



4th Summer Datathon on Linguistic Linked Open  
Data  
Cercedilla, Spain  
30th May – 3 June 2022



# Metadata

Andon Tchechmedjiev (IMT Mines Alès, EuroMov Digital Health in Motion)

- ▶ EU Open Data Support
  - General Introduction: <http://bit.ly/2VjCNt8>
  - Metadata and licensing: <http://bit.ly/2E5wqnL>
  - Quality of metadata: <http://bit.ly/2VzDTpC>
- ▶ DCAT Specification: <http://bit.ly/2JBY7bp>
- ▶ VOID Specification: <http://bit.ly/2LEXGjf>
  - <https://www.slideshare.net/cygri/void-metadata-for-rdf-datasets>
- ▶ Lime extension specification: <http://bit.ly/2LEj47W>
- ▶ Meta-share documentation: <http://bit.ly/30qCqRg>
- ▶ LOD-cloud: <https://lod-cloud.net/>

# OUTLINE

1. WHAT IS METADATA?
2. MANAGING METADATA
3. SOME COMMON  
METADATA  
VOCABULARIES FOR  
LINGUISTIC RESOURCES

# WHAT IS METADATA?

DEFINITION, EXAMPLES AND REUSABLE STANDARDS.

## A general definition

“Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.”

-- National Information Standards Organization

<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

Metadata provides information enabling to make sense of data (e.g. documents, images, datasets), concepts (e.g. classification schemes) and real-world entities (e.g. people, organisations, places, paintings, products).

## Types of metadata

- ▶ **Descriptive metadata**, describe a resource for purposes of discovery and identification.
- ▶ **Structural metadata**, e.g. data models and reference data.
- ▶ **Administrative metadata**, provides information to help manage a resource.

What we will focus on is descriptive metadata that will help you annotate your datasets.

By this definition, the Ontolex model itself could be considered a form of structural metadata.

# WHAT IS METADATA?

7

## Examples of metadata

label

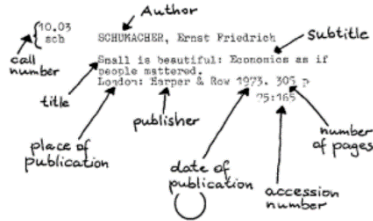


Provides metadata on

can



card



book



## Dataset description (DCAT)

```
:weather1-7 a dcat:Dataset ;
dct:title "Measurements from weather stations 1-7" ;
dct:description "Data from seven weather stations
showing temperature, humidity,
wind direction and wind speed" ;
dct:modified "2013-07-01" ;
dct:publisher <http://myweather.com/id/myweather> ;
dct:keyword "weather" ;
dcat:landingpage <http://myweather.com/stations1-7.html> ;
dcat:distribution :weatherdata-xlsx .

:weatherdata1-7-xlsx a dcat:Distribution ;
dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX>
dct:licence <http://creativecommons.org/licenses/CC0> ;
dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
```

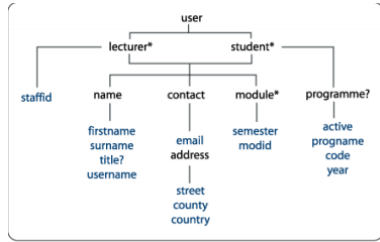
dataset

	Temp. °C	Humidity %	Wind direction	Wind speed km/h
Station 1	18.1	60	WSW	18
Station 2	17.5	59	WSW	20
Station 3	18.2	55	SW	22
Station 4	19.0	62	SW	18
Station 5	18.0	65	WSW	19
Station 6	18.2	63	SSW	21
Station 7	17.9	61	SW	22

# WHAT IS METADATA?

8

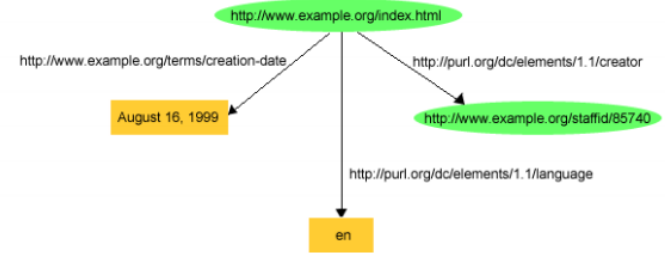
## Two approaches to provide metadata on the web



XML

```
<?xml version="1.0"?>
<!DOCTYPE user SYSTEM "users.dtd">
<user>
  <student>
    <name>
      <firstname>Joe</firstname>
      <surname>Smith</surname>
      <title>Mr.</title>
      <username>smithj</username>
    </name>
    <contact>
      <address>
        <street>54 Maple Rise, Santry</street>
        <county>Dublin</county>
        <country>Ireland</country>
      </address>
      <email>smithj@dcu.ie</email>
    </contact>
    <programme active="true">
      <progrname>M.Eng in Electronic Systems</progrname>
      <code>MEN</code>
      <year>1</year>
    </programme>
    <module semester="2">
      <modid>EE557</modid>
    </module>
    <module semester="1">
      <modid>EE553</modid>
    </module>
  </student>
```

RDF



ex:index.html	dc:creator	exstaff:85740 .
ex:index.html	exterms:creation-date	"August 16, 1999" .
ex:index.html	dc:language	"en" .



# MANAGING METADATA

## Managing your metadata is important

### Metadata needs to be managed to ensure:

- ▶ **Availability:** metadata needs to be stored where it can be accessed and indexed so it can be found.
- ▶ **Quality:** metadata needs to be of consistent quality so users know that it can be trusted.
- ▶ **Persistence:** metadata needs to be kept over time.
- ▶ **Open License:** metadata should be available under a public domain license to enable its reuse.

### The metadata lifecycle is larger than the data lifecycle:

- ▶ Metadata may be **created before data is created** or captured, e.g. to inform about data that will be available in the future.
- ▶ Metadata needs to be **kept after data has been removed**, e.g. to inform about data that has been decommissioned or withdrawn.

“A labelling, tagging or coding system used for recording cataloguing information or structuring descriptive records. A metadata schema establishes and defines data elements and the rules governing the use of data elements to describe a resource.”

