JSON schema extension for semantic data validation (topic #9)

Simon Jupp

Ontology Project Lead

Samples, Phenotypes and Ontologies

EMBL-EBI

ELIXIR Biohackathon

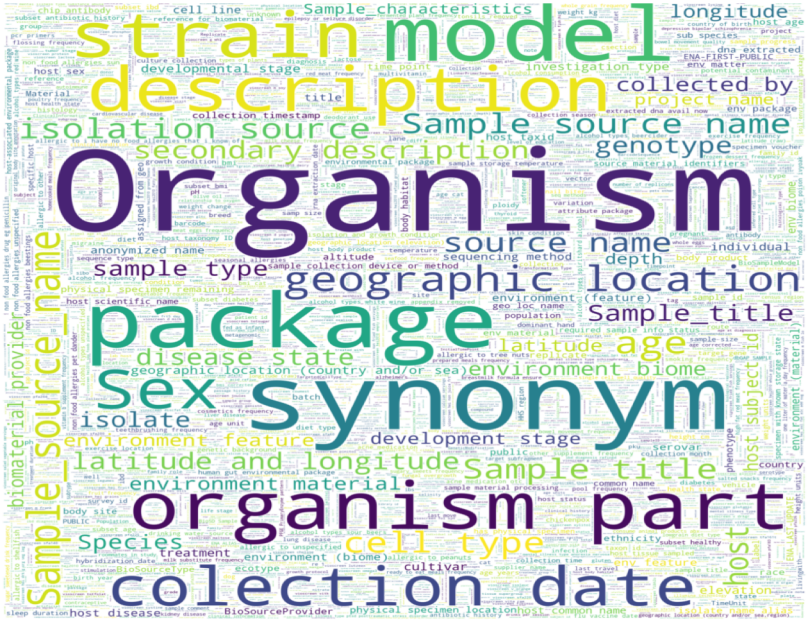
November 12th 2018

Paris, France

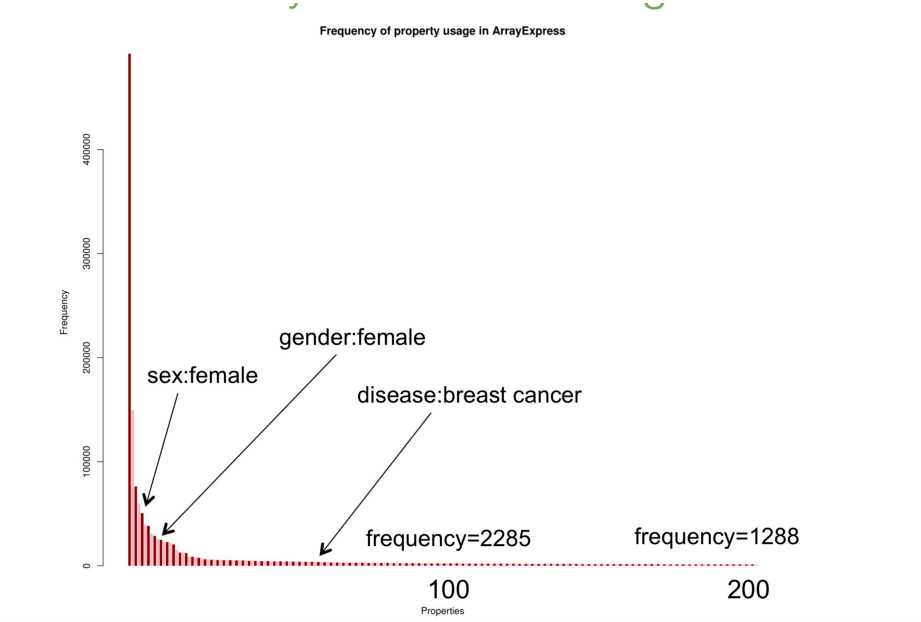
EMBL-EBI

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

```json
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Sample Schema",
  "properties": {
    "sample_id": {
      "type": "string",
    },
    "sample_name": {
      "type": "string",
    },
    "taxon": {
      "type": "integer",
    },
    "donor_age": {
      "type": "integer",
    },
    "donor_age_unit": {
      "type": "string",
      "enum": ["years", "months", "weeks", "days"]
    },
    "tissue": {
      "type": "string",
    },
    "diseases": {
      "type": "array",
      "items": {
        "type": "string",
      }
```

