European Commission

OPEN DATA SUPPORT

Training Module 1.5

Promoting the reuse of Open Government Data through the Open Data Interoperability Platform (ODIP)

pwc

Presentation metadata

Open Data Support is funded by the European Commission under SMART 2012/0107 'Lot 2: Provision of services for the Publication, Access and Reuse of Open Public Data across the European Union, through existing open data portals' (Contract No. 30-CE-0530965/00-17).

This presentation has been created by PwC

Authors:
Michiel De Keyzer, Nikolaos Loutas and Stijn Goedertier

Disclaimers

1. The views expressed in this presentation are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission.
The European Commission does not guarantee the accuracy of the information included in this presentation, nor does it accept any responsibility for any use thereof.
Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission.
All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.

2. This presentation has been carefully compiled by PwC, but no representation is made or warranty given (either express or implied) as to the completeness or accuracy of the information it contains. PwC is not liable for the information in this presentation or any decision or consequence based on the use of it.. PwC will not be liable for any damages arising from the use of the information contained in this presentation. The information contained in this presentation is of a general nature and is solely for guidance on matters of general interest. This presentation is not a substitute for professional advice on any particular matter. No reader should act on the basis of any matter contained in this publication without considering appropriate professional advice.

Learning objectives

By the end of this training module you should have an understanding of:

- How you can overcome the barriers of reuse for your datasets.
- How Open Data Support can promote the reuse of datasets.
- What the DCAT Application Profile is and how it can be used.
- What Open Data Interoperability Platform (ODIP) is and how it can be used.

Content

This module contains…

- An outline of the context of Open Government Data in Europe.
- An outline of the Open Data Support project.
- Information about the DCAT Application Profile for Data Portals in Europe as a homogenised metadata model.
- Information on how to use the Open Data Interoperability Platform.

There are more than 160 portals in Europe hosting Open Government Data

Provenance?

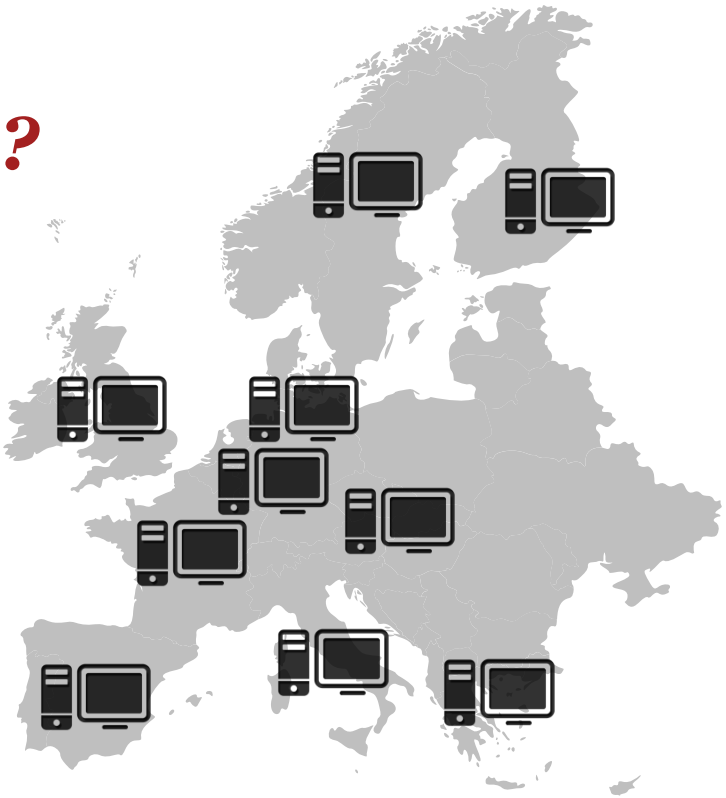
Licence?

Persistence?

Trust?

Availability?

Quality?

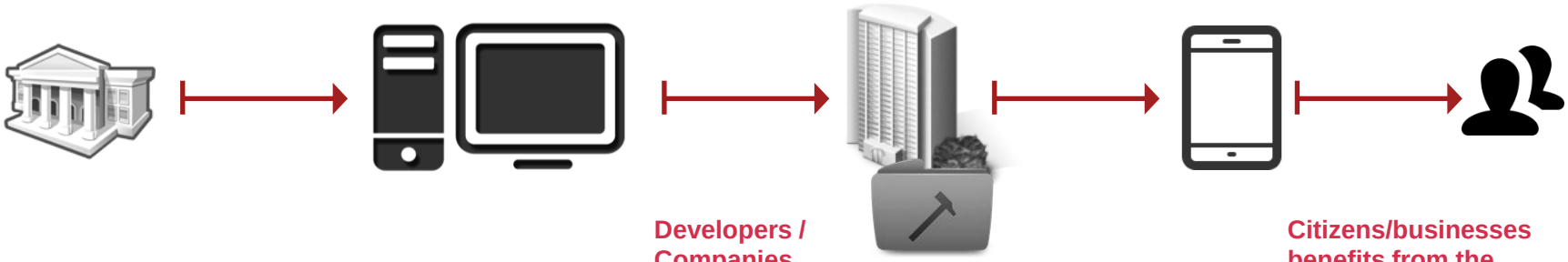


160+

Open Data has a great potential to create social and economic value

Publishing data

Public administrations share data online



Developers / Companies integrate data into app (services)

Developers / Companies search for data

Citizens/businesses benefits from the app (services)

Reusing data

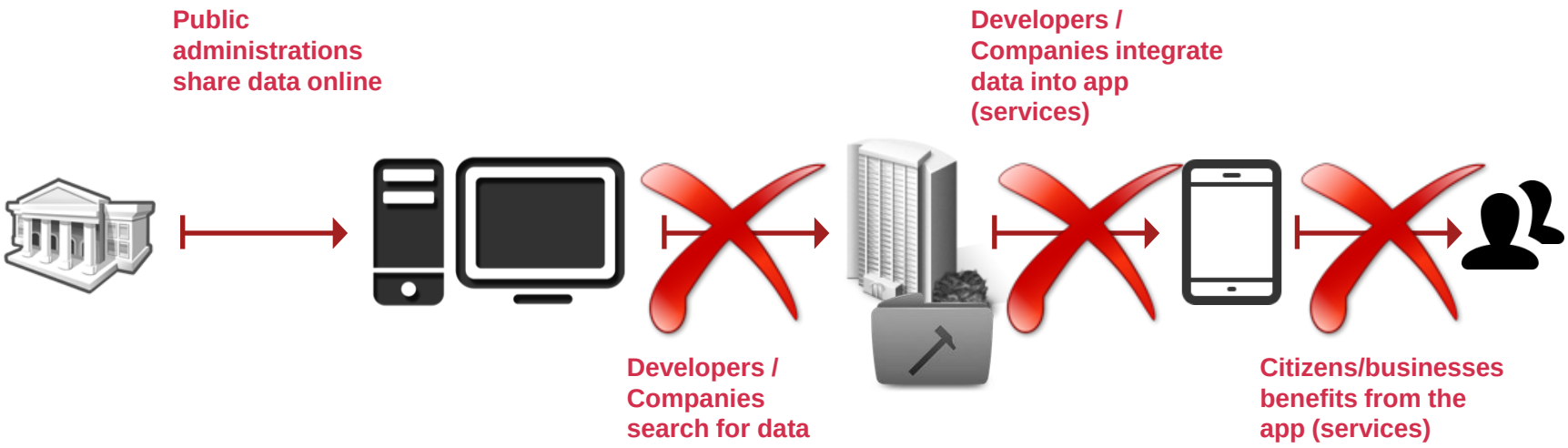
Barriers to Open Data publishing and reuse

Data publishers	Data reusers
No view on which data is more likely to be reused / has a higher ROI potential.	Lack of overview of existing/available datasets.
Unclear business model for publishing Open Data.	Unclear business model for reusing Open Data.
Limited tool support.	Data is often of low quality, outdated, unstructured and/or not machine-readable.
Competing licenses for datasets.	Lack of licensing information or incompatible licenses.
Competing vocabularies for describing datasets.	Different vocabularies when searching for datasets.
Domain-specific metadata needs.	Lack of (good quality) metadata.
Effort required for keeping the metadata up-to-date.	Lack of provenance information.

Metadata

Metadata

No reuse = No social and economic value



Open Data Support

...funded by the European Commission, DG CONNECT, aims at lowering accessibility and awareness barriers.

Open Data Support mission…

To Improve the visibility and facilitate the access to datasets published on local and national Open Data portals in order to increase their reuse within and across borders.

