“Open Data Web” – A Linked Open Data Repository Built with CKAN

Cheng-Jen Lee

Andrea Wei-Ching Huang

Tyng-Ruey Chuang

Institute of Information Science, Academia Sinica, Taiwan

CKANCon 2016@Madrid

2016/10/04

Slide and Transcript

Slide

Transcript

https://hackmd.io/s/rJIcV6Op

..or search for #CKANCon on

Outline

- Data Source
- Linked Data
- From Archive Catalog to Linked Data
- Linked Open Data Repository: Open Data Web
- System Architecture
- Implementation
- Limitations
- Future Work

Data Source

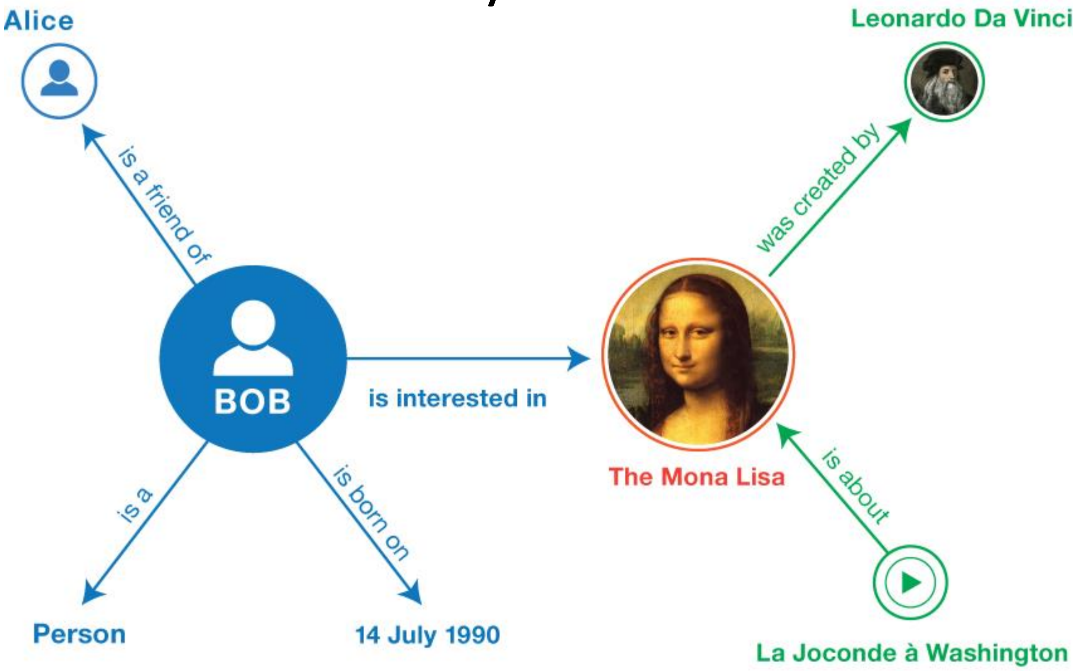
- Union Catalog of Digital Archives Taiwan
  - http://catalog.digitalarchives.tw
- Web catalog for digitized archives in 14 domains from many institutions.
- Part of the catalog is released under CC licenses
  - About 840,000 catalog records.
  - Free to copy and redistribute.
- Represent resources in a linked data format
  - Provide semantic query for time, place, object, etc.
  - Enrich resources by linking them to third-party datasets.

Linked Data

- Linked Data (from Wikipedia)
  - A method of publishing structured data.
  - It can be interlinked and become more useful through semantic queries.
  - **Linked Open Data** is linked data that is open content.
  - Mostly in the form of **RDF**.
- RDF (from W3C RDF 1.1 Primer)
  - Resource Description Framework
  - A framework for expressing information about resources.
  - RDF can enrich a dataset by linking it to third-party datasets.
  - Ex. Enrich a dataset about paintings by linking them to the corresponding artists in *Wikidata*.

