Data Repositories Specification 0.1

Bioschemas specification for describing data repositories and data catalogues in the life-sciences

This document covers the [schema.org](http://schema.org) and [bioschemas.org](http://bioschemas.org) properties necessary to describe data repositories, catalogues and databases in the life sciences, organised by their requirement level (M for mandatory, R for recommended and O for optional) and driven by a set of [use cases](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit#gid=837395607). A simple example in JSON-LD is also provided. The majority of these fields already exist in schema.org, others are pending.

Note that this draft will be improved progressively through the bioschemas process.

# Properties

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Based on schema.org/DataCatalog | | | | |
| **Property** | **Expected Type** | **Description** | **CN** | **MG** |
| [name](http://schema.org/name) | [Text](http://schema.org/Text) | The name of the item. | One | M |
| [description](http://schema.org/description) | [Text](http://schema.org/Text) | A description of the item. | One | M |
| [url](http://schema.org/url) | [URL](http://schema.org/URL) | URL of the item. | One | M |
| [keywords](http://schema.org/keywords) | [Text](http://schema.org/Text) | Keywords or tags used to describe this content. Use terms from Controlled Vocabularies where possible. | One | M |
| [provider](http://schema.org/provider) | [Organization](http://schema.org/Organization) or [Person](http://schema.org/Person) | Contact information for this data repository. | Many | M |
| [dataset](http://schema.org/dataset) | [Dataset](http://schema.org/Dataset) | A dataset contained in this catalog.  Inverse property: includedInDataCatalog. | Many | R |
| [identifier](http://schema.org/identifier) | [PropertyValue](http://schema.org/PropertyValue) or  [Text](http://schema.org/Text) or  [URL](http://schema.org/URL) | Unique identifier for the data catalog. Use identifiers.org namespace prefix to uniquely identify the data catalog where possible. | One | R |
| [alternateName](http://schema.org/alternateName) | [Text](http://schema.org/Text) | An alias for the item. | Many | R |
| [license](http://schema.org/license) | [CreativeWork](http://schema.org/CreativeWork) or [URL](http://schema.org/URL) | A license document that applies to this content, typically indicated by URL. | One | R |
| [citation](http://schema.org/citation) | [CreativeWork](http://schema.org/CreativeWork) or [Text](http://schema.org/Text) | A citation or reference to another creative work, such as another publication, web page, scholarly article, etc. | Many | R |
| [sourceOrganization](http://schema.org/sourceOrganization) | [Organization](http://schema.org/Organization) | The Organization on whose behalf the creator was working. | Many | R |
| [dateModified](about:blank) | [Date](http://schema.org/Date) or [DateTime](http://schema.org/DateTime) | The date on which the data repository was most recently modified. | One | R |
| [publication](http://schema.org/publication) | [PublicationEvent](http://schema.org/PublicationEvent) | A publication event associated with the item.  A PublicationEvent corresponds indifferently to the event of publication for a CreativeWork of any type e.g. a broadcast event, an on-demand event, a book/journal publication via a variety of delivery media. | Many | R |
| [datePublished](http://schema.org/datePublished) | [Date](http://schema.org/Date) | Date of first broadcast/publication. | One | O |
| [fileFormat](http://schema.org/fileFormat) | [Text](http://schema.org/Text) or [URL](http://schema.org/URL) | Media type, typically MIME format of the content. | Many | O |

**Legend:** *CN: Cardinality (one, many)*

*MG: Marginality (M: minimum; R: recommended; O: optional)*

# Examples

Schema.org [suggests](http://schema.org/docs/gs.html) implementing metadata using JSON-LD, RDFa or Microdata. JSON-LD is the recommended format by Google, but any of these formats can be used for embedding information about tools in a web page or other online resource. The example below with additional fields can be found on the Identifiers.org landing page.

|  |
| --- |
| Example 1 - JSON-LD. Data repository description for Identifiers.org using minimum fields. |
| <script type="application/ld+json">  {  "@context": "http://schema.org",  "@type": "DataCatalog",  "name": "Identifiers.org",  "description": "The Identifiers.org registry contains registered namespace and provider prefixes with associated access URIs for a large number of high quality data collections. These prefixes are used in web resolution of compact identifiers of the form PREFIX:ACCESSION or PROVIDER/PREFIX:ACCESSION commonly used to specify bioinformatics and other data resources.",  "url": "https://identifiers.org/",  "dataset": [  {  "@type" : "Dataset",  "name" : "Collection",  "description": "Collection catalog with each assigned a unique namespace.",  "identifier": "https://identifiers.org/MIR:00100005",  "url": "https://identifiers.org/MIR:00100005",  "keywords":"prefix"  },  {  "@type" : "Dataset",  "name" : "Resource",  "description": "Resources (physical locations) providing access to the data.",  "identifier": "https://identifiers.org/MIR:00100245",  "url": "https://identifiers.org/MIR:00100245",  "keywords": "provider"  }  ],  "keywords": ["registry","life science"],  "provider": [  {  "@type":"Person",  "name": "Identifiers.org support",  "email": "identifiers-org@ebi.ac.uk"  }  ]  }  </script> |

# Adoption by data repositories and registries

In production:

* [Identifiers.org](http://identifiers.org/)
* [FAIRsharing.org](https://fairsharing.org/)

In development:

* [GigaDB.org](http://gigadb.org) (NB - NOT included dataset list)
* Image Data Resource idr.openmicroscopy.org (also not included dataset list yet)
* [Human Protein Atlas](https://docs.google.com/a/scilifelab.se/document/d/1JxeMDgMuFMPwKlGzu9m9Et-EIpjDIB9BHqUXu82l9Dw/edit?usp=sharing)
* EGA
* FAIRDOM
* [Neuroscience Information Framework (NIF)](https://neuinfo.org) / [NIDDK Information Network (dkNET)](https://dknet.org) [SciCrunch.org]

# References

* [Shared folder](https://drive.google.com/open?id=0Bw_p-HKWUjHoNDJNUWltYVBkV1k)
* [Use-cases](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit#gid=837395607)
* [Mapping](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit#gid=292464567)
* [Crosswalk document](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit)
* [DataRepository Group](http://bioschemas.org/groups/datarepositories/)
* [DataRepository issue tracker](https://github.com/BioSchemas/bioschemas/labels/type%3A%20datarepository)
* [Schema.org](http://schema.org/)
* [BioSchemas](http://bioschemas.org/)
* [Google structured data testing tool](https://search.google.com/structured-data/testing-tool)

# Contacts for comments on the draft specification:

Rafael Jimenez (ELIXIR) - [rafael.jimenez@elixir-europe.org](mailto:rafael.jimenez@elixir-europe.org)

Peter McQuilton (FAIRsharing.org) - [peter.mcquilton@oerc.ox.ac.uk](mailto:peter.mcquilton@oerc.ox.ac.uk)

Sarala Wimalaratne (identifiers.org) - [sarala@ebi.ac.uk](mailto:sarala@ebi.ac.uk)