See also:
http://www.slideshare.net/OpenDataSupport

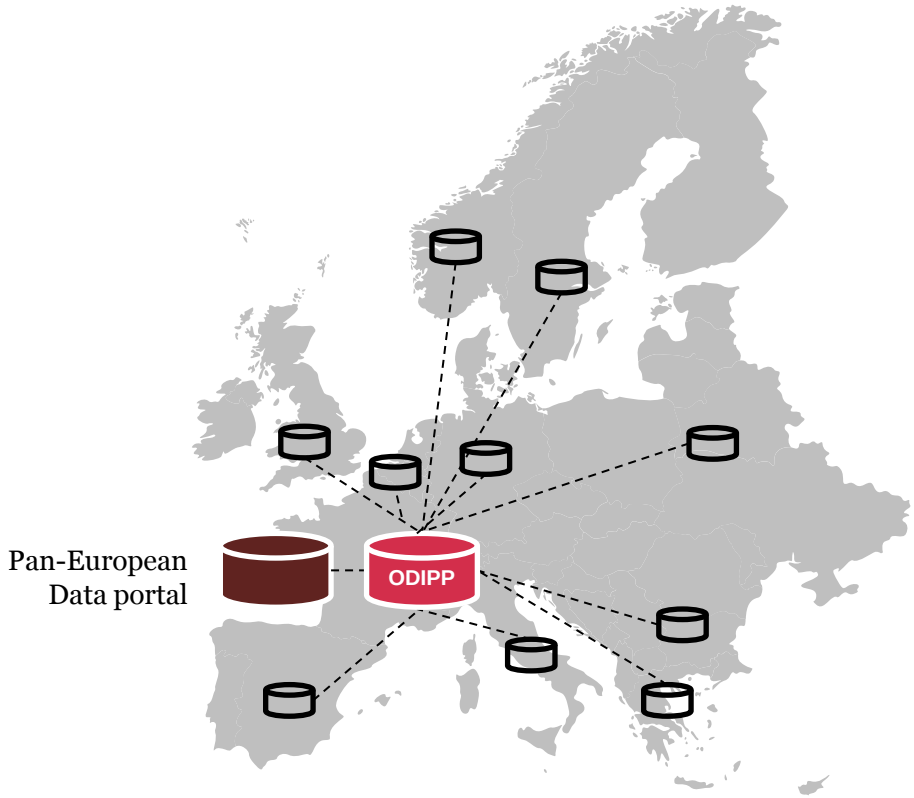
By …

Providing

homogenised access

to metadata descriptions of open datasets via a

single point of access



DCAT Application Profile

A common vocabulary for describing datasets hosted in data portals in Europe, based on the Data Catalogue vocabulary (DCAT).

A shared initiative of …

isa
Interoperability Solutions
for European Public Administrations

Publications Office

European Commission
Communications Networks, Content and Technology
European Commission Directorate General
DG CONNECT

SEMIC
SEMANTIC
INTEROPERABILITY
COMMUNITY

Funded by the ISA Programme under Action 1.1. "Improving semantic interoperability in European eGovernment systems" (a.k.a the SEMIC project).

An international Working Group of experts

- Chair: Antonio Carneiro (Publications Office)
- 59 Working Group members representing:
  - 15 different European Member States (UK,IT,ES,DK,DE,SK,BE,AT,SE,FI,FR,IE,NL,GR,SI)
  - US
  - Several European Institutions and international organisations
  - 40 different Data Portals

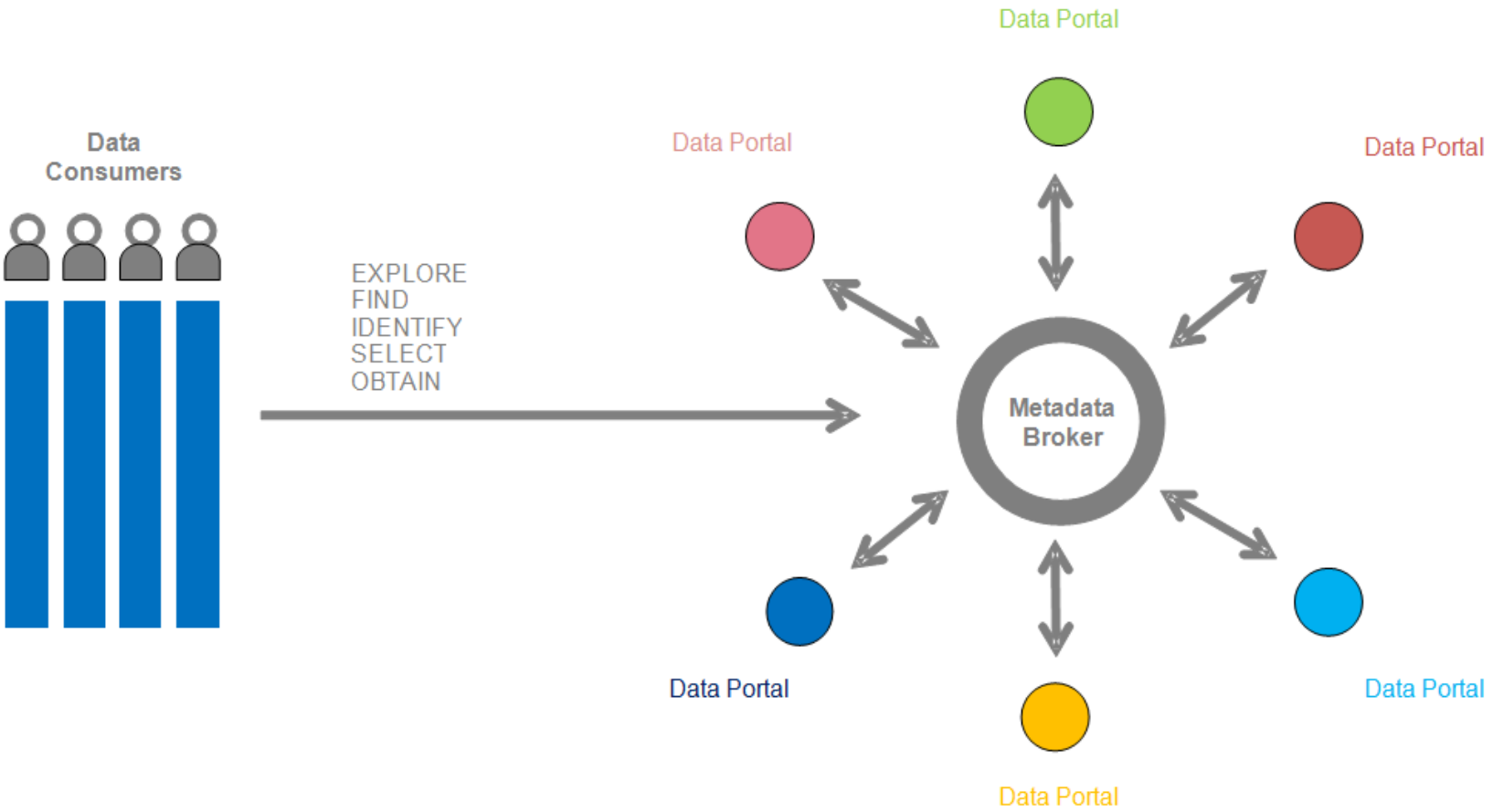
See also:
https://joinup.ec.europa.eu/asset/dcat_application_profile/description

By using a common metadata schema to describe datasets and sharing metadata…

- **Data publishers** increase discoverability and thus reuse of their data.
- **Data reusers** can uniformly search across platforms without facing difficulties caused by the use of separate models or language differences.

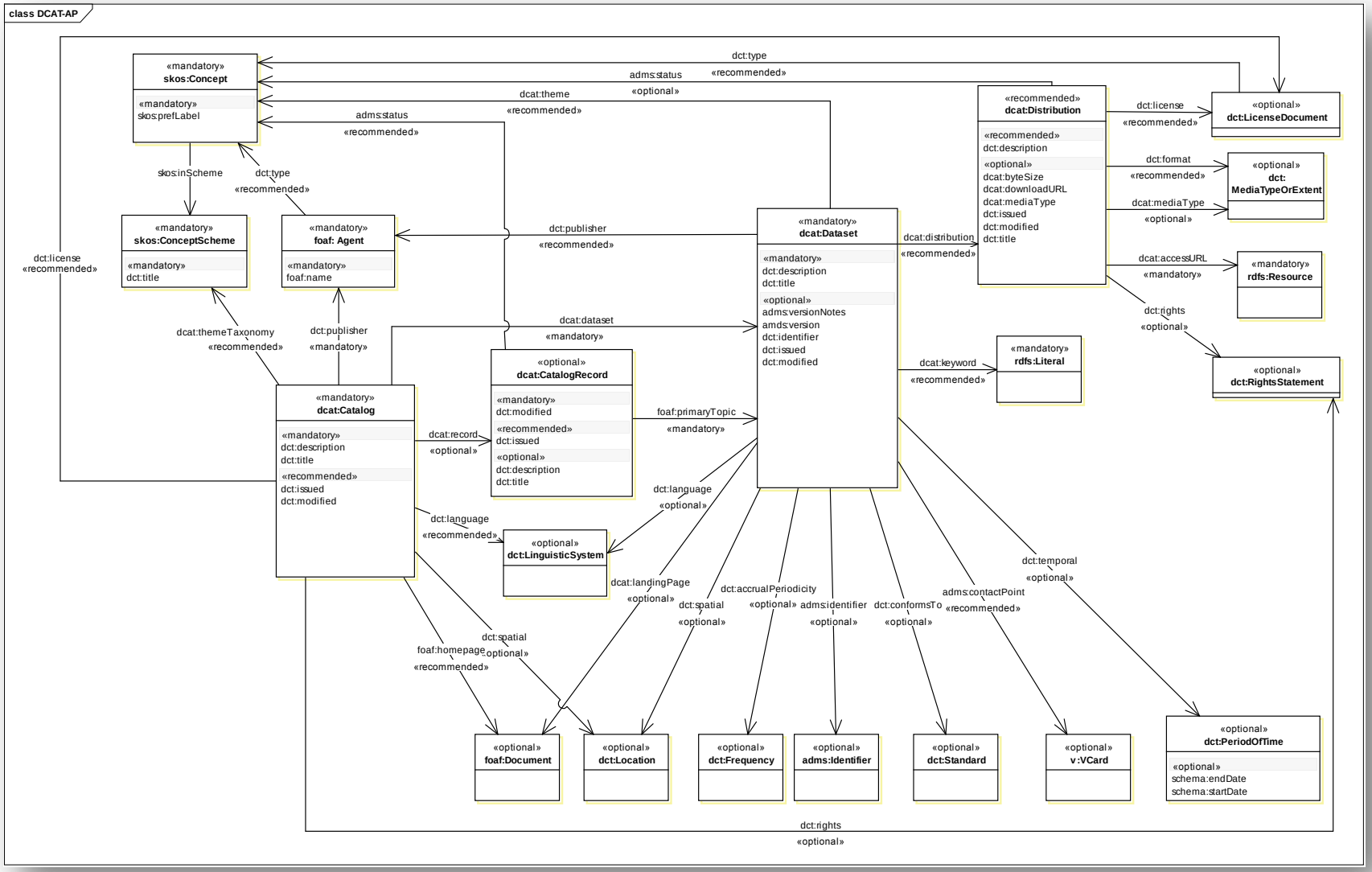
The quality and the availability of the description metadata directly affects how easily datasets can be found!