XML  
Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCTYPE RDF>
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:rd="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/dc/"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  - <rd:Description rdf:about="http://purl.org/dc/terms/"
    <dc:terms:title xml:lang="en">DCMI Metadata Terms - other</dc:terms:title>
    <dc:terms:publisher rdf:resource="http://purl.org/dc/aboutdc#DCMI"/>
    <dc:terms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2012-06-14</dc:terms:modified>
  </rd:Description>
  - <rd:Description rdf:about="http://purl.org/dc/terms/title">
    <rdfs:label xml:lang="en">Title</rdfs:label>
    <rdfs:comment xml:lang="en">A name given to the resource.</rdfs:comment>
    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/terms/" />
    <dc:terms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2008-01-14</dc:terms:issued>
    <dc:terms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2010-10-11</dc:terms:modified>
    <rd:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
    <dc:terms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#title-002"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
    <rdfs:subPropertyOf rdf:resource="http://purl.org/dc/elements/1.1/title"/>
  </rd:Description>
  - <rd:Description rdf:about="http://purl.org/dc/terms/creator">
    <rdfs:label xml:lang="en">Creator</rdfs:label>
    <rdfs:comment xml:lang="en">An entity primarily responsible for making the resource.</rdfs:comment>
    <dc:terms:description xml:lang="en">Examples of a Creator include a person, an organization, or a service.</dc:terms:description>
    <rdfs:isDefinedBy rdf:resource="http://purl.org/dc/terms/" />
    <dc:terms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2008-01-14</dc:terms:issued>
    <dc:terms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2010-10-11</dc:terms:modified>
    <rd:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
    <dc:terms:hasVersion rdf:resource="http://dublincore.org/usage/terms/history/#creatorT-002"/>
    <rdfs:range rdf:resource="http://purl.org/dc/terms/Agent"/>
    <rdfs:subPropertyOf rdf:resource="http://purl.org/dc/elements/1.1/creator"/>
    <rdfs:subPropertyOf rdf:resource="http://purl.org/dc/terms/contributor"/>
    <owl:equivalentProperty rdf:resource="http://xmlns.com/foaf/0.1/maker"/>
  </rd:Description>
  - <rd:Description rdf:about="http://purl.org/dc/terms/subject">
```

RDFS

## Reusing existing vocabularies to define metadata

### General purpose standards and specifications

- ▶ **Dublin Core**: for published material (text, images),  
<http://dublincore.org/documents/dcmi-terms/>
- ▶ **FOAF**: for people and organisations, <http://xmlns.com/foaf/spec/>
- ▶ **SKOS**: for concept collections, <http://www.w3.org/TR/skos-reference>
- ▶ Data Catalog Vocabulary **DCAT**, <http://www.w3.org/TR/vocab-dcat/>
  - **ADMS**: for interoperability assets in e-government systems,  
<http://www.w3.org/TR/vocab-adms/>
  - **DCAT-AP**:

RDF schema is particularly good in combining terms from different standards and specifications.

- ▶ **Do not re-invent** terms that are already defined somewhere else, when designing RDF schemas – reuse terms where possible.

# SOME COMMON METADATA VOCABULARIES FOR LINGUISTIC RESOURCES

## FOAF (Friend Of A Friend)

@prefix dbr: <http://dbpedia.org/resource/> .  
@prefix dbo: <http://dbpedia.org/ontology/> .  
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  
@prefix foaf: <http://xmlns.com/foaf/0.1/> .  
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84\_pos#> .  
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .  
@prefix schema: <http://schema.org/> .

### dbr:Bob\_Marley

a foaf:Person ;  
rdfs:label "Bob Marley"@en ;  
rdfs:label "Bob Marley"@fr ;  
rdfs:seeAlso dbr:Rastafari ;  
dbo:birthPlace dbr:Jamaica .

### dbr:Jamaica

a schema:Country ;  
rdfs:label "Jamaica"@en ;  
rdfs:label "Giamaica"@it ;  
geo:lat "17.9833"^^xsd:float ;  
geo:long "-76.8"^^xsd:float ;  
foaf:homepage <http://jis.gov.jm/> .

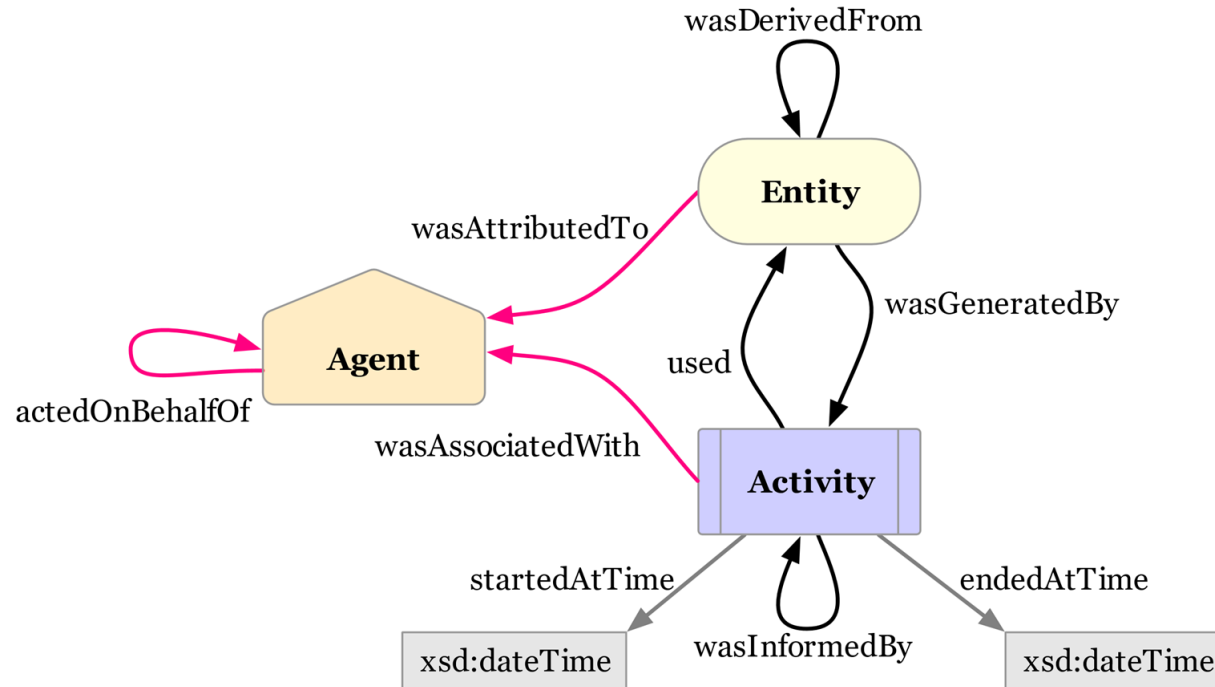
## DUBLIN CORE

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://www.w3schools.com">
    <dc:title>D-Lib Program</dc:title>
    <dc:description>W3Schools - Free tutorials</dc:description>
    <dc:publisher>Refsnes Data as</dc:publisher>
    <dc:date>1999-09-01</dc:date>
    <dc:type>Web Development</dc:type>
    <dc:format>text/html</dc:format>
    <dc:language>en</dc:language>
  </rdf:Description>
</rdf:RDF>
```