Biological Sample Schema

Valid

```json
{
    "sample_id" : "1234",
    "sample_name" : "Bio rep #1",
    "taxon" : "9606",
    "donor_age" : "22",
    "donor_age_unit" : "years",
    "tissue" : "liver",
    "diseases" : [
      "Type II diabetes"
    ]
}
```

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

```
{
  "sample_id" : "1234",
  "sample_name" : "Bio rep #1",
  "taxon" : "9606",
  "donor_age" : "22",
  "donor_age_unit" : "years",
  "tissue" : "liver",
  "tissue_ontology" : [
      "UBERON:0002107"
  ],
  "diseases" : [
    "Type II diabetes mellitus"
  ],
  "disease_ontology" : [
    "DOID:9352"
  ]
}
```

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**

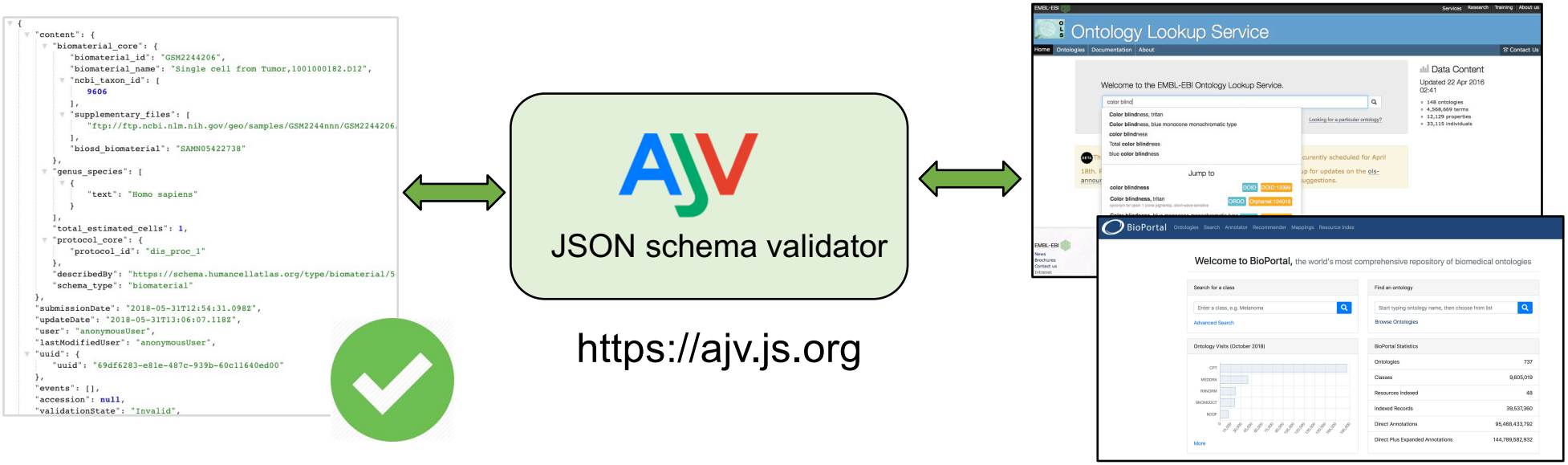
Hacking tasks 1

- 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

```
"ontology": {
  "description": "A term from the ontology [UBERON]",
  "type": "string",
  "graph_restriction": {
    "ontologies": [
      "obo:hcao",
      "obo:uberon"
    ],
    "classes": [
      "UBERON:0000465"
    ],
    "relations": [
      "rdfs:subClassOf"
    ],
    "direct": false,
    "include_self": true
  }
```

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
- Additional functionality
  - Connect to BioPortal API
  - Other types of identifier validation (e.g. CURIEs against identifiers.org)
- Integration with JSON schema documentation generators