The DCAT-AP enables the exchange of description metadata between data portals



What's in the specification?

The DCAT Application Profile data model



Usage of the DCAT Application Profile

Mandatory class: a receiver of data MUST be able to process information about instances of the class; a sender of data MUST provide information about instances of the class.

Recommended class: a receiver of data MUST be able to process information about instances of the class; a sender of data MUST provide information about instances of the class, if it is available.

Optional class: a receiver MUST be able to process information about instances of the class; a sender MAY provide the information but is not obliged to do so.

Mandatory property: a receiver MUST be able to process the information for that property; a sender MUST provide the information for that property.

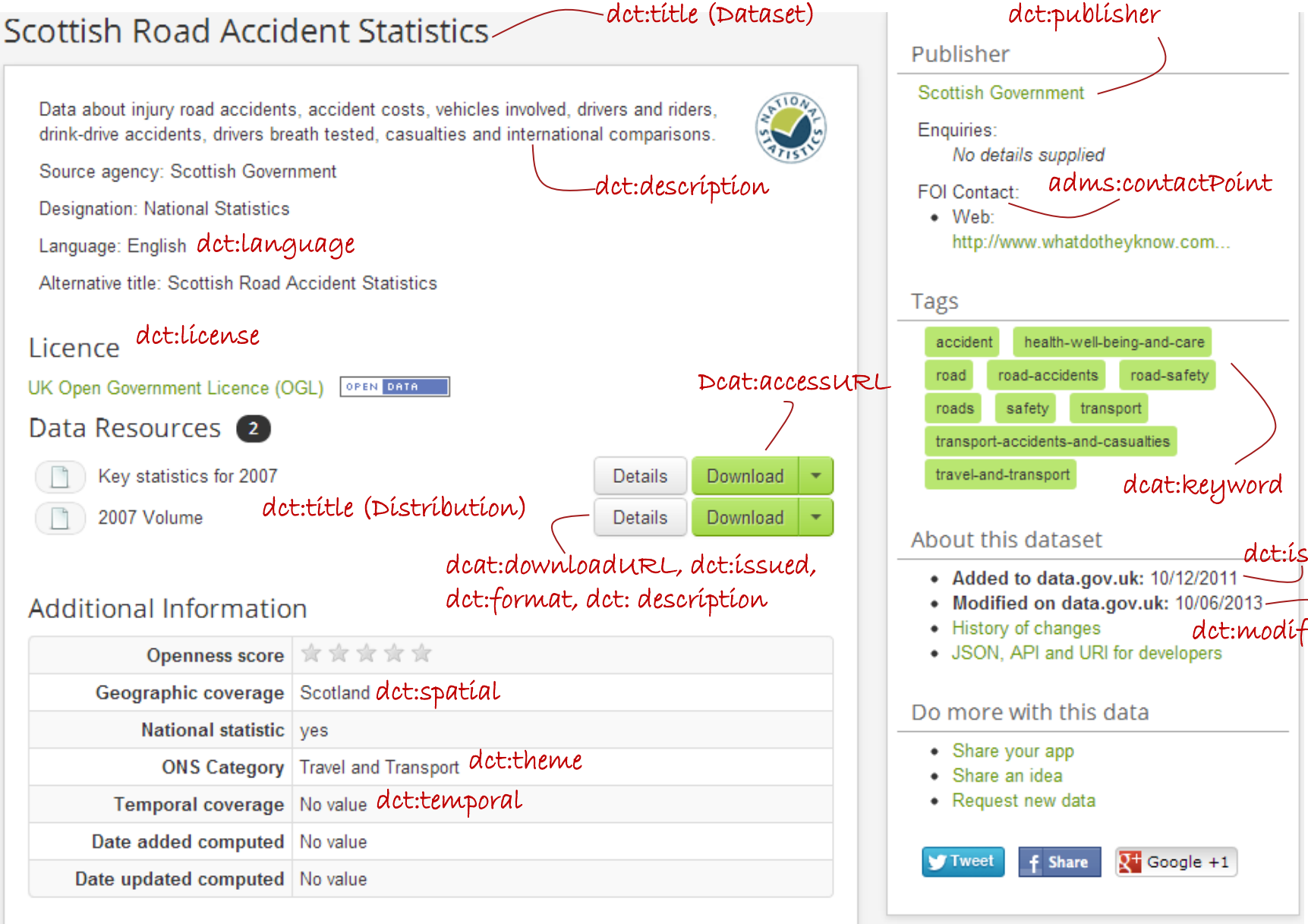
Recommended property: a receiver MUST be able to process the information for that property; a sender SHOULD provide the information for that property if it is available.

Optional property: a receiver MUST be able to process the information for that property; a sender MAY provide the information for that property but is not obliged to do so.

Controlled vocabularies

Property URI	Used for Class	Proposed vocabulary
dcat:mediaType	Distribution	MDR File types Name Authority List
dcat:theme	Dataset	EuroVoc domains
dcat:themeTaxonomy	Catalog	EuroVoc
dct:accrualPeriodicity	Dataset	Dublin Core Collection Description Frequency Vocabulary
dct:format	Distribution	MDR File Type Named Authority List
dct:language	Catalog, Dataset	MDR Languages Named Authority List
dct:publisher	Catalog, Dataset	MDR Corporate bodies Named Authority List
dct:spatial	Catalog, Dataset	MDR Countries Named Authority List, MDR Places Named Authority List
adms:status	CatalogRecord	ADMS change type vocabulary
dct:type	License Document	ADMS license type vocabulary

Mapping example – data.gov.uk



Example description of dataset with the DCAT-AP

<rdf:Description rdf:about="http://data.gov.uk/data ">

<rdf:type rdf:resource="http://www.w3.org/ns/dcat#Catalog"/>

<dct:title xml:lang="en">data.gov.uk</dct:title>

<dct:description xml:lang="en">Description of the data portal</dct:description>

<dct:license rdf:resource=" http://www.nationalarchives.gov.uk/doc/open-government-licence"/>

</rdf:Description>

<rdf:Description rdf:about="http://data.gov.uk/dataset/east-sussex-county-council-election-results"/>

<rdf:type rdf:resource="http://www.w3.org/ns/dcat#Dataset"/>

<dct:title xml:lang="en">East Sussex County Council election results</dct:title>

<dct:description xml:lang="en">A list of elections to East Sussex County Council, which leads to data about candidates, parties, electoral divisions and votes cast. Uses the Open Election Data RDF vocabulary from http://openelectiondata.org/

</dct:description>

</rdf:Description>

<rdf:Description rdf:adbout="http://www.eastsussex.gov.uk/yourcouncil/localelections/election2009/default.aspx"/>

<rdf:type rdf:resource="http://www.w3.org/ns/dcat#Distribution"/>

<dct:title xml:lang="en">East Sussex County Council election 4 June 2009, and subsequent bi-elections</dct:title>

<dcat:accessURL rdf:resource="http://www.eastsussex.gov.uk/yourcouncil/localelections/election2009/default.aspx "/>

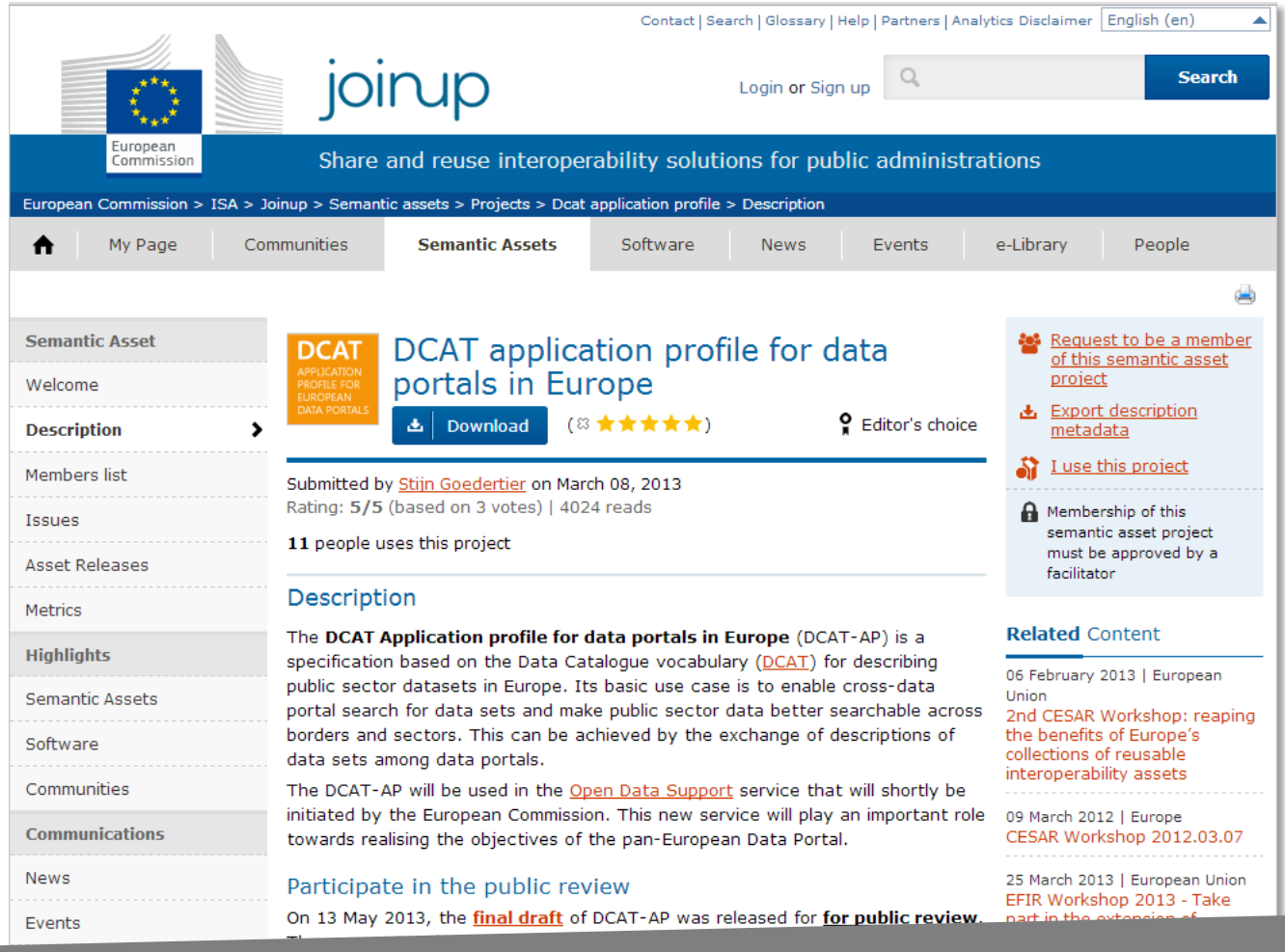
<dct:license rdf:resource="http://www.nationalarchives.gov.uk/doc/open-government-licence"/>