RDF Data Model

- A Triple: <subject> <predicate> <object>
  - <Bob> <is a> <person>.
  - <Bob> <is interested in> <the Mona Lisa>.
  - <the Mona Lisa> <was created by> <Leonardo da Vinci>.



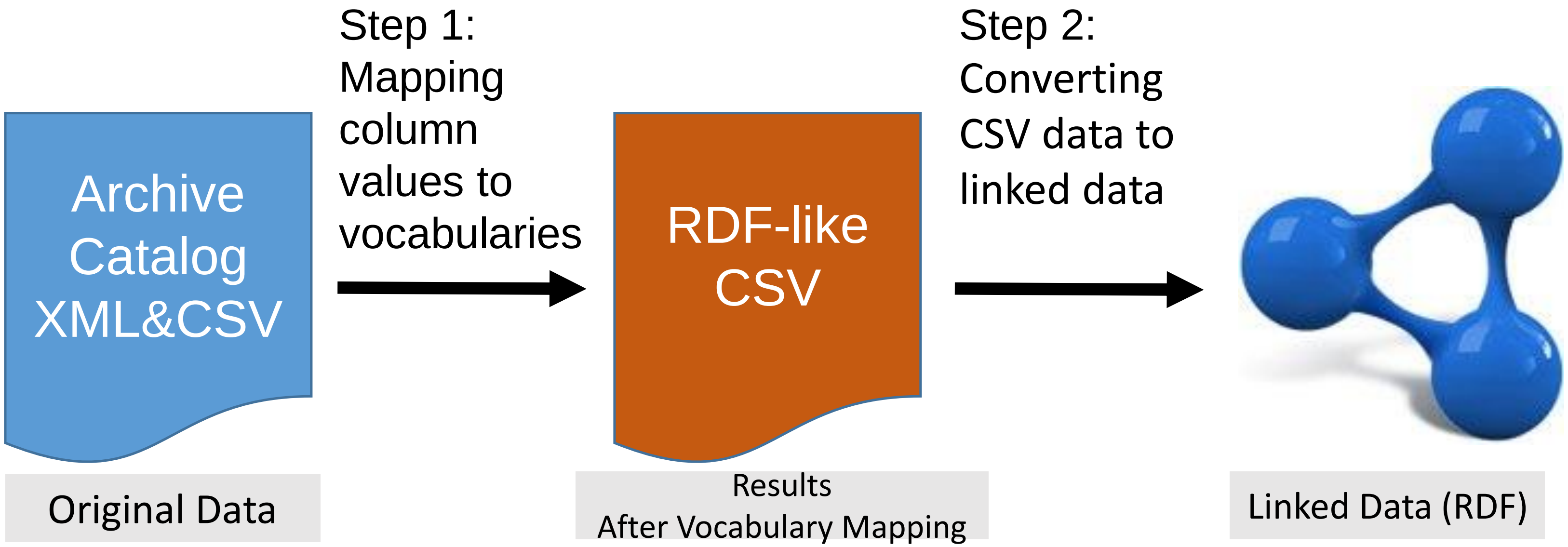
Source: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-triple

From Archive Catalog to Linked Data

- We converted archive catalog to two versions of linked data.
- Version D: triples with just Dublin Core descriptions from the catalog
  - D means *Dublin Core*
- Version R: mapping column values in the catalog to external datasets (with domain vocabularies) to give enriched semantics
  - R means *Refined*
  - Extract place names from "Coverage" column (dc:coverage) in the catalog and map them to place IDs on geonames.org.
  - Normalize values in "Date" column (dc:date) to ISO8601 format, or map them to Wikidata IDs.
  - Map titles of biology archives to entries on Encyclopedia of Life.

