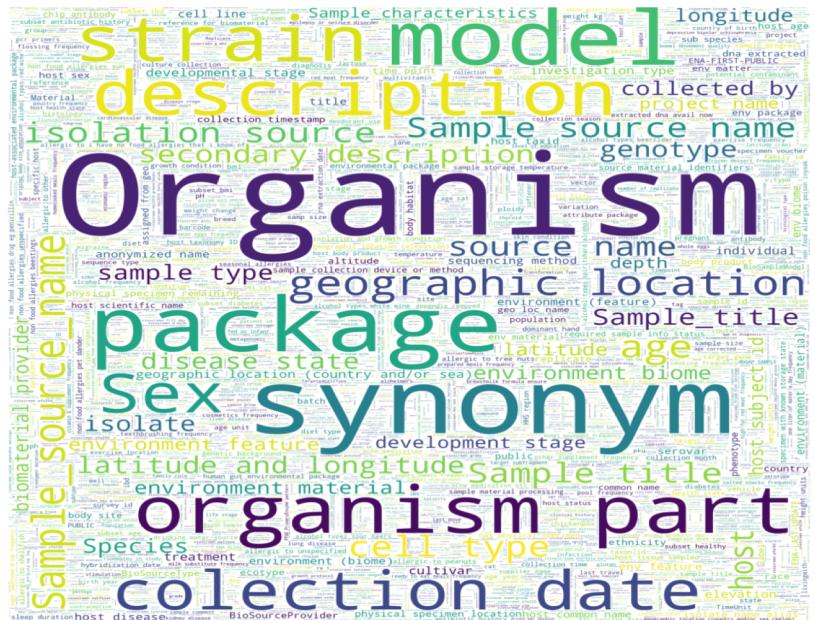
ELIXIR Biohackathon

November 12th 2018

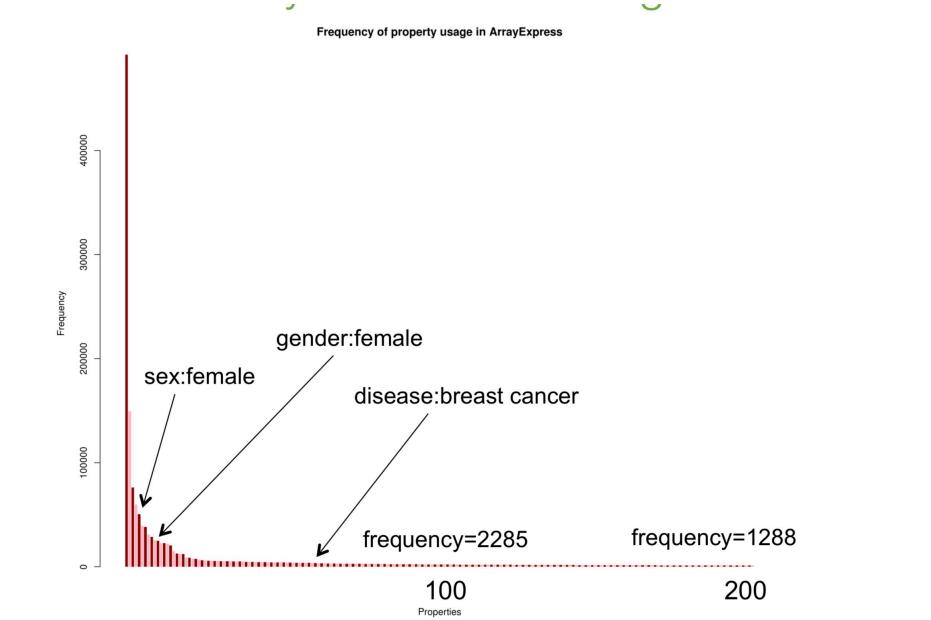
Paris, France

Metadata standards and validation

- Metadata quality in biological data archives is low making data hard to Find, Interoperate and Reuse
 - If we can improve the quality of data coming in, we can lower the cost of data curation



Data attributes from EMBL-EBI Biosamples



Long tail of “unknown” attributes

JSON schema

- JSON schema is a popular mechanism for communities to define metadata standards

```
1  {
2    "$schema": "http://json-schema.org/draft-07/schema#",
3    "title": "Sample Schema",
4    "properties": {
5      "sample_id": {
6        "type": "string",
7      },
8      "sample_name": {
9        "type": "string",
10     },
11      "taxon": {
12        "type": "integer",
13      },
14      "donor_age": {
15        "type": "integer",
16      },
17      "donor_age_unit": {
18        "type": "string",
19        "enum": ["years", "months", "weeks", "days"]
20      },
21      "tissue": {
22        "type": "string",
23      },
24      "diseases": {
25        "type": "array",
26        "items": {
27          "type": "string",
28        }
29    }
```