</rdf:Description>

Creating mappings to the DCAT-AP

Dataset Properties			
Raw Predicate	Example Value	Harmonized Predicate	Generated SPARQL
http://data.gov.uk/predicate/title	Government Major Projects data for the Foreign and Commonwealth Office 2012	dct:title	prefix dct:<http://purl.org/dc/terms/> INSERT { ?harmds dct:title ?d.} where { ?ds a <http://www.w3.org/ns/dcat#Dataset>. ?ds <http://data.gov.uk/predicate/title> ?d. ?harmrecord <http://xmlns.com/foaf/0.1/primaryTopic> ?harmds. ?harmrecord <http://data.opendatasupport.eu/ontology/harmonisation.owl#raw_ ?ds. }
http://data.gov.uk/predicate/unpublished	FALSE		
http://data.gov.uk/predicate/update_frequency	other	dct:accrualPeriodicity	prefix dct:<http://purl.org/dc/terms/> INSERT { ?harmds dct:accrualPeriodicity ?d.} where { ?ds a <http://www.w3.org/ns/dcat#Dataset>. ?ds <http://data.gov.uk/predicate/update_frequency> ?d. ?harmrecord <http://xmlns.com/foaf/0.1/primaryTopic> ?harmds. ?harmrecord <http://data.opendatasupport.eu/ontology/harmonisation.owl#raw_ ?ds. }
http://data.gov.uk/predicate/update_frequency-other	quarterly	dct:accrualPeriodicity	prefix dct:<http://purl.org/dc/terms/> INSERT { ?harmds dct:accrualPeriodicity ?d.} where { ?ds a <http://www.w3.org/ns/dcat#Dataset>. ?ds <http://data.gov.uk/predicate/update_frequency-other> ?d. ?harmrecord <http://xmlns.com/foaf/0.1/primaryTopic> ?harmds. ?harmrecord <http://data.opendatasupport.eu/ontology/harmonisation.owl#raw_ ?ds. }

Where can you find it?



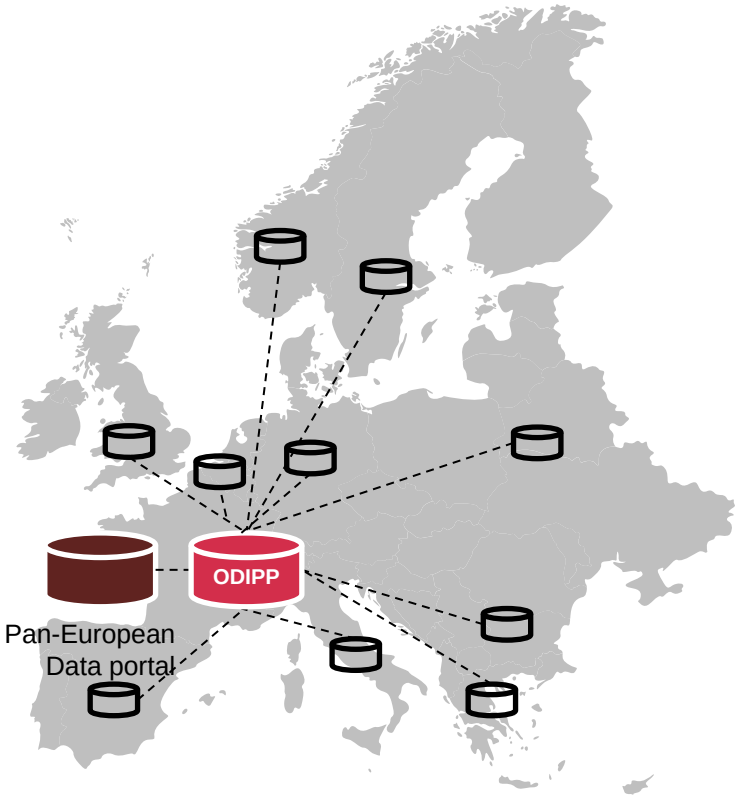
https://joinup.ec.europa.eu/asset/dcat_application_profile/description

Share the metadata of you datasets on ODIP

The Open Data Interoperability Platform (ODIP) enables you to share metadata of datasets described using the DCAT-AP, thus improving the discoverability and visibility of your datasets, eventually leading to wider reuse.

What can ODIP do?

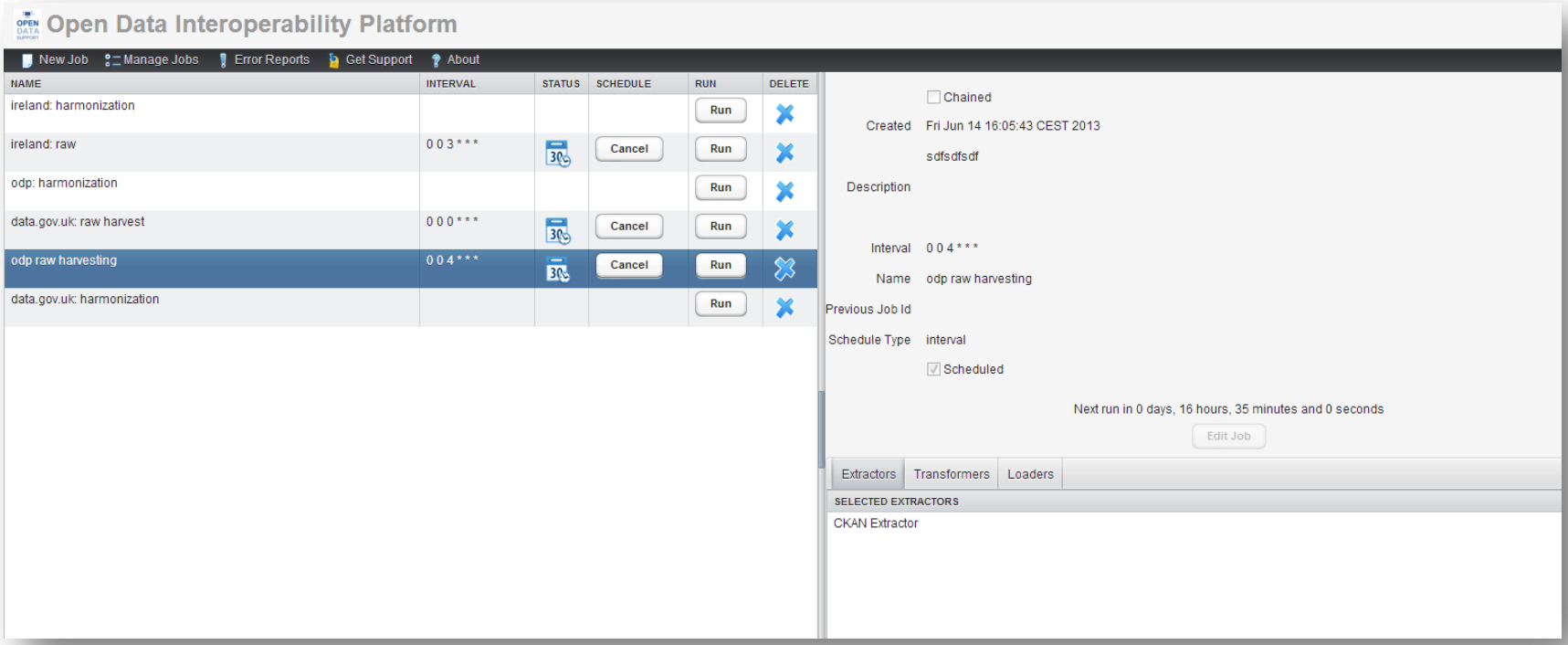
- **Harvest** metadata from an Open Data portal.
- **Transform** the metadata to RDF.
- **Harmonise** the RDF metadata produced in the previous steps with DCAT-AP.
- **Validate** the harmonised metadata against the DCAT-AP.
- **Publish** the description metadata as Linked Open Metadata.
- **Translate** metadata automatically in English



How can ODIP help you improve your metadata?