### Dublin Core Elements

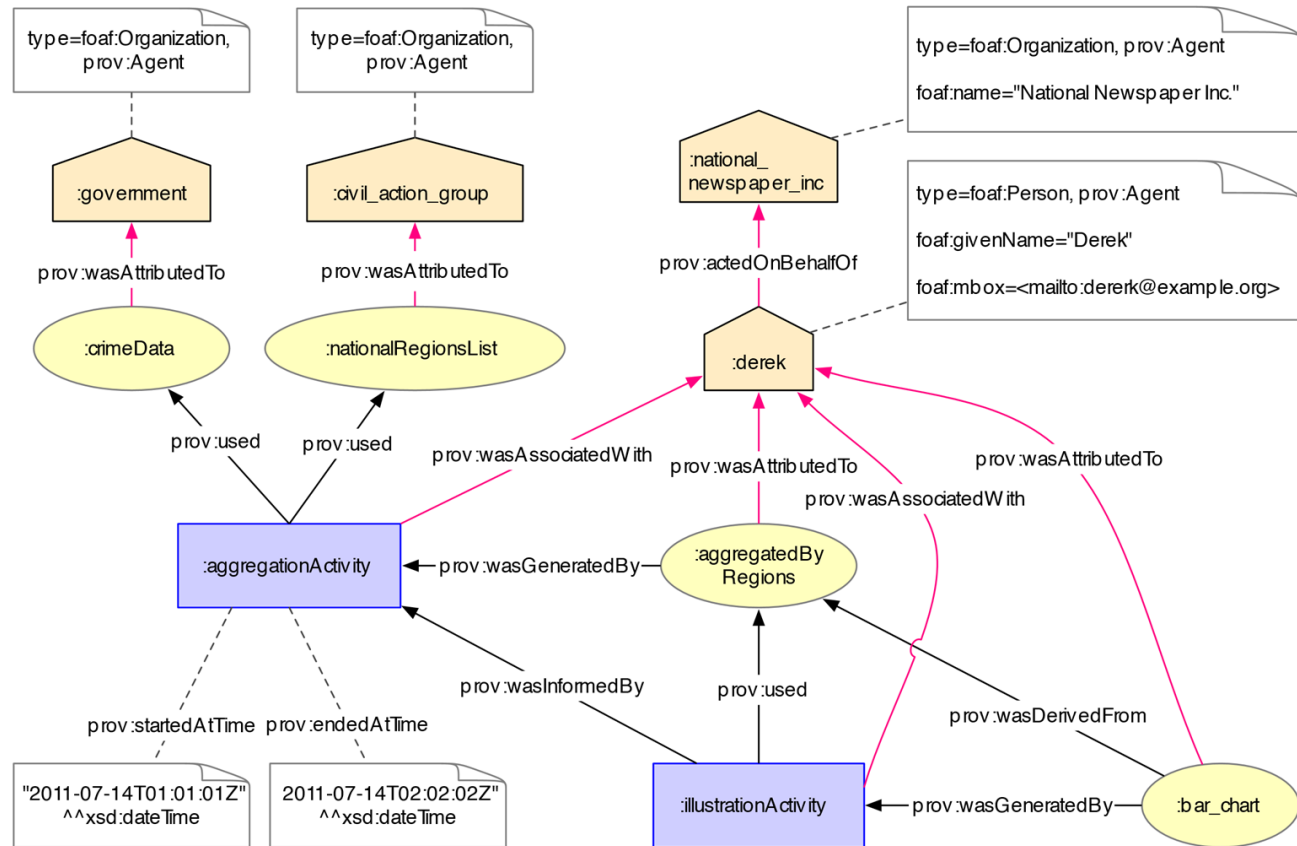
Rights	Contributor	Creator
Subject	Coverage	Title
Publisher	Identifier	Description
Type	Date	Source
Relation	Format	Language

## PROVO (PROVenance Ontology)

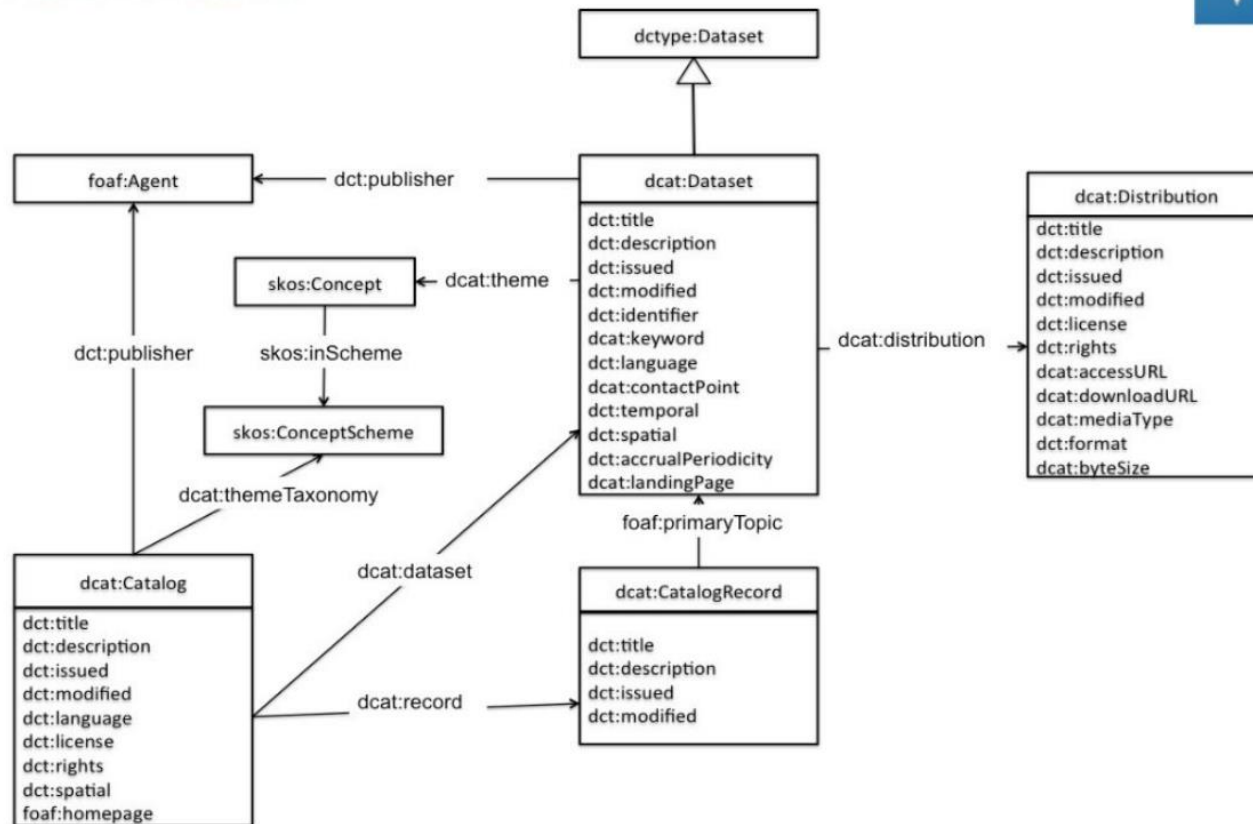




## PROVO (PROVenance Ontology)



## W3C DCAT



## DCAT

### Description of the Catalogue

```
:catalog
  a dcat:Catalog ;
  dct:title "Imaginary Catalog" ;
  rdfs:label "Imaginary Catalog" ;
  foaf:homepage <http://example.org/catalog> ;
  dct:publisher :transparency-office ;
  dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
  dcat:dataset :dataset-001 , :dataset-002 , :dataset-003 ;
  .
```