Valid

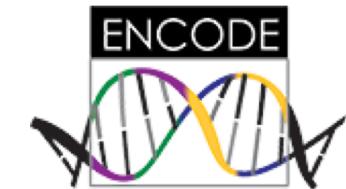
Biological Sample Schema

```
1  {
2    "sample_id" : "1234",
3    "sample_name" : "Bio rep #1",
4    "taxon" : "9606",
5    "donor_age" : "22",
6    "donor_age_unit" : "years",
7    "tissue" : "liver",
8    "diseases" : [
9      "Type II diabetes"
10   ]
11 }
```

Instance data

Example projects using JSON schema

- Human Cell Atlas metadata model described using strict JSON schema.
<http://schema.data.humancellatlas.org>
- Functional Annotation of Animal Genome JSON schema for experiment and sample metadata
https://www.ebi.ac.uk/vg/faang/rule_sets
- ENCODE: Encyclopedia of DNA elements.
<https://www.encodeproject.org/profiles/>
- EMBL-EBI unified submissions system to ENA, BioStudies and BioSamples



Expressing an ontology constraint in JSON schema

```
1  {
2      "sample_id" : "1234",
3      "sample_name" : "Bio rep #1",
4      "taxon" : "9606",
5      "donor_age" : "22",
6      "donor_age_unit" : "years",
7      "tissue" : "liver" →
8      "tissue_ontology" : [
9          "UBERON:0002107"
10         ],
11     "diseases" : [
12         "Type II diabetes mellitus" →
13     ],
14     "disease_ontology" : [
15         "DOID:9352"
16     ]
17 }
```

tissue_ontology:
should be a CURIE
from UBERON
ontology and a
subclass or part of
anatomical entity

disease_ontology:
should be a CURIE
from Human
Disease Ontology
and a subclass of
disease

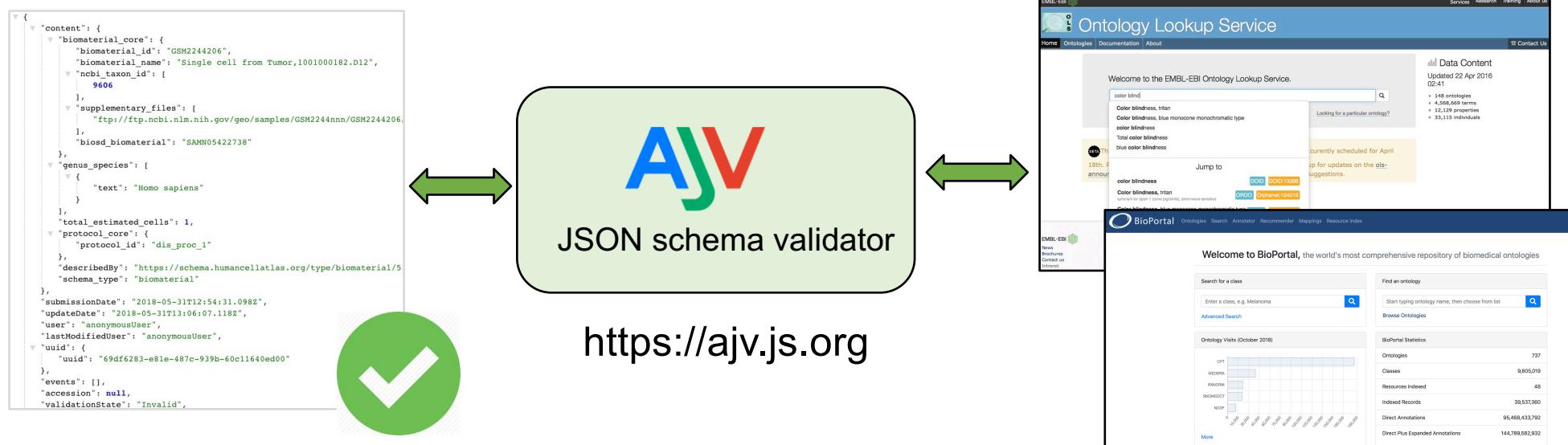
Hacking tasks 1

- Agree on a representation in JSON schema
- Some prior work at EMBL-EBI and Human Cell Atlas project
- Collect use-cases with data clearinghouse and validation group

```
2 "ontology": {  
3     "description": "A term from the ontology [UBERON]",  
4     "type": "string",  
5     "graph_restriction": {  
6         "ontologies": [  
7             "obo:hcao",  
8             "obo:uberon"  
9         ],  
10        "classes": [  
11            "UBERON:0000465"  
12        ],  
13        "relations": [  
14            "rdfs:subClassOf"  
15        ],  
16        "direct": false,  
17        "include_self": true  
18    }  
19}
```

Hacking tasks 2

- Implement a validator in JavaScript as AJV plugin
 - Working prototype already done
- Use EMBL-EBI Ontology Lookup Service to validate terms
 - Develop fast validation service with Node JS



Get involved

- Additional use cases or test datasets
- Anyone with experience working with or building JSON schemas
- Developers who have used AJV or built services with Node.js
- Ontologists or Semantic Web developers