- ODIP maps your metadata to a standard model, i.e. the DCAT-AP.
- ODIP helps you reuse standardised multilingual controlled vocabularies in your metadata, replacing error-prone text values or tailor-made lists.
- By means of its validation services, ODIP allows you to detect inconsistencies and errors in your metadata.
- ODIP assigns persistent URIs to your metadata.
- ODIP links your metadata with other metadata, thus adding context to it and enriching its meaning.
- ODIP automatically translates the title and description of the metadata to English.

How does ODIP look like?



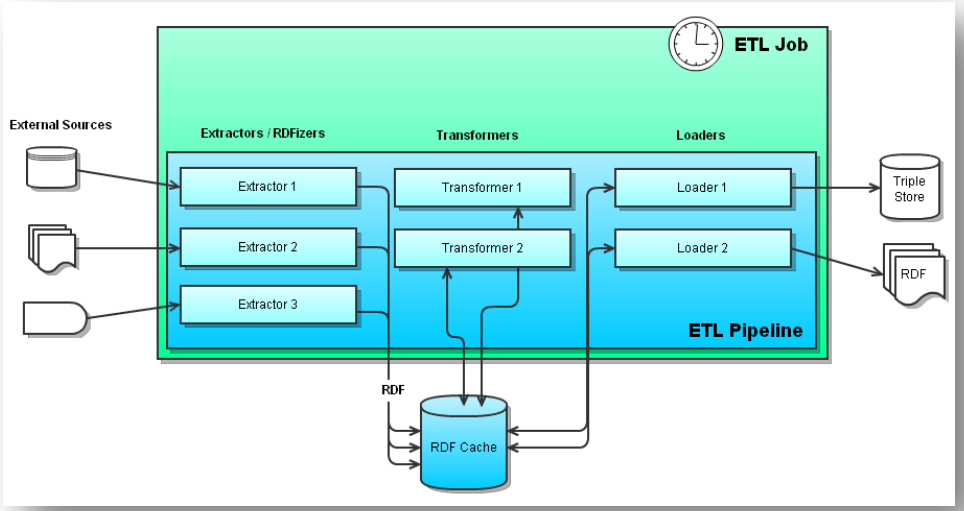
http://odip.opendatasupport.eu

An ODIP Job

The ODIP job consists of three possible phases which need to be ran in order and that are composed of several plug-ins :

1. **Extraction**
2. **Transformation**
3. **Loading**

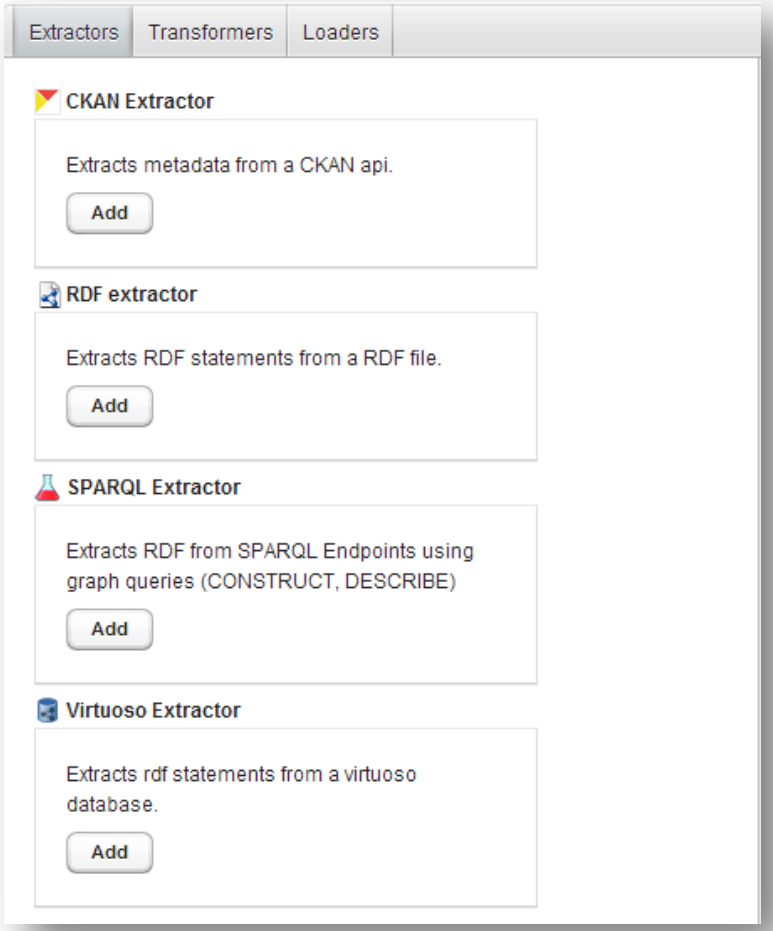
Furthermore these jobs can be scheduled to be launched periodically, in succession or manually.



Overview of ODIP's Extract-Transform-Load process

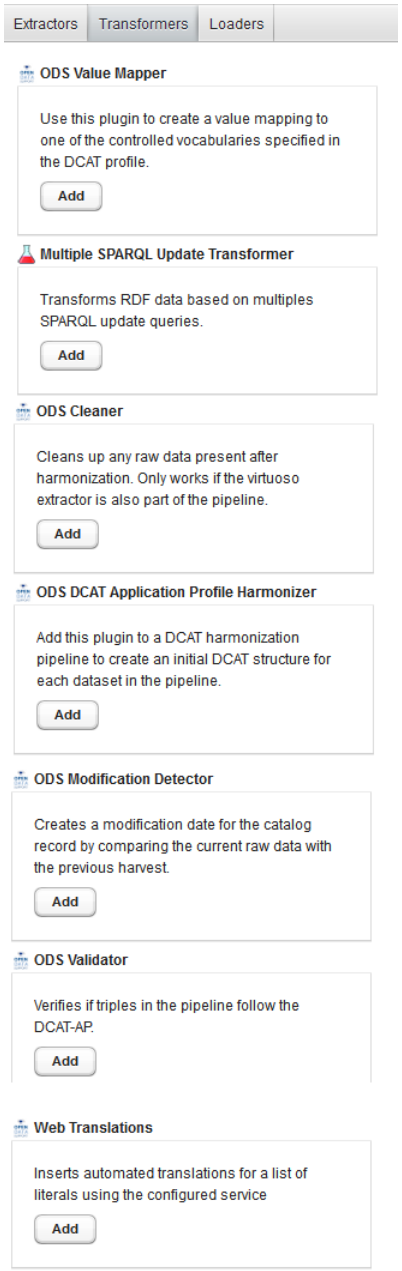
1. Extraction

- The extraction phase entails retrieving (extracting) raw data from a given source Open Data portal using the appropriate plug-in, depending on the technology of the source.
- Available extractors:
  - CKAN Extractor
  - RDF extractor
  - SPARQL Extractor
  - Virtuoso Extractor
  - CSV Extractor



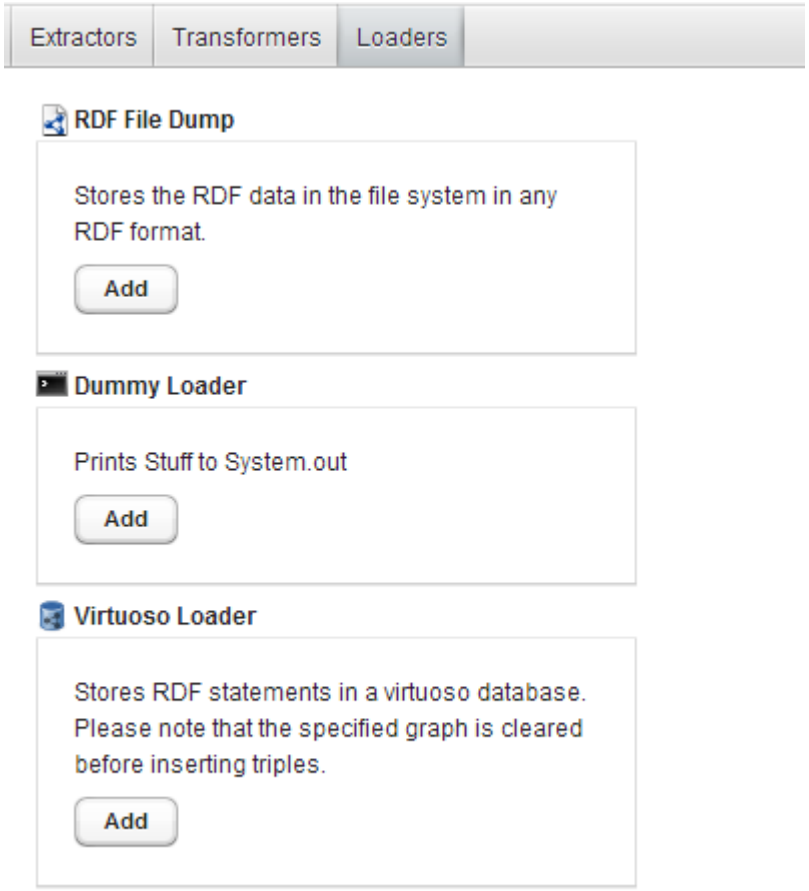
2. *Transformation (1/3)*

- The goal of the transformation phase is to harmonise, cleanse and prepare for storing on ODIP metadata harvested from Open Data portals.
- Available transformers:
  - ODS Value Mapper.
  - SPARQL Update Query Transformer.
  - ODS Cleaner.
  - ODS DCAT Application Profile Harmoniser.
  - ODS Modification Detector.
  - ODS Validator.
  - Web Translations.