### Description of the Dataset

```
:dataset-001
  a dcat:Dataset ;
  dct:title "Imaginary dataset" ;
  dcat:keyword "accountability", "transparency" , "payments" ;
  dct:issued "2011-12-05"^^xsd:date ;
  dct:modified "2011-12-05"^^xsd:date ;
  dct:publisher :finance-ministry ;
  dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
  dcat:distribution :dataset-001-csv ;
  .
```

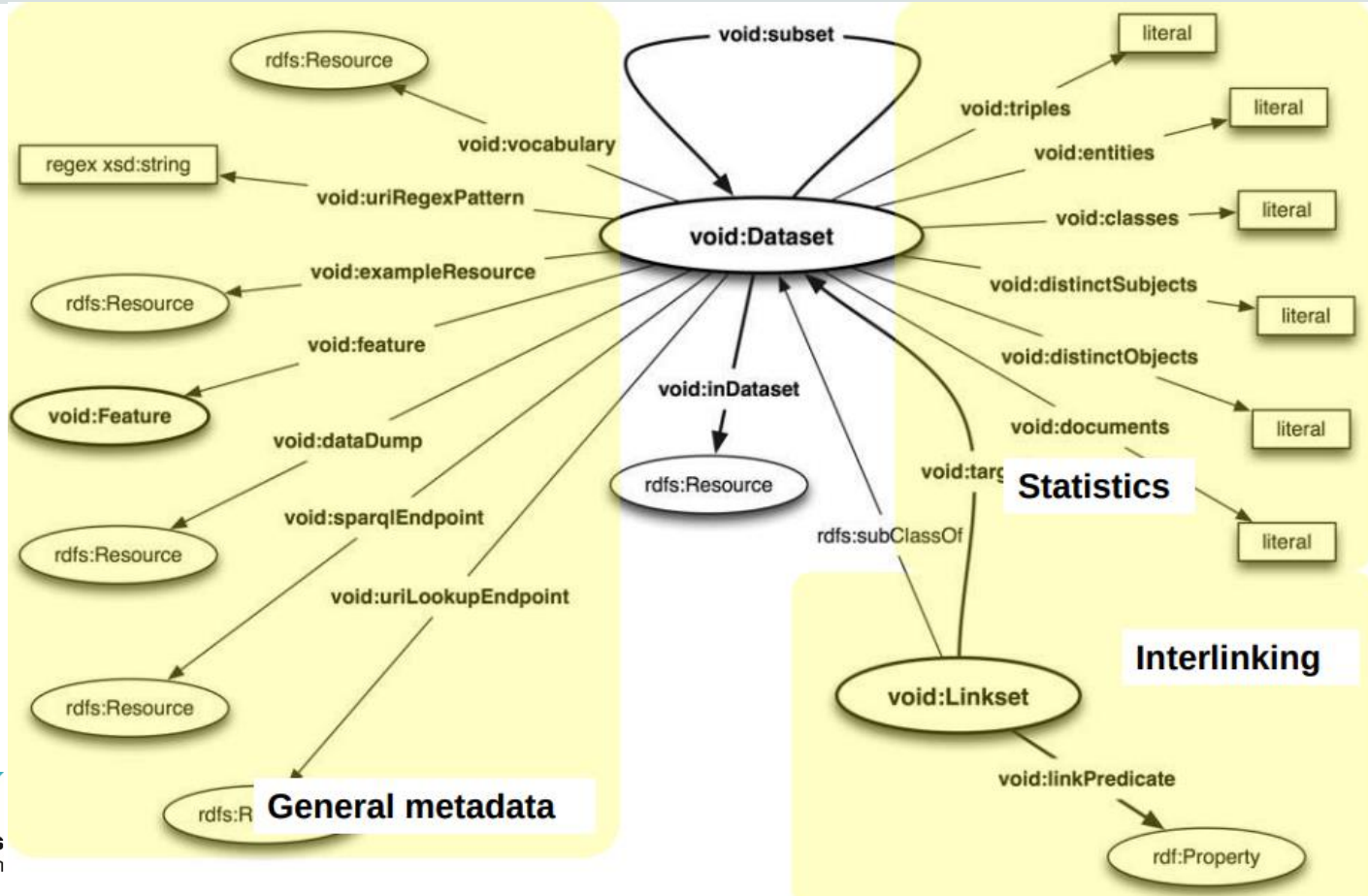
### Description of the Distribution

```
:dataset-001-csv
  a dcat:Distribution ;
  dcat:downloadURL <http://www.example.org/files/001.csv> ;
  dct:title "CSV distribution of imaginary dataset 001" ;
  dcat:mediaType "text/csv" ;
  dcat:byteSize "5120"^^xsd:decimal ;
  .
```

# SOME COMMON METADATA VOCABULARIES

20

VOID (Vocabulary Of Interlinking Datasets)



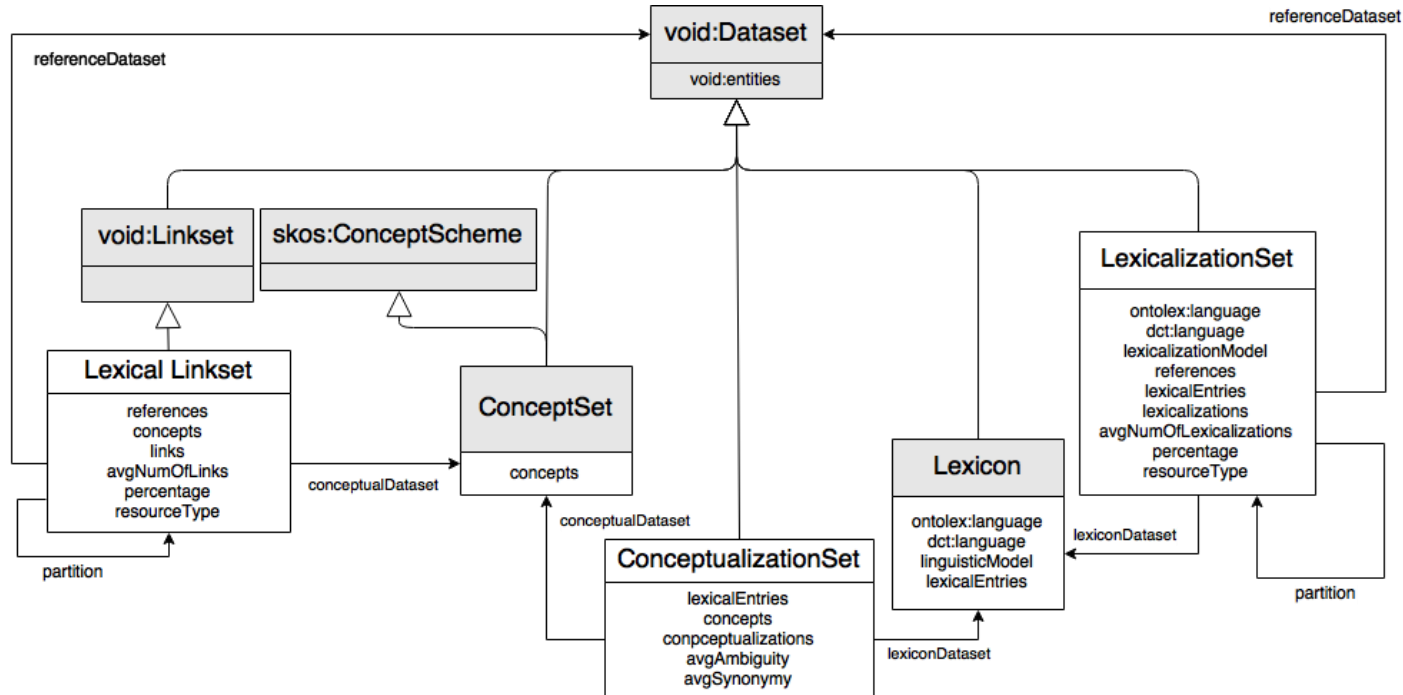
## VOID (Vocabulary Of Interlinking Datasets)