Vocabulary Mapping and Data Conversion

Python Scripts: https://gitlab.com/iislod/dat2ld



Archive Catalog XML&CSV
Step 1: Mapping column values to vocabularies
RDF-like CSV
Step 2: Converting CSV data to linked data
Original Data
Results After Vocabulary Mapping
Linked Data (RDF)

Title	台灣一葉蘭
Date::field	採集日期
Date	1993-04-25

txn:hasEOLPage	eol:1134120
rdf:type	schema:CreateAction
skos:editorialNote	採集日期
dwc:eventDate	1993-04-25

```
txn:hasEOLPage
<http://eol.org/pages/1134120> ;
-----------------------------------------------
skos:editorialNote "採集日期" ;
dwc:eventDate "1993-04-25" ;
```

- "採集日期" means *date collected* in English.

Linked Open Data Repository:

Open Data Web (ODW)

http://data.odw.tw

Ontology* for Open Data Web (Draft)

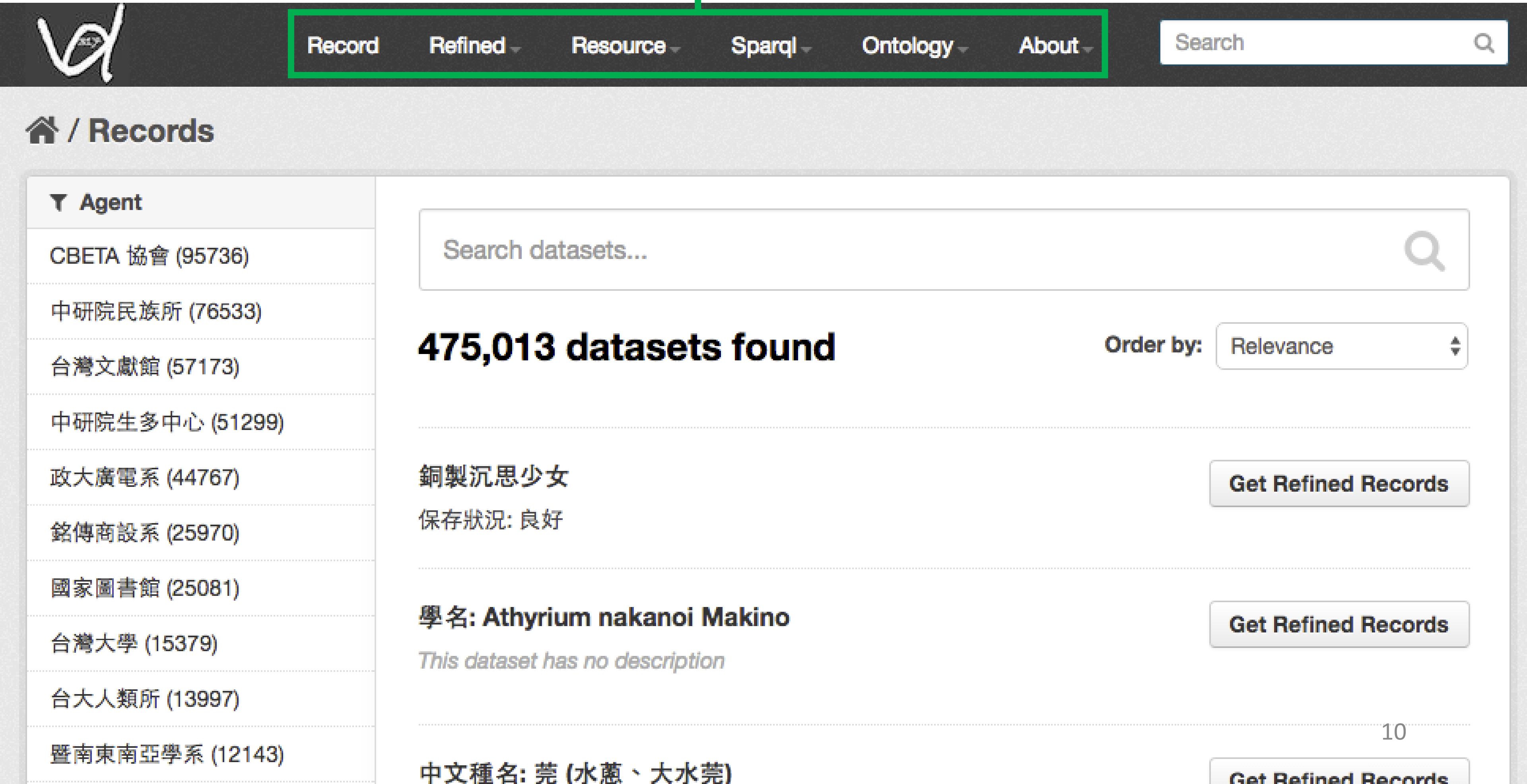
http://voc.odw.tw

* Definitions of the vocabularies used to describe objects in RDF.

Feature (1): Linked Data Browsing

http://data.odw.tw/record/

Main Menu
Records: D version
Refined: R version (still uploading)



/ Records

Agent

CBETA 協會	(95736)
中研院民族所	(76533)
台灣文獻館	(57173)
中研院生多中心	(51299)
政大廣電系	(44767)
銘傳商設系	(25970)
國家圖書館	(25081)
台灣大學	(15379)
台大人類所	(13997)
暨南東南亞學系	(12143)