Loading

- In the loading phase, the harvested and harmonised metadata is stored on Virtuoso's RDF repository using the Virtuoso Loader.



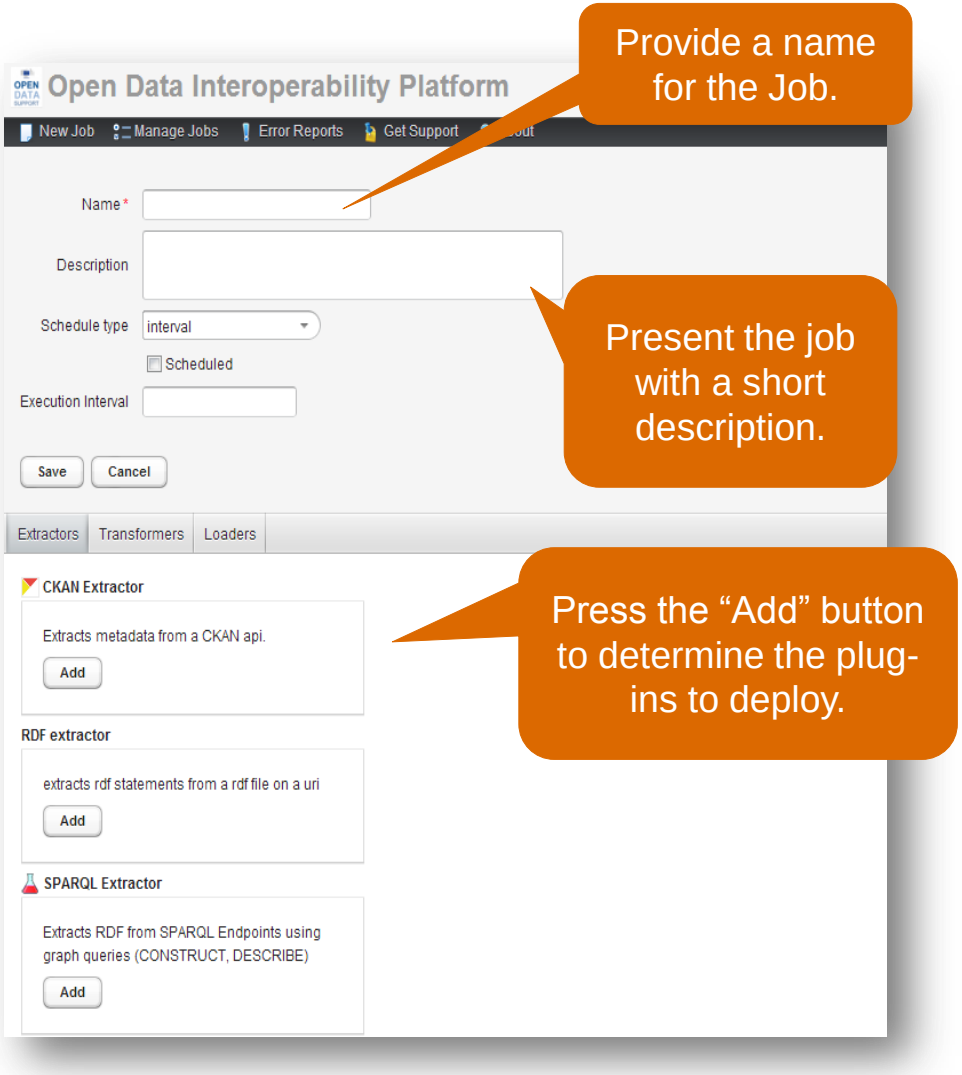
Example

Harvesting a CKAN-based Open Data portal

1. Create a new job on ODIP
2. Extraction phase
   - Add and Configure a CKAN Extractor to harvest data from a CKAN API.
3. Transformation phase
   - Add ODS Value mapper
   - Add a SPARQL Update Query Transformer with the pertinent queries
   - Add ODS Cleaner
   - Add and configure DCAT Application Profile Harmoniser
   - Add Modification detector
   - Add ODS Validator
   - Add Web Translations
4. Loading phase
   - Load the extracted data in a Virtuoso RDF Store via the Virtuoso Loader
5. Scheduling the job on ODIP

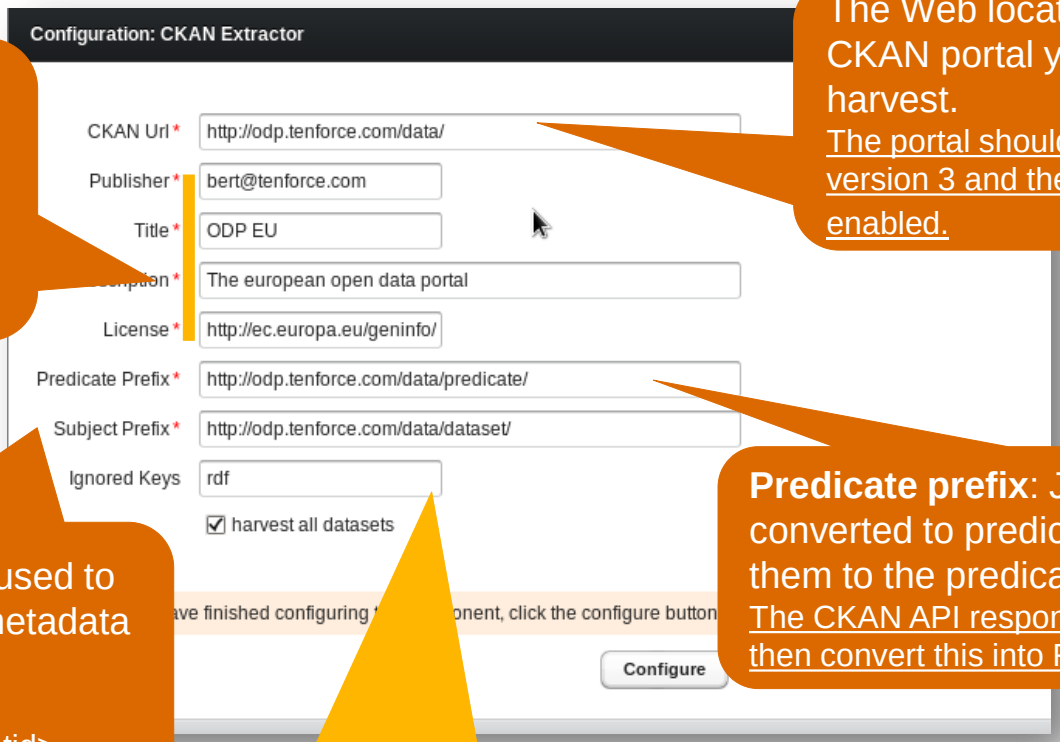
Example – 1. Create Job : Creating a job on ODIP

- To create a new job, click on "*New Job*".
- At the bottom part of the screen you can configure the actual tasks within each of the three phases by selecting a tab.
- For each phase you can add and configure modules accordingly.



Example – 2.Extraction : Adding and Configuring a CKAN Extractor to harvest data from a CKAN API

After adding the CKAN extractor plugin you will be prompted to fill out the following form:



Publisher, license, title and description: Used in the stored catalog for the dct:publisher, dct:license, dct:title and dct:description properties.

The Web location of the CKAN portal you wish to harvest.
The portal should support API version 3 and the API must be enabled.

Subject prefix: The prefix used to create a URI for each the metadata of harvested dataset.
The subject is created as <subjectprefix>/dataset/<datasetid>

Predicate prefix: JSON attributes are converted to predicates by appending them to the predicate prefix.
The CKAN API response is in JSON, we then convert this into RDF.

Ignored keys: A comma seperated list of JSON attributes that should not be converted to RDF triples.

Example – 3. Transformation : Adding and configuring plug-ins to harmonise data(1/3)

- Start by adding the **ODS DCAT Application Profile Harmonizer**.

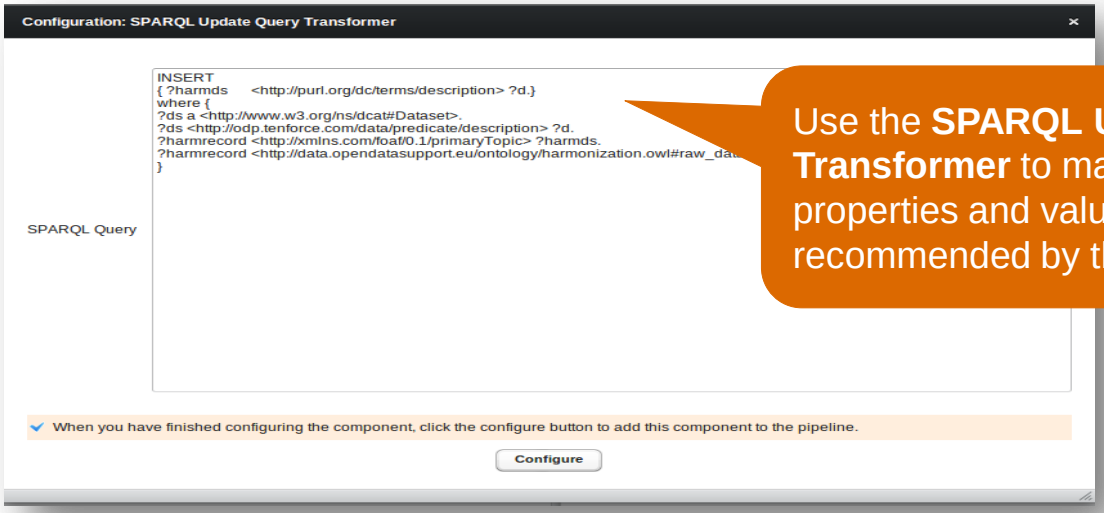
  ✓ This plugin will create the harmonized catalog data and a basic skeleton for each dataset it identifies.