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:sc="http://sw.deri.org/2007/07/sitemapextension/scschema.xsd"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:scovo="http://purl.org/NET/scovo#"
  xmlns:void="http://rdfs.org/ns/void#">
  <void:Dataset rdf:about="http://example.com/catalog.rdf#catalog">
    <dcterms:comment>Example Corp. Product Catalog</dcterms:comment>
    <foaf:homepage rdf:resource="http://example.org/dataset.html"/>
    <dcterms:subject rdf:resource="http://dbpedia.org/resource/EXAMPLE"/>
    <dcterms:creator rdf:resource="http://example.org/creator#me"/>
    <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/3.0/">
    <void:statItem rdf:parseType="Resource">
      <scovo:dimension rdf:resource="http://rdfs.org/ns/void#numOfTriples"/>
      <rdf:value rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">0</void:value>
    </void:statItem>
    <void:exampleResource rdf:resource="http://example.com/products/widgets/X42"/>
    <void:exampleResource rdf:resource="http://example.com/products/categories/all"/>
    <void:sparqlEndpoint rdf:resource="http://example.com/sparql"/>
    <void:dataDump rdf:resource="http://example.com/data/catalogdump.rdf.gz"/>
    <void:dataDump rdf:resource="http://example.org/data/catalog_archive.rdf.gz"/>
    <void:dataDump rdf:resource="http://example.org/data/product_categories.rdf.gz"/>
    <void:uriPattern>^http://example.com/products/$</void:uriPattern>
  </void:Dataset>
</rdf:RDF>
```

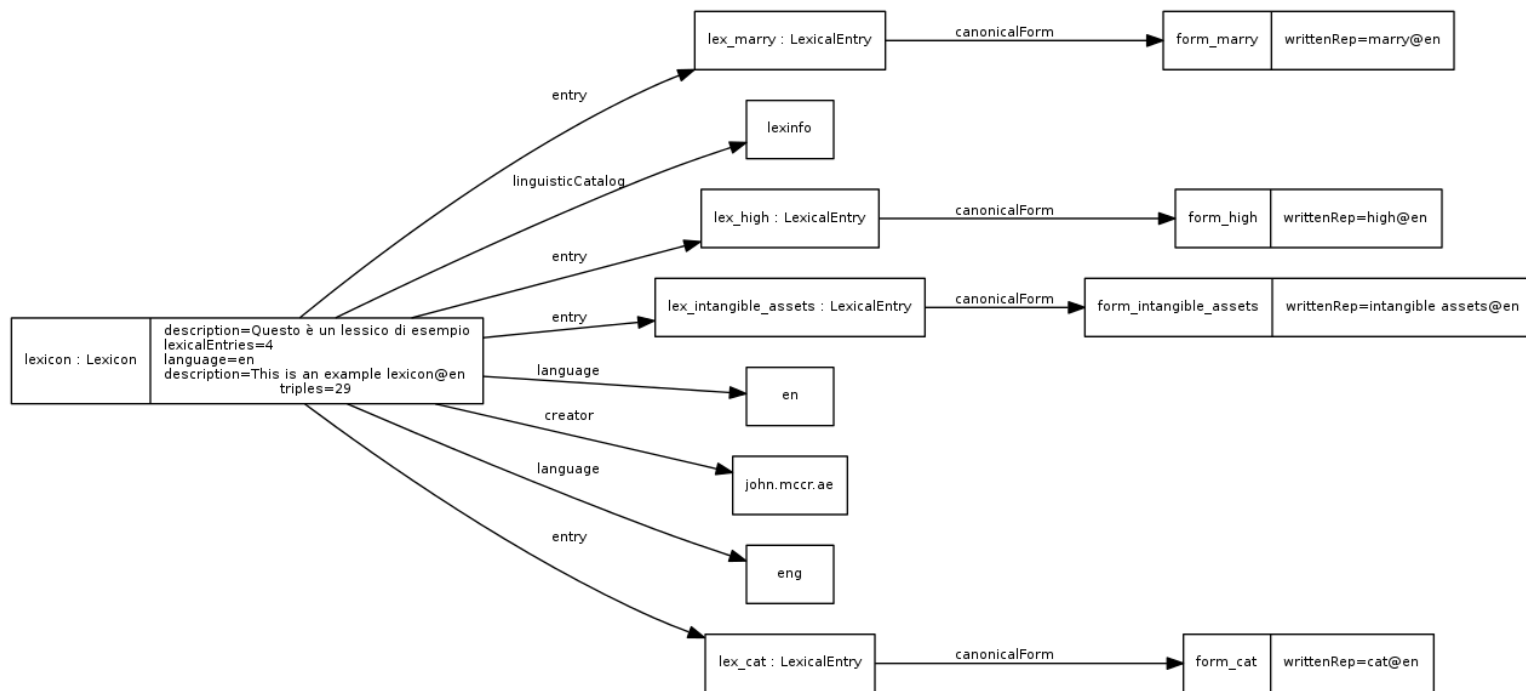
### Specifies SPARQL Endpoint capabilities

- ▶ Some properties pertinent to describe named graphs and datasets,
  - **sd:Dataset** Equivalent class in DCAT, VOID, Schema.org and PROV-O
  - **sd:defaultGraph** property of **sd:Dataset**, gives URI of default graph (**sd:Graph**)
  - For an **sd:Graph**, the last modification date can be expressed with **dct:modified**
  - **sd:namedGraph** property of **sd:Dataset**, gives URI of a particular **sd:NamedGraph**
  - A **sd:NamedGraph** can further be specified with the property **sd:name**
  - An **sd:Service** (similar to a **dcat:DataService**) describes a SPARQL endpoint
  - The URI of the service is defined with **sd:endpoint**