475,013 datasets found

銅製沉思少女

保存狀況: 良好

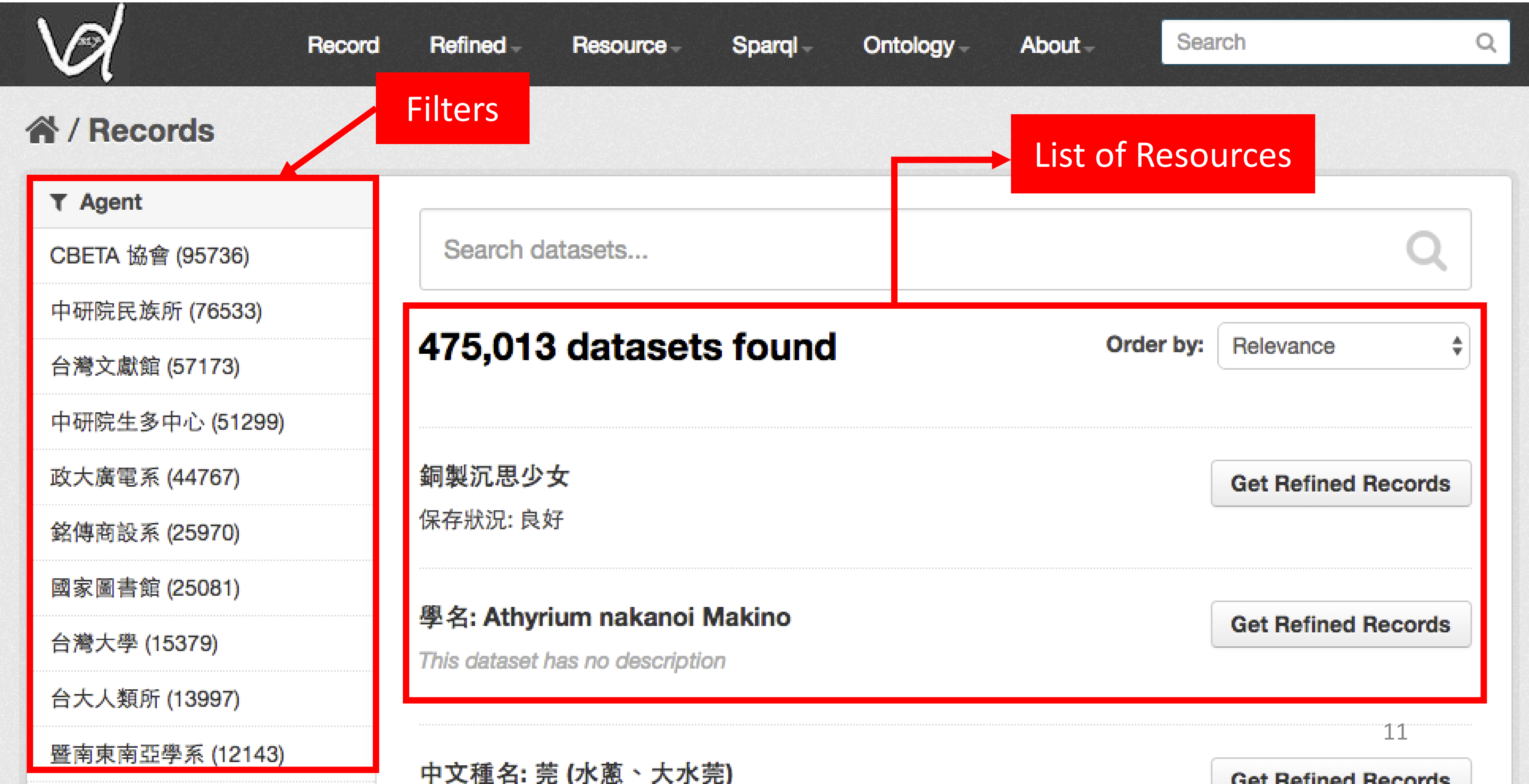
學名: Athyrium nakanoi Makino

This dataset has no description

中文種名: 莞 (水蔥、大水莞)

Feature (1): Linked Data Browsing

http://data.odw.tw/record/



Filters

List of Resources

/ Records

Agent

- CBETA 協會 (95736)
- 中研院民族所 (76533)
- 台灣文獻館 (57173)
- 中研院生多中心 (51299)
- 政大廣電系 (44767)
- 銘傳商設系 (25970)
- 國家圖書館 (25081)
- 台灣大學 (15379)
- 台大人類所 (13997)
- 暨南東南亞學系 (12143)

Search datasets...