- Use the **Modification Detector** to compare provenance data generated by the CKAN extractor between the current and previous version of the raw data to set the dct:modified field of the catalog records.

  ✓ No configuration is required.

Example – 3. Transformation : Adding and configuring plug-ins to harmonise data (2/3)

- Mapping the description of dataset to dct:description as required by the DCAT-AP.



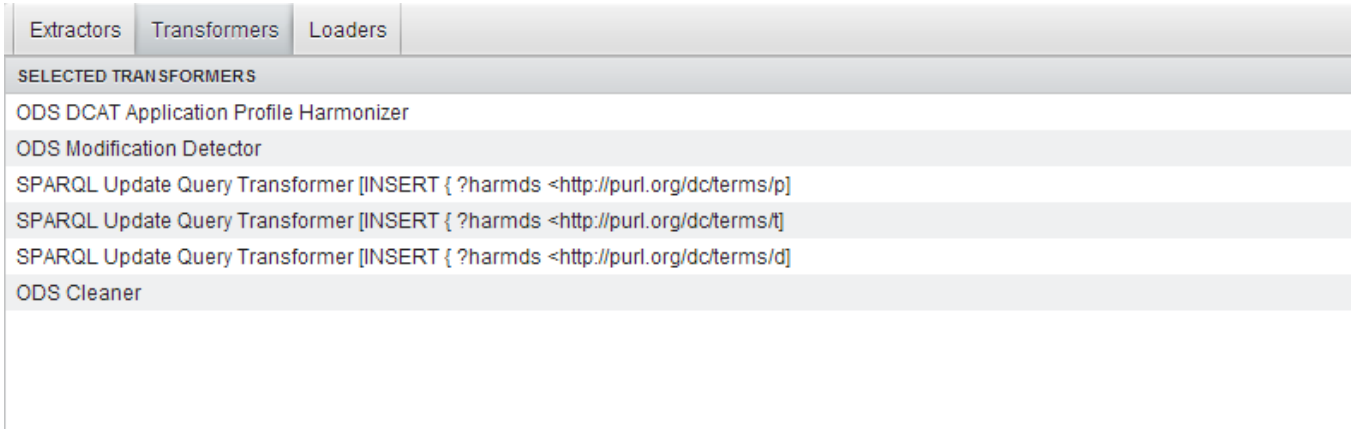
- Use the **ODS Cleaner Plugin** to remove raw data loaded into the working set before storing it into a harmonized graph.

  ✓ No configuration is required.

Example – 3. Transformation : Adding and configuring plug-ins to harmonise data (3/3)

Result

The final result of your harmonisation pipeline should look similar to the following :

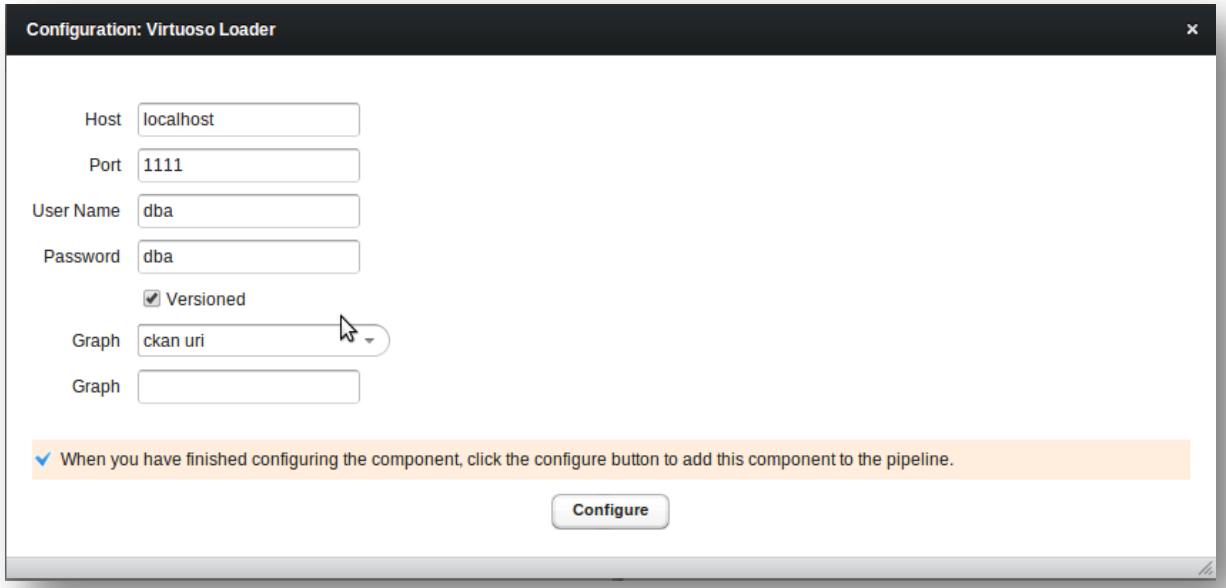


Configure the Virtuoso Loader to load the harmonized data into Virtuoso.

Example – 4. Loading: Load the extracted data in a Virtuoso RDF Store via the Virtuoso Loader

The Virtuoso Loader will store the generated triples in the Virtuoso RDF store. The triples will be inserted into a graph of your choice.

The Virtuoso Loader needs host, port and user credentials to connect to your Virtuoso server.



5. Scheduling a job on ODIP

A job can be scheduled to run at a set interval or chained after another job:

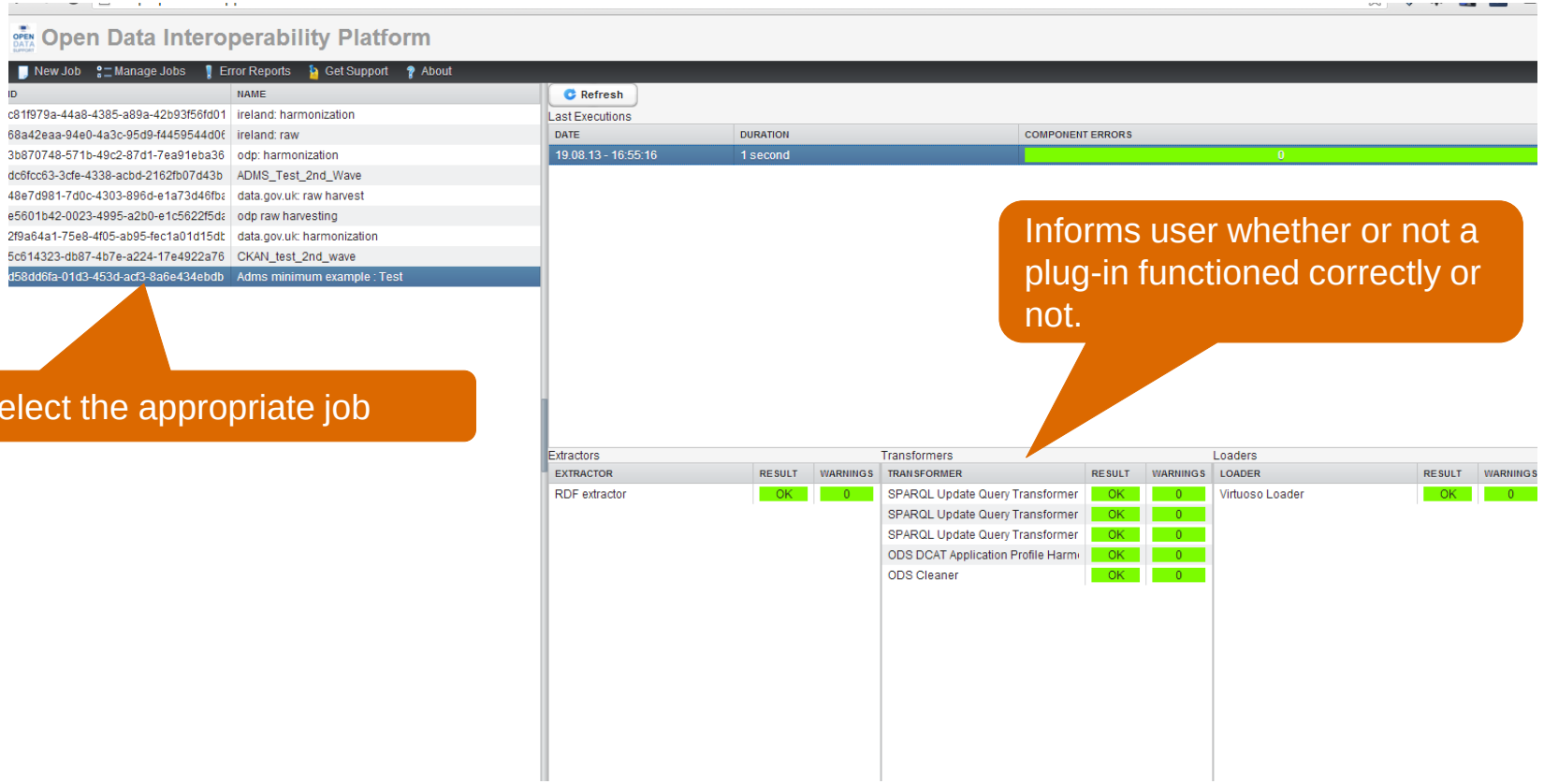
- **Interval Scheduling**:
  <sec> <min> <hour> <day-of-month> <month> <day-of-week>

  ❑ Example:

  - 0 0 4 * * * - each day at 4 am
  - 0 0 0 * * 1 - each Monday at midnight
  - 0 30 * * * - every half past the hour
- **Chained scheduling:** Select a job after which this job should be executed.