  - Available graphs and the default dataset can be respectively specified by **sd:availableGraphs** and **sd:defaultDataset**

## LIME (LInguistic MEtadata)



## LIME (LInguistic MEtadata)

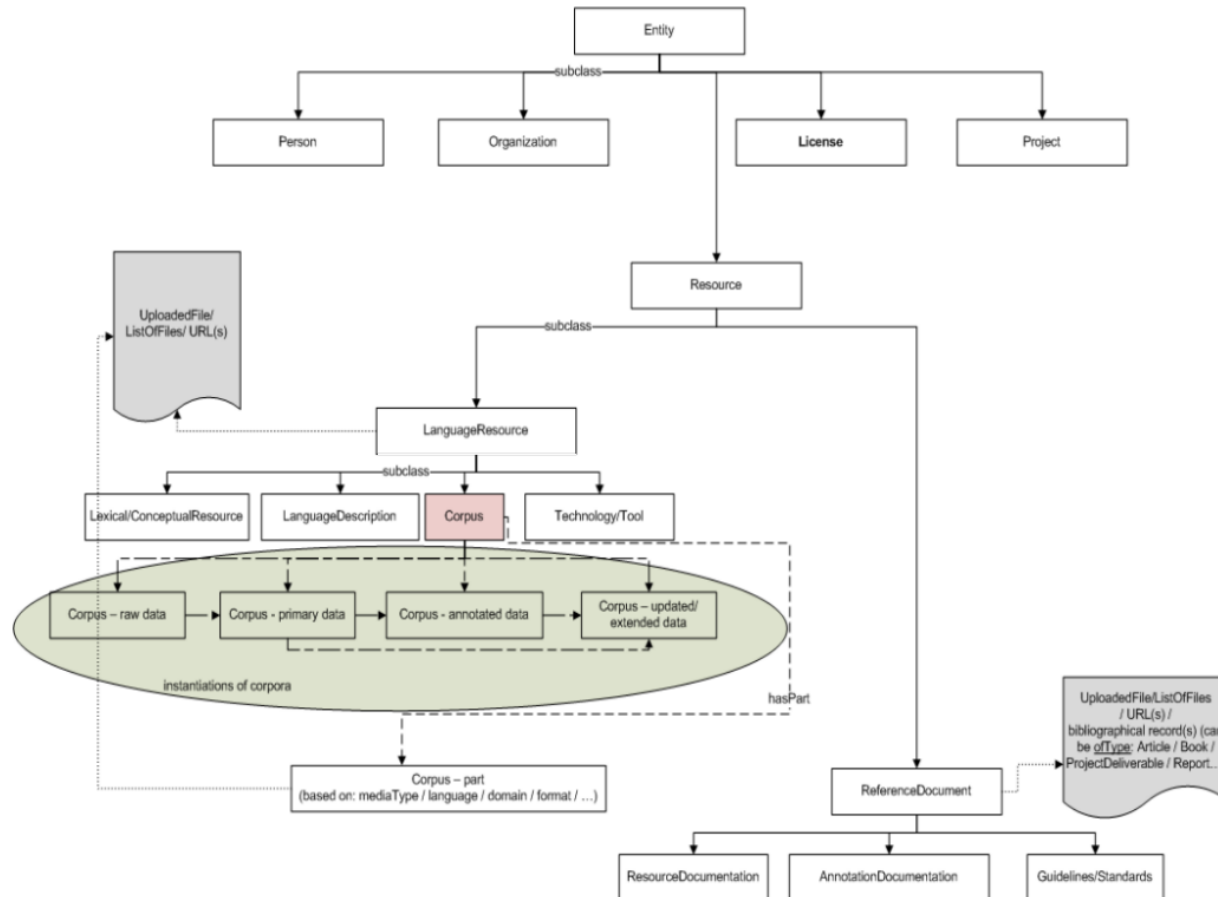




# SOME COMMON METADATA VOCABULARIES

25

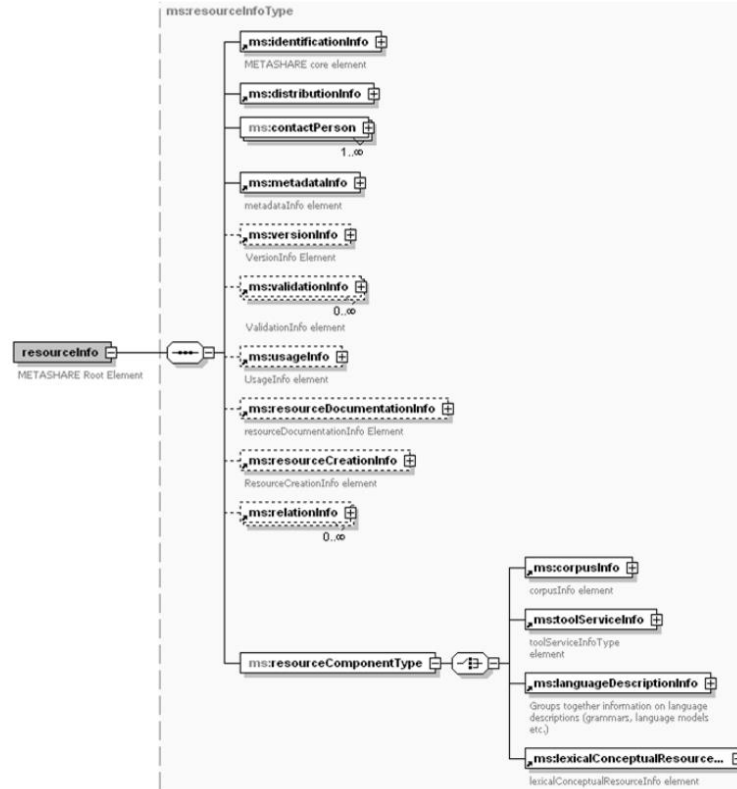
Meta-Share



# SOME COMMON METADATA VOCABULARIES

26

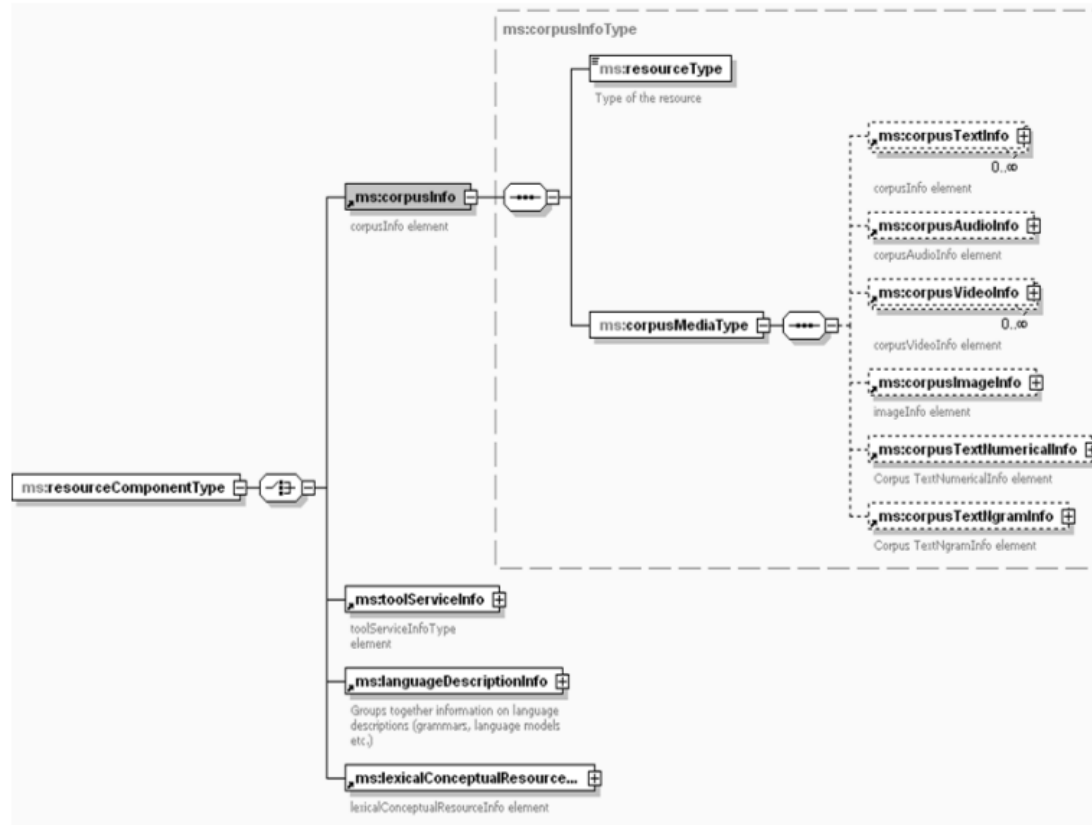
## Meta-Share



# SOME COMMON METADATA VOCABULARIES

27

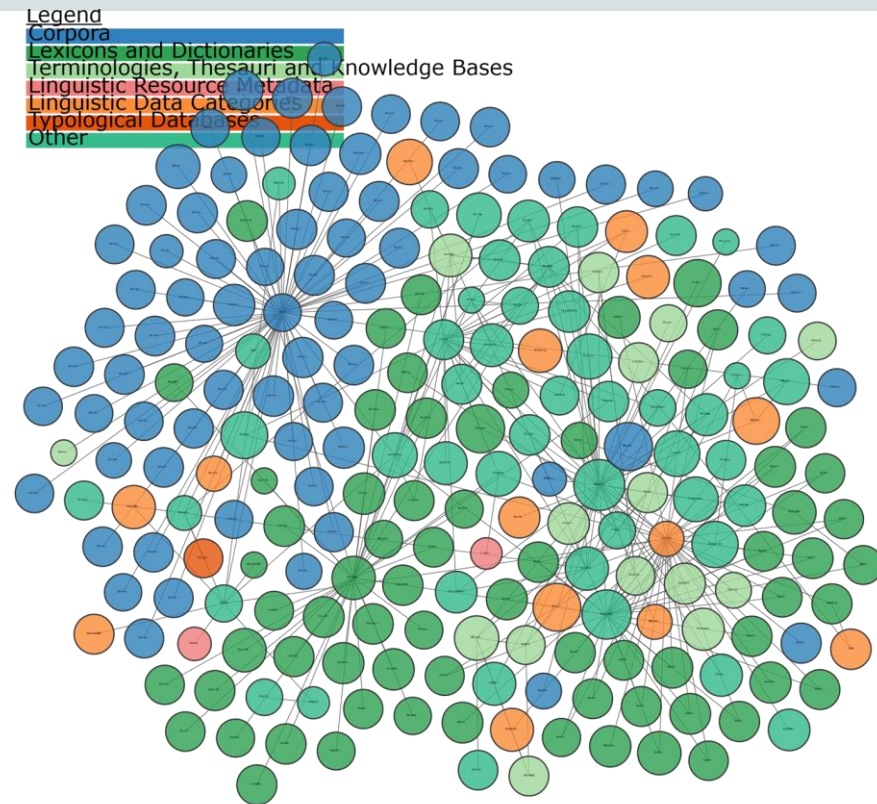
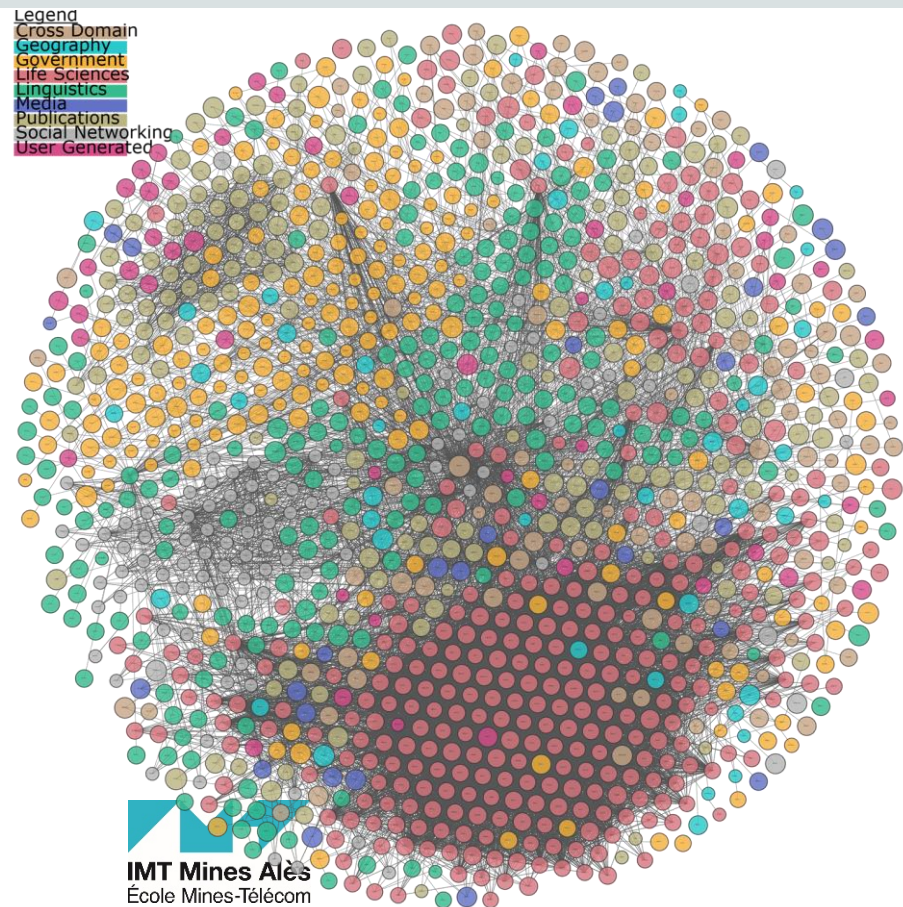
Meta-Share



# EXAMPLE USE-CASE

## The LOD/LLOD Cloud

28



The Linked Open Data Cloud

[Browse](#)

[Submit a dataset](#)

[Diagram](#)

[Subclouds](#)

[About](#)

[Logout](#)

### Edit dataset

Identifier

Title

Dataset title

Description

Dataset description

Full Download

[+]

SPARQL Endpoint

[+]

Other Download

[+]

Example

[+]

Keywords

Domain

Website

Valid URL

Contact Point

Name:

Email:

Name

Email

Links

[+]

Size

0

## General purpose template

<https://github.com/Wimmics/dekalog/blob/master/template-description.ttl>

