No longer maintained

Content has been moved to https://docs.google.com/document/d/1U51c2408X\_j8itLt8b5j8aTLrab6OEuucXJUXYCV4y8/edit#

Data Repositories Specification 0.1

**!! DRAFT - Work in progress !! Date 19th June 2017**

*Brief description of the specification*

# Pete.

# 

# Recommendation 30 June 2016

This version: [github link]

Latest published version: [github link]

Previous version: [github link]

Editors: [Names]

Authors: Peter McQuilton, Sarala Wimalaratne, Rafael Jimenez

Contributors: [Names]

# Abstract

This document describes a set of fields (properties, types and description) for the discovery and detailing of data repositories in the life sciences, organized by their requirement level and driven by a set of use cases. The majority of these fields already exist in the schema.org [Datacatalog](http://schema.org/DataCatalog). Note that this draft will be iteratively reviewed in conjunction with other Bioschema specifications.

The sections below outline the mandatory (minimum) and the recommended (optional) set of properties. These fields have been identified via a [crosswalk exercise](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit#gid=1261485211).

# Status of this Document

*Who will work on this document, who will review it and where do the ideas behind it come from (workshops, meetings, etc.).*

This document is the culmination of 3 Bioschemas workshops held in Harpenden, Hertfordshire, UK and the Wellcome Genome Campus, Cambridgeshire, UK.

# Table of Contents

[Recommendation [DAY] [MONTH] 2016](#_jdeykt68ile4)

[Abstract](#_iwpofulctdmg)

[Status of this Document](#_nrwnogeyt0xk)

[Table of Contents](#_vt2ihcfvn38e)

[Specification Template](#_ril5bbnyppbk)

[Type definition](#_uc6mc8iukqe)

[Data fields](#_2fbsxcd60sp9)

# Introduction

## **Problem statement**

The main reasons why this specification is needed. - Pete

Use cases - Pete

## **Proposed solution - Sarala**

### **Rationale**

The various steps needed to adopt this specification.

### **Goals**

What are the goals of this specification

* Increase discoverability

### **Scope**

What this document is intended for.

# Data model

## **Type definition**

### **Data fields**

**Legend:**

*CN: Cardinality (one, many)*

*CG: Content Guideline (M: minimum; R: recommended; O: optional)*

*CV: Controlled Vocabulary*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Property** | **Expected Type** | **Description** | **CN** | **CG** | **CV** |
| **Proposed new properties for this Type** | | | | | |
| accessURL | Text | A URL-like string specifying a rule for resolving the dataset identifiers. | One | M |  |
| identifierPattern | Text | Regular expression describing alphanumeric strings used to identify items (or records) in a dataset. | One | M |  |
| exampleIdentifeir | Text | An example identifier used by one item (or record) from a dataset. | One | M |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Existing properties in schema.org/<Type>** | | | | | |
| dataset | Dataset | A dataset contained in this catalog.  Inverse property: includedInDataCatalog. | Many | M |  |
| identifier | PropertyValue or  Text or  URL | The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details. | Many | M |  |
| name | Text | The name of the item. | One | M |  |
| description | Text | A description of the item. | One | M |  |
| url | URL | URL of the item. | One | M |  |
| keywords | Text | Keywords or tags used to describe this content. Description of the existing Property. | One | M | CV |
| provider | Organization or Person | The service provider, service operator, or service performer; the goods producer. Another party (a seller) may offer those services or goods on behalf of the provider. A provider may also serve as the seller. Supersedes carrier. | Many | M | ? |
| alternateName | Text | An alias for the item. | Many | R |  |
| license | CreativeWork or URL | A license document that applies to this content, typically indicated by URL. | One | R |  |
| citation | CreativeWork or Text | A citation or reference to another creative work, such as another publication, web page, scholarly article, etc. | Many | R |  |
| sourceOrganization | Organization | The Organization on whose behalf the creator was working. | Many | R |  |
| dateModified | Date or DateTime | The date on which the CreativeWork was most recently modified or when the item's entry was modified within a DataFeed. | One | R |  |
| publication | PublicationEvent | A publication event associated with the item. | Many | R |  |
| datePublished | Date | Date of first broadcast/publication. | One | O |  |
| fileFormat | Text or URL | [Media type, typically MIME format (see IANA site) of the content e.g. application/zip of a SoftwareApplication binary. In cases where a CreativeWork has several media type representations, 'encoding' can be used to indicate each MediaObject alongside particular fileFormat information. Unregistered or niche file formats can be indicated instead via the most appropriate URL, e.g. defining Web page or a Wikipedia entry.](http://www.iana.org/assignments/media-types/media-types.xhtml) | Many | O |  |

A full mapping between Schema.org, bioschemas, identifiers.org and BioSharing.org can be found [here](https://docs.google.com/spreadsheets/d/1H12h5VpVNJFzNs2RQJWjXkauCEn3qEsVFzKRoiHHffY/edit#gid=1261485211).

## **Implementation Guidelines**

Schema.org [suggests](http://schema.org/docs/gs.html) implementing metadata using Microdata, RDFa, or JSON-LD. Depending on the context, any of these can be used for embedding the Data Repository metadata on the repository provider’s landing web page. BioSchema community recommends using JSON-LD.

### **JSON-LD**

[JSON-LD](https://en.wikipedia.org/wiki/JSON-LD) (JavaScript Object Notation for Linked Data), is a method of transporting [Linked Data](https://en.wikipedia.org/wiki/Linked_Data) using [JSON](https://en.wikipedia.org/wiki/JSON). Example below represents an Data Repositories described in JSON-LD format.

|  |
| --- |
| **Example 5. Representing an Data Repositories in JSON-LD format** |
| <script type="application/ld+json">  {  "@context": "http://schema.org",  "@type": "DataCatalog",  "identifier":"https://identifiers.org/MIR:00100800",  "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/",  "dateModified": "June 6, 2017",  "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"  }  ],  "license": {  "@type":"CreativeWork",  "name": "Creative Commons CC4 Attribution",  "url": "https://creativecommons.org/licenses/by/4.0/"  }  ,  "publication":[  {  "@type":"PublicationEvent",  "name": "Uniform Resolution of Compact Identifiers for Biomedical Data",  "url": "http://biorxiv.org/content/early/2017/01/18/101279"  },  {  "@type":"PublicationEvent",  "name": "Identifiers.org and MIRIAM Registry: community resources to provide persistent identification.",  "url": "https://identifiers.org/pubmed:22140103"  }  ],  "sourceOrganization": {  "@type":"Organization",  "name": "The European Bioinformatics Institute (EMBL-EBI)",  "url": "http://www.ebi.ac.uk/"  }  ,  "alternateName": ["Identifiers.org Registry", "MIRIAM Registry"],  "accessURL":"https://identifiers.org/",  "identifierPattern":"^MIR:\\d{8}$",  "exampleIdentifier":"MIR:00100037"  }  </script> |

For more information, please refer to the [JSON-LD specification](http://www.w3.org/TR/json-ld/).

# Glossary

Describe here any specific terms that need further explanation.

# Further examples

Insert here more examples showing how to use this specification in real life cases.