ODIP Reporting tool

Whenever a “job” is ran, a report is created and can be reviewed as can be seen in the following screenshot:

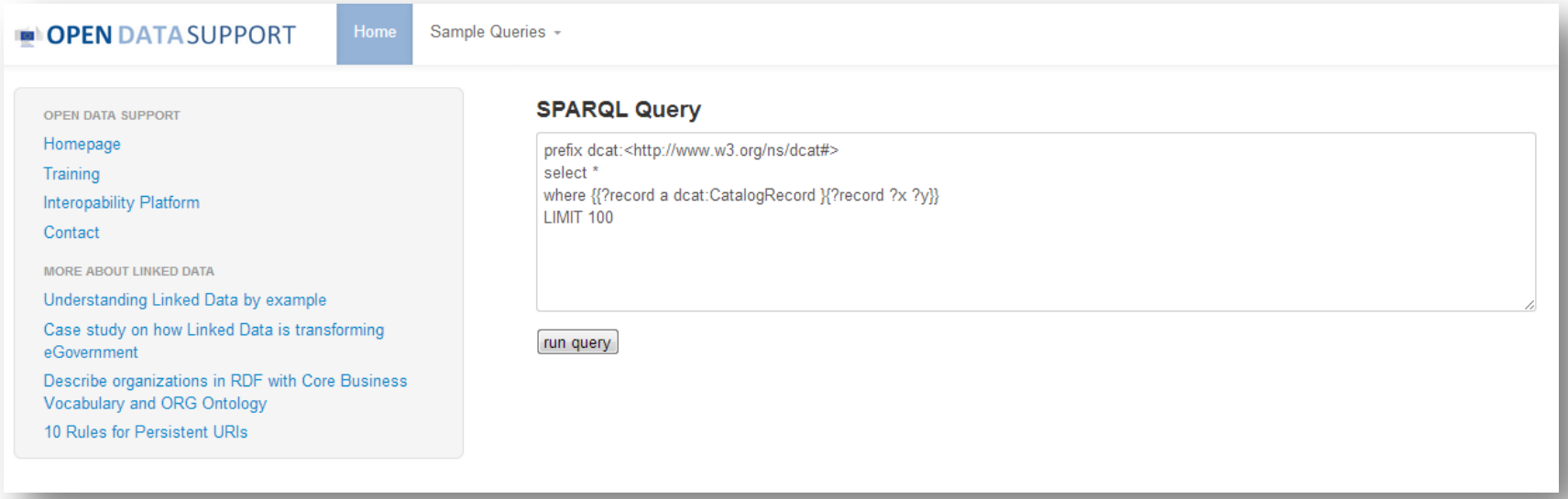


Discover datasets through ODIP

The Open Data Interoperability Platform (ODIP) enables you to share metadata of datasets described using the DCAT-AP, thus improving the discoverability and visibility of your datasets, eventually leading to wider reuse.

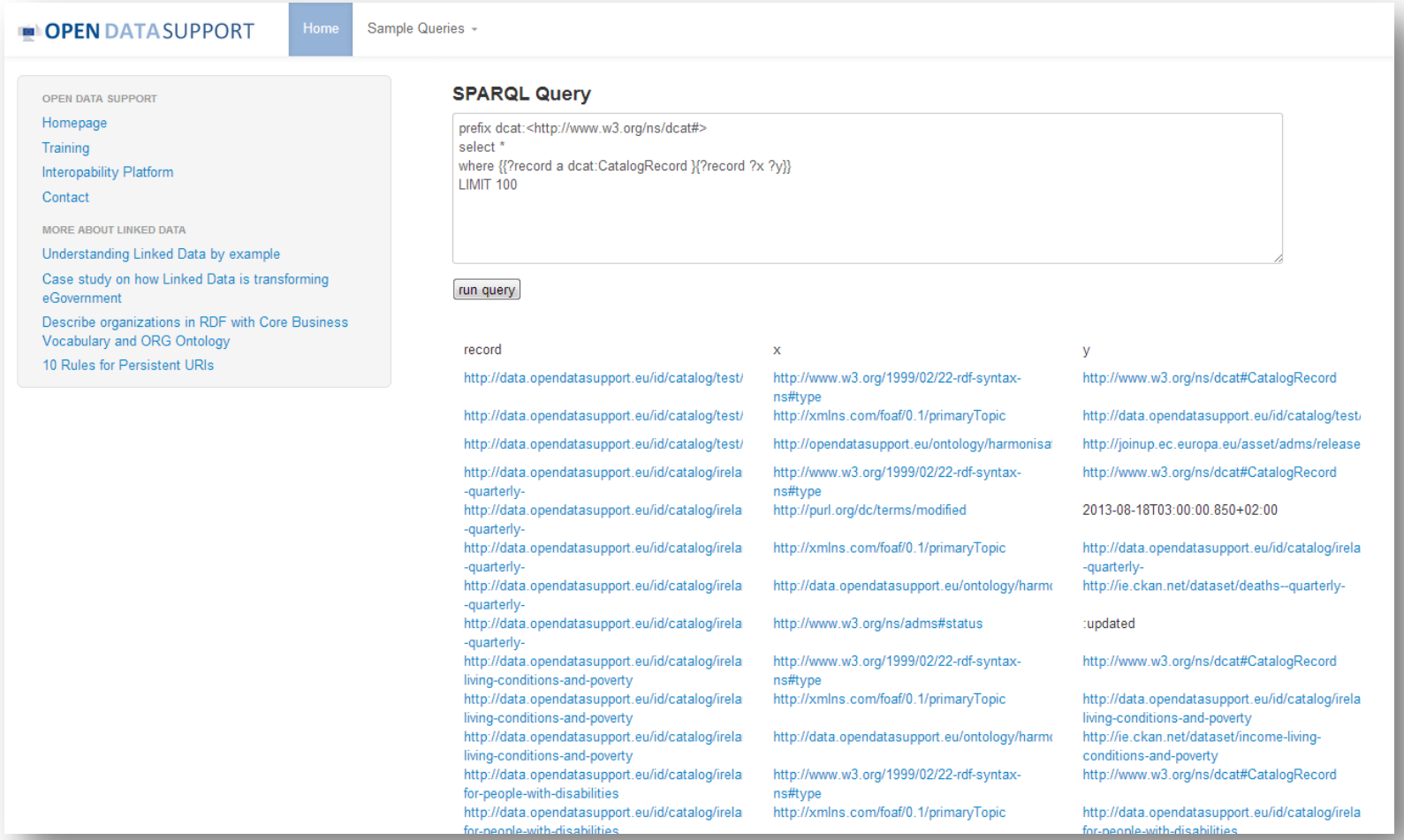
The public SPARQL endpoint of ODIP

Query interface



http://data.opendatasupport.eu

The public SPARQL endpoint of ODIP
Result set



More about ODIP

- ODIP is based on the LOD Management Suite, originally created by the Semantic Web Company in the context of LOD2 FP7 project.

- The LOD Manager Suite was further extended by TenForce in the context of Open Data Support for the deployment of ODIP.

- It will be made available on GitHub under GPLv2.

Conclusions

- Good quality description metadata can improve the discoverability of open datasets.
- DCAT-AP can be used for homogenising metadata of datasets hosted on different Open Data portals and allows for querying them using a uniform vocabulary.
- ODIP can support harvesting, harmonising according to the DCAT-AP and publishing as linked data metadata of datasets published on different Open Data portals.
- ODIP, through its public SPARQL endpoint, provides a single point of access to datasets from all over Europe.
- Easier access to datasets means higher reuse of datasets.

Group questions

http://www.visualpharm.com

How many Open Government Data portals do you know in your country?

http://www.visualpharm.com

In your country, are you aware of any applications or services that were built upon Open Government Data?

http://www.visualpharm.com

How would you compare the visibility of Open Government Data portals with that of traditional data providers such as national statistics offices?

http://www.visualpharm.com

Have you heard about the Open Government Data initiatives of the European Commission?

Take also the online test here!

Thank you!

...and now YOUR questions?

References

Slide 4, 6, 9, 10, 11 & 12:

- Open Data Support: How can we help you?. Open Data Support. http://www.slideshare.net/OpenDataSupport/open-data-support-service-description

Slide 12:

- Data Catalogue Vocabulary. http://www.w3.org/TR/vocab-dcat/

Slide 13-21:

- DCAT Application Profile for data portals in Europe Community. ISA Programme. https://joinup.ec.europa.eu/asset/dcat_application_profile/description https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/all

Slide 23-35:

- LODMS User Manual for Open Data Support. Open Data Support

Slide 29:

- Figure from http://www.semantic-web.at/linked-open-data-management-suite-lodms

Related projects and initiatives

DCAT Application Profile for Data Portals in Europe, https://joinup.ec.europa.eu/asset/dcat_application_profile/description

Publicdata.eu, http://www.w3.org/2011/gld/wiki/Main_Page

LOD2 FP7 Project, http://lod2.eu/

The Semantic Web Company, http://www.semantic-web.at/

Linked Open Data Management Suite, http://www.semantic-web.at/linked-open-data-management-suite-lodms

OpenLink Virtuoso, http://virtuoso.openlinksw.com/

Data Catalog Interoperability Protocol, http://spec.datacatalogs.org/

Be part of our team…

Find us on

Open Data Support
http://www.slideshare.net/OpenDataSupport

Open Data Support
http://goo.gl/y9ZZI

Join us on

http://www.opendatasupport.eu

Follow us

@OpenDataSupport

Contact us

contact@opendatasupport.eu