475,013 datasets found

Order by: Relevance

銅製沉思少女

保存狀況: 良好

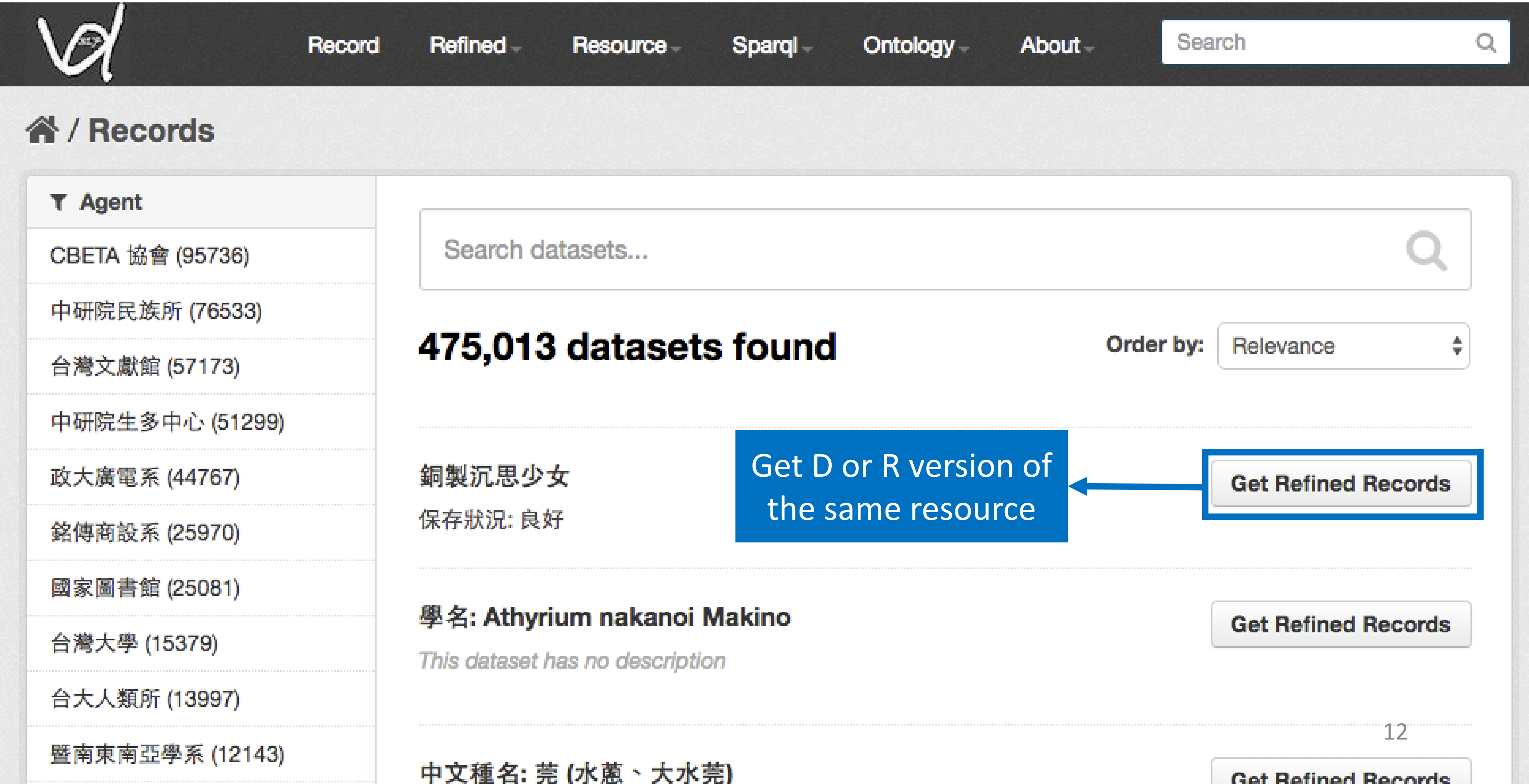
學名: Athyrium nakanoi Makino

This dataset has no description

中文種名: 莞 (水蔥、大水莞)

Feature (1): Linked Data Browsing

http://data.odw.tw/record/



/ Records

Agent

- CBETA 協會 (95736)
- 中研院民族所 (76533)
- 台灣文獻館 (57173)
- 中研院生多中心 (51299)
- 政大廣電系 (44767)
- 銘傳商設系 (25970)
- 國家圖書館 (25081)
- 台灣大學 (15379)
- 台大人類所 (13997)
- 暨南東南亞學系 (12143)