```
@prefix rdfs:      <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:    <http://www.w3.org/2002/07/owl#> .
@prefix xsd:      <http://www.w3.org/2001/XMLSchema#> .
@prefix dcat:     <http://www.w3.org/ns/dcat#> .
@prefix foaf:     <http://xmlns.com/foaf/0.1/> .
@prefix prov:     <http://www.w3.org/ns/prov#> .
@prefix schema:   <http://schema.org/> .
@prefix void:     <http://rdfs.org/ns/void#> .
@prefix sd:       <http://www.w3.org/ns/sparql-service-description#> .
@prefix dce:      <http://purl.org/dc/elements/1.1/> .
@prefix dct:      <http://purl.org/dc/terms/> .
@prefix skos:     <http://www.w3.org/2004/02/skos/core#> .
```

```
# Below is a non-exhaustive template for the description of a RDF dataset and its associated endpoint.
# This template tries to use as much of the most common vocabularies as possible to create a basic description.
# The resources linked to this description should also be described in the Dataset when relevant.
# Repeated information as objects of different properties will facilitate the access to the description for future users and should be kept.
```

## General purpose template

<dataset-uri>

```
a    dcat:Dataset , void:Dataset , schema:Dataset , prov:Entity , sd:Dataset ;  
dct:title      "Name of the dataset" ;  
schema:name    "Name of the dataset" ;  
skos:prefLabel "Name of the dataset" ;  
rdfs:label     "Name of the dataset" ;  
dct:description "Long description of the dataset" ;  
schema:description "Long description of the dataset" ;  
rdfs:comment    "Long description of the dataset" ;  
  
foaf:homepage   "Url of a web page describing this dataset, if any" ;  
dcat:landingPage "Url of a web page describing this dataset, if any" ;  
  
dcat:contactPoint "Contact information for this dataset, preferably an email adress or a VCard" ;  
  
dct:creator      <Uri of creator(s)> , "Avoid string literals when possible" ;  
schema:author    <Uri of creator(s)> , "Avoid string literals when possible" ;  
dct:publisher     <Uri of creator(s)> , "Avoid string literals when possible" ;  
prov:wasAttributedTo <Uri of creator(s)> , "Avoid string literals when possible" ;  
schema:publisher  <Uri of creator(s)> , <publisher(s)> , "Avoid string literals when possible" ;  
schema:editor     <Uri of creator(s)> , <editor(s)> , "Avoid string literals when possible" ;  
dct:contributor   <Uri of contributor(s)> , "Avoid string literals when possible" ;
```



## General purpose template

```
dct:language      "Lang" , "tags" , "used" , "in" , "the" , "literals" , "if" , "any" ;
schema:inLanguage "Lang" , "tags" , "used" , "in" , "the" , "literals" , "if" , "any" ;

dct:issued        "Date time of creation"^^xsd:date ;
prov:wasGeneratedAtTime "Date time of creation"^^xsd:date ;
schema:datePublished "Date time of creation/publication"^^xsd:date ;
dct:modified       "Date time of the last modification"^^xsd:date ;

dct:subject       "Keywords" , "describing" , "the" , "content" , "of" , "the" , "dataset" , "strings" , "or" , <URIs> ;
dcat:keyword       "Keywords" , "describing" , "the" , "content" , "of" , "the" , "dataset" , "strings" , "or" , <URIs> ;
schema:keywords    "Keywords" , "describing" , "the" , "content" , "of" , "the" , "dataset" , "strings" , "or" , <URIs> ;

void:uriSpace      "Namespace of the uris of the resources created in the dataset, if any" ;
void:uriRegexPattern "Namespace of the uris of the resources created in the dataset, if any, as a regex pattern" ;
```



## General purpose template

```
void:sparqlEndpoint <Url of the SPARQL endpoint> ;
dcat:distribution [
    a                dcat:Distribution ;
    dct:title        "This dataset's endpoint" ;
    dcat:accessURL   <Url of the SPARQL endpoint> ;
    dcat:mediaType   "application/sparql-query"
] ;

void:dataDump        <Url where the dataset can be downloaded as a file, if any> ;
dcat:distribution [
    a                dcat:Distribution ;
    dct:title        "This dataset's archive title" ;
    dcat:downloadUrl <Url where the dataset can be downloaded as a file, if any> ;
    dct:format       "Format of the dump file"
] ;

dcat:service <endpoint-uri> ;

sd:defaultGraph <Resource used to describe the default graph> ;
sd:namedGraph <Resource used to describe a named graph> .

<Resource used to describe the default graph> a sd:Graph ;
    dct:modified        "Date time of the last modification"^^xsd:date .

<Resource used to describe a named graph> a sd:Graph, sd:NamedGraph ;
    sd:name             <URI of the named graph> .
```

<endpoint-uri>

a sd:Service, dcat:DataService , prov:Entity ;

dcat:servesDataset <dataset-uri> ;

sd:endpoint <Url of the SPARQL endpoint> ;

dcat:endpointUrl <Url of the SPARQL endpoint> .

dct:creator <Uri of creator(s) of the endpoint> , "Avoid string literals when possible" ;

prov:wasAttributedTo <Uri of creator(s) of the endpoint> , "Avoid string literals when possible" ;

sd:availableGraphs <dataset-uri> ;

sd:defaultDataset <dataset-uri> .

## General purpose template

```
# The given template can be expanded using the DCAT, PROV-O, Schema, VoID, SPARQL-SD, DCTERMS, or any other vocabulary.
# As much as possible, every URI should have a label or a title.
# As an example, here are some elements that can be use to expand the dataset description:
#
#<dataset-uri>
#   schema:isBasedOn      <Uri of the source of the dataset, if relevant> ;
#   prov:wasDerivedFrom   <Uri of the source of the dataset, if relevant> ;
#   prov:wasGeneratedBy   <Uri of the process at the origin of the dataset> ;

#   void:vocabulary      <Uri(s) of the vocabularies/ontologies used in the dataset> .
#
#   void:triples          "Number of triples in the dataset"^^xsd:integer ;
#   void:classes          "Number of classes in the dataset"^^xsd:integer ;
#   void:properties       "Number of properties in the dataset"^^xsd:integer ;
#
#<Resource used to describe the default graph>
#   void:triples          "Number of triples in the graph"^^xsd:integer ;
#   void:classes          "Number of classes in the graph"^^xsd:integer ;
#   void:properties       "Number of properties in the graph"^^xsd:integer .
#
#<Resource used to describe a named graph>
#   void:triples          "Number of triples in the graph"^^xsd:integer ;
#   void:classes          "Number of classes in the graph"^^xsd:integer ;
#   void:properties       "Number of properties in the graph"^^xsd:integer .
```