Bioschemas Dataset

**!! DRAFT - Work in progress !! Version of May 25th 2017**

Summary

This document describes a set of fields (properties, types and description) for discovering datasets, organized by their requirement level and driven by a set of use cases. The majority of these fields already exist in the [schema.org Dataset](http://schema.org/Dataset), others are pending. Note that this is just a draft that will be iteratively reviewed especially with the other Bioschemas groups.

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/1XzrZxFIuG3TS9RU8vACoUjAvaADLmI_FrIk7O3BEkxY/edit?usp=sharing).

# 

# Specification

## Mandatory Fields

The mandatory fields are the minimum set (MUST) that is most relevant criteria to discover datasets, and within Bioschemas we have agreed to keep them to a maximum of six.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mandatory Fields - MUST** | | | |
| ***Property*** | ***Expected Type*** | ***Description***  ***(source: schema.org Dataset)*** | ***Presence in schema.org Dataset*** |
| 1. name | Text | A descriptive name of a dataset | yes |
| 2. description | Text | A short summary describing a dataset | yes |
| 3. url | URL | Location of a page describing the dataset | yes |
| 4. keywords | Text | Keywords summarizing the dataset | yes |
| 5. includedInDataCatalog | DataCatalog | The catalog to which this dataset belongs to | yes |
| 6. identifier | PropertyValue, Text, URL | Any kind of identifier for any kind of thing | [pending](http://pending.webschemas.org/identifier) |

## Recommended Fields

The recommended fields are optional (SHOULD), but providing them is encouraged because they enhance the discoverability.

|  |  |  |  |
| --- | --- | --- | --- |
| **Recommended Fields - SHOULD** | | | |
| ***Property*** | ***Expected Type*** | ***Description***  ***(source: schema.org Dataset)*** | ***Presence in schema.org Dataset*** |
| 7. creator | Text | The name of the dataset creator (person or organization) | yes |
| 8. version | Text, Number | The version number for this dataset | yes |
| 9. variablesMeasured | Text, PropertyValue | What does the dataset measure? (e.g., temperature, pressure) | [pending](http://pending.webschemas.org/variableMeasured) |
| 10. measuramentTechnique | Text | A technique or technology used for measuring the corresponding variable(s) (described using variablesMeasured) | [pending](http://pending.webschemas.org/measurementTechnique) |
| 11. citation | Text | A citation for a publication that describes the dataset | yes |
| 12. license | Text, URL | A license under which the dataset is distributed | yes |

## Example (not to use - to be updated)

|  |
| --- |
|  |
| <script type="application/ld+json">  {  **"@context"**: **"http://schema.org/"**,  **"@type"**: **"Dataset"**,  **"identifier"**:**"MTBLS234"**,  **"url"**: **"http://www.ebi.ac.uk/metabolights/MTBLS234"**,  **"name"**: **"Automated Label-free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data (Plasma)"**,  **"dateCreated"**: **"2015-09-23T"**,  **"datePublished"**: **"2016-12-15T"**,  **"dateModified"**: **"2016-12-15T"**,  **"description"**: **"Liquid chromatography coupled to mass spectrometry (LC-MS) has become a standard technology in metabolomics...."**,  **"fileFormat"**:**"ISA-Tab"**,  **"license"**:**"<SOME\_URL>"**,  **"Author"**:{  **"@type"**: **"Person"**,  **"familyName"**: **"Kohlbacher"**,  **"givenName"**: **"Oliver"**  },  **"locationCreated"**: {  **"@type"**: **"Place"**,  **"name"**: **"Wilhelm Schickard Institute for Computer Science"**,  **"address"**: {  **"@type"**: **"PostalAddress"**,  **"streetAddress"**: **"University of Tübingen"**,  **"addressCountry"**: **"Germany"**,  **"addressRegion"**: **"Tübingen"**,  **"postalCode"**: **""**  },  **"geo"**: {  **"@type"**: **"GeoCoordinates"**,  **"latitude"**: **"TBD"**,  **"longitude"**: **"TDB"**  }  },  **"variableMeasured"**: **"metabolite concentration"**,  **"keywords"**: [  **"Mass spectrometry"**, **"plasma"**,**"spike-in"**,**"control"**  ],  **"distribution"**: {  **"@type"**: **"DataDownload"**  },  **"includedinDataCatalog"**: {  **"name"**: **"EMBL-EBI Metabolights"**  }  }  </script> |
|  |

# Motivation

Most dataset repositories and registries do not provide structured data easily crawlable by search engines. Registries like DataMed, OMICsDI and BioSamples do automated ingestion of content mainly through APIs but not all the data repositories have a programmatic interface and the existing variety of programmatic interfaces are subject to changes which break integration workflows.

The purpose of this Bioschemas data group is to:

* Facilitate the ingestion of datasets metadata from data repositories (databases) into search engines and dataset registries like OMICsDI and DataMed via Bioschemas;
* Automate the linking of datasets metadata to samples in dataset registries like Biosamples, and identify cases where samples are missing or metadata is absent;
* Engage and help data providers to test and adopt the exposure of dataset metadata Bioschemas;
* Contribute to increase the number of indexed data repositories via Bioschemas;
* Make dataset registries compliant with Bioschemas.

# Use Cases

The overall aim is findability and not structured queries. Based on this, the key use cases that have driven the selection of the fields are the following:

* Search on a title of dataset
* Enable discovery by indexing on free text description
* Enable direct access, resolution of dataset
* Allow restriction to specific dimensions and variables specifically recorded in a dataset (e.g. get all climate datasets which have monitors CO2 concentration or datasets where metabolite concentration was recorded)
* Search according to the repository of the datasets
* Search according to creator, author
* Allow restriction to how a dataset has been generated (e.g. get all climate datasets if metabolite concentration was recorded if acquired using mass spectrometry)
* ….

# Reference Material

Refer to the [crosswalk](https://docs.google.com/spreadsheets/d/1XzrZxFIuG3TS9RU8vACoUjAvaADLmI_FrIk7O3BEkxY/edit?usp=sharing) file where some of these are mapped to schema.org Dataset.

* [NIH BD2K DataMed DATS annotated with schema.org](https://biocaddie.org/group/working-group/working-group-3-descriptive-metadata-datasets)
* [OmicsDI model](http://biorxiv.org/content/early/2016/04/18/049205)
* [DataCite schemas](http://schema.datacite.org/)
* [A Data Citation Roadmap for Scholarly Data Repositories](http://biorxiv.org/content/early/2016/12/28/097196)
* [Automating the collection of “data repositories” metadata](https://docs.google.com/document/d/1N9-FTKL1GX2kmndIaSjbiJBxi0oykRYdJUjDkN2bFjE/edit)
* [Schema.org biological dataset via Bioschemas.org](https://docs.google.com/document/d/1N9-FTKL1GX2kmndIaSjbiJBxi0oykRYdJUjDkN2bFjE/edit)
* [Porting JATS into Schema.org](https://docs.google.com/document/d/1N9-FTKL1GX2kmndIaSjbiJBxi0oykRYdJUjDkN2bFjE/edit)
* [Google science dataset documentation](https://developers.google.com/search/docs/data-types/datasets)
* [Ontology-based Dataset Exploration](http://scim.brad.ac.uk/~dthakker/event/iesd2016/iesd16_submission_2.pdf)
* [The healthcare and life sciences community profile for dataset descriptions](https://peerj.com/articles/2331/)
* …..