

Dataset description with DataID

DataID is a way of describing datasets including content, authors, licenses and provenance. DataID extends standards such as DCAT and is a precondition to enhance discoverability, automatic data processing and to explore the data itself.

Target audience

Dataset creators and publishers wanting to describe dataset content and location. Basic knowledge of linked data technology (URLs, RDF format) is needed. Must be able to upload the DataID together with the described dataset.

Scope

This reference card describes basic concepts of DataID descriptions, their creation, distribution and common use-cases.

How to create a DataID

- ❑ visit <http://dataid.dbpedia.org/>
- ❑ Use the Generator to enter your metadata. Try to be as verbose as possible, to ensure good documentation of your data
- ❑ Take care to include links to your actual data: Link the data files, not your website
- ❑ Upload the metadata to datahub with the integrated “Upload to datahub” feature
- ❑ Download the DataID file and upload it on your server, either next to the data files or next to the robots.txt: <http://example.org/dataid.json>
- ❑ Maintain your metadata as you maintain your dataset. Data quality depends quality metadata

Dataset: Main Dataset

The screenshot shows the DataID Generator web form. It is titled "Dataset: Main Dataset". The form is divided into several sections:

- General Information:** Contains fields for "Dataset/Subset URI", "Title", "Label", "Homepage", "Example Resource", "Language", "Root Resource", and "Ontology Location". Each field has a corresponding Data ID (e.g., dct:title, rdfs:label, dcat:landingPage, void:exampleResource, dct:language, void:rootResource, dcat:ontologyLocation).
- Description and rights:** Contains fields for "Issued date", "Version Info", "Description", "License", "Rights", and "Keywords". Each field has a corresponding Data ID (e.g., dct:issued, dataid:versionInfo, dct:description, owl:license, dct:rights, dcat:keyword).
- Contact Information:** Contains sections for "Contact Information", "Provenance", "WasDerivedFrom", and "WasGeneratedBy". Each section has a table with columns for "Contact Label", "Contact Property", "Edit", and "Remove". Each section also has an "Add Contact" button.

DataID Generator

Example

The following example provides a DataID on the right and a description of the concepts, resources and properties on the left. **Bold properties are mandatory** for valid DataIDs.

• DatasetDescription

This resource describes the DataID itself and links to the dataset descriptions.

• Dataset

Datasets are resources that describe the conceptual datasets, i.e. the metadata describing the dataset. They contain name and label of the dataset, date of modification, license and contact information. They also link at least one data distribution.

• License

The dc:license of the dataset should be stated by an [ODRL](http://purl.org/NET/rdflicense) resource. ODRL resources for most licenses can be found in this repository: <http://purl.org/NET/rdflicense>

Additional clarification using dc:rights can be used to cite scientific publications or special conditions.

• Provenance and contact

prov:wasDerivedFrom should be used to link to the original data which the dataset was derived from. It can either link another DataID dataset resource or provide textual description.

At least one associated agent must be provided with name and a contact address of some sort.

• Additional metadata

Triple count, example resource and other VoID properties may be added.

• Distribution

Distributions describe the actual, physical data aggregation, i.e. the data itself. They can describe any file (like zipped dump files) or service (like a SPARQL endpoint) identified by a URI.

They must be described with name, dates and download URL. The MediaType describes the mime-type of the resource, format describes the inner file format if applicable.

```
<http://example.org/DataID_URI>
  a void:DatasetDescription;
  foaf:primaryTopic <http://example.org/DataID_URI#dataset1>.
```

```
<http://example.org/DataID_URI#dataset1>
  a dataid:Dataset ;
  dc:title "Detailed Dataset Name" ;
  rdfs:label "Short Dataset Name" ;
  dc:description "Descriptive text" ;
  dc:modified "10-28-2014"^^xsd:date ;

  dataid:licenseName "Name of the license" ;
  dc:license <http://purl.oclc.org/NET/rdflicense/cc-by-sa> ;
  dc:rights "Further clarification of license / citations" ;
  void:subset <http://example.org/DataID_URI#links> ;
  dc:language "Dataset language" ;
  void:vocabulary <http://example.org/ontology> ;
  dcat:landingPage <http://example.org/homepage> ;
  dcat:distribution <http://example.org/DataSet_URI> ;

  prov:wasDerivedFrom [
    a dataid:Entity ;
    dcat:landingPage <http://example.org/homepage> ;
  ] ;
  dataid:associatedAgent [
    a dataid:Maintainer ;
    dataid:agentName "Maintainer Name" ;
    dataid:agentEmail "Maintainer email address"
  ] ;

  void:triples "Triplecount" ;
  void:exampleResource <http://example.org/DataSet_URI> .
```

```
<http://example.org/DataSet_URI>
  a dataid:Distribution ;
  dc:modified "10-28-2014"^^xsd:date ;
  dc:title "Name of the resource (i.e. 'Dataset RDF dump')" ;
  dc:description "Describe distribution if necessary" ;
  dcat:downloadURL <http://example.org/DataSet_URI> ;
  dc:format "text/turtle" ;
  dcat:mediaType "application/gzip".
```

After creation, upload the DataID in turtle or jsonld format on your server, either next to the data files or next to the robots.txt: <http://example.org/dataid.json>