475,013 datasets found

銅製沉思少女

保存狀況: 良好

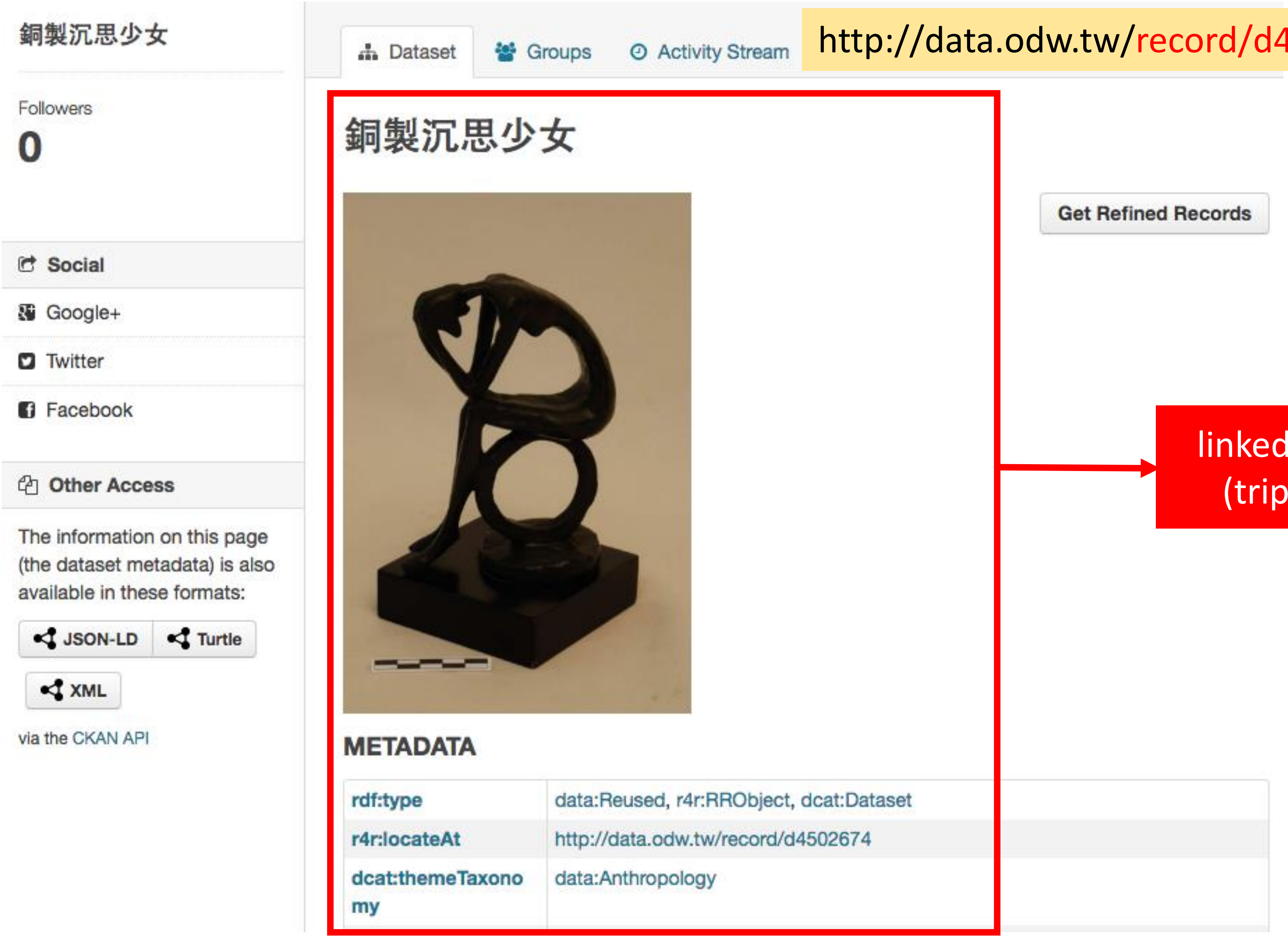
Get D or R version of the same resource

學名: Athyrium nakanoi Makino

This dataset has no description

中文種名: 莞 (水蔥、大水莞)

Example: “Girl Lost in Thought”



銅製沉思少女

Followers

0

Social

- Google+
- Twitter
- Facebook

Other Access

The information on this page (the dataset metadata) is also available in these formats:

JSON-LD | Turtle | XML

via the CKAN API

Dataset | Groups | Activity Stream

http://data.odw.tw/record/d4502674

銅製沉思少女

Get Refined Records

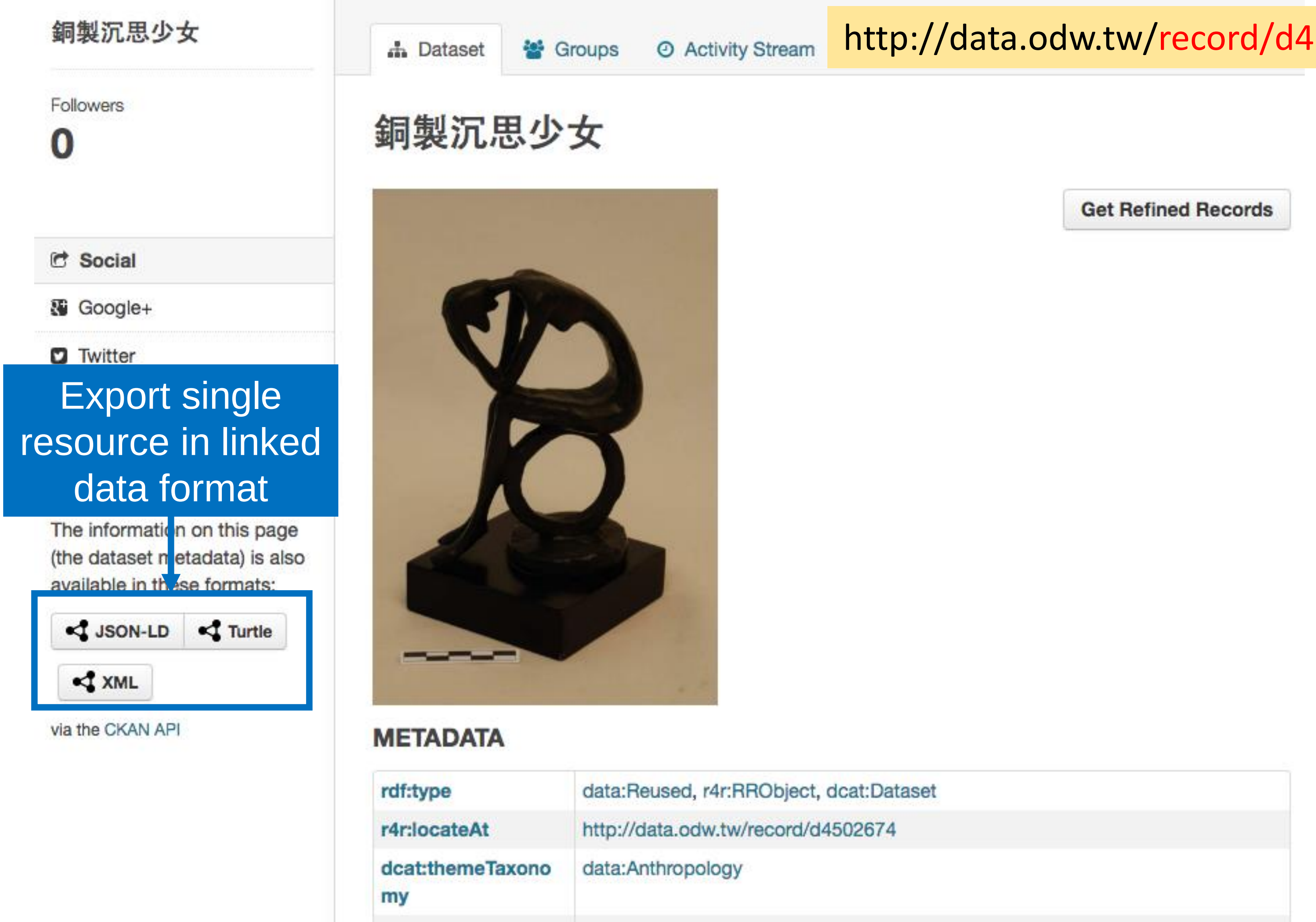
METADATA

rdf:type	data:Reused, r4r:RRObject, dcat:Dataset
r4r:locateAt	http://data.odw.tw/record/d4502674
dcat:themeTaxonomy	data:Anthropology

linked data (triples)

Example: “Girl Lost in Thought”

http://data.odw.tw/record/d4502674



銅製沉思少女

Followers
0

Social
Google+
Twitter

Export single resource in linked data format

The information on this page (the dataset metadata) is also available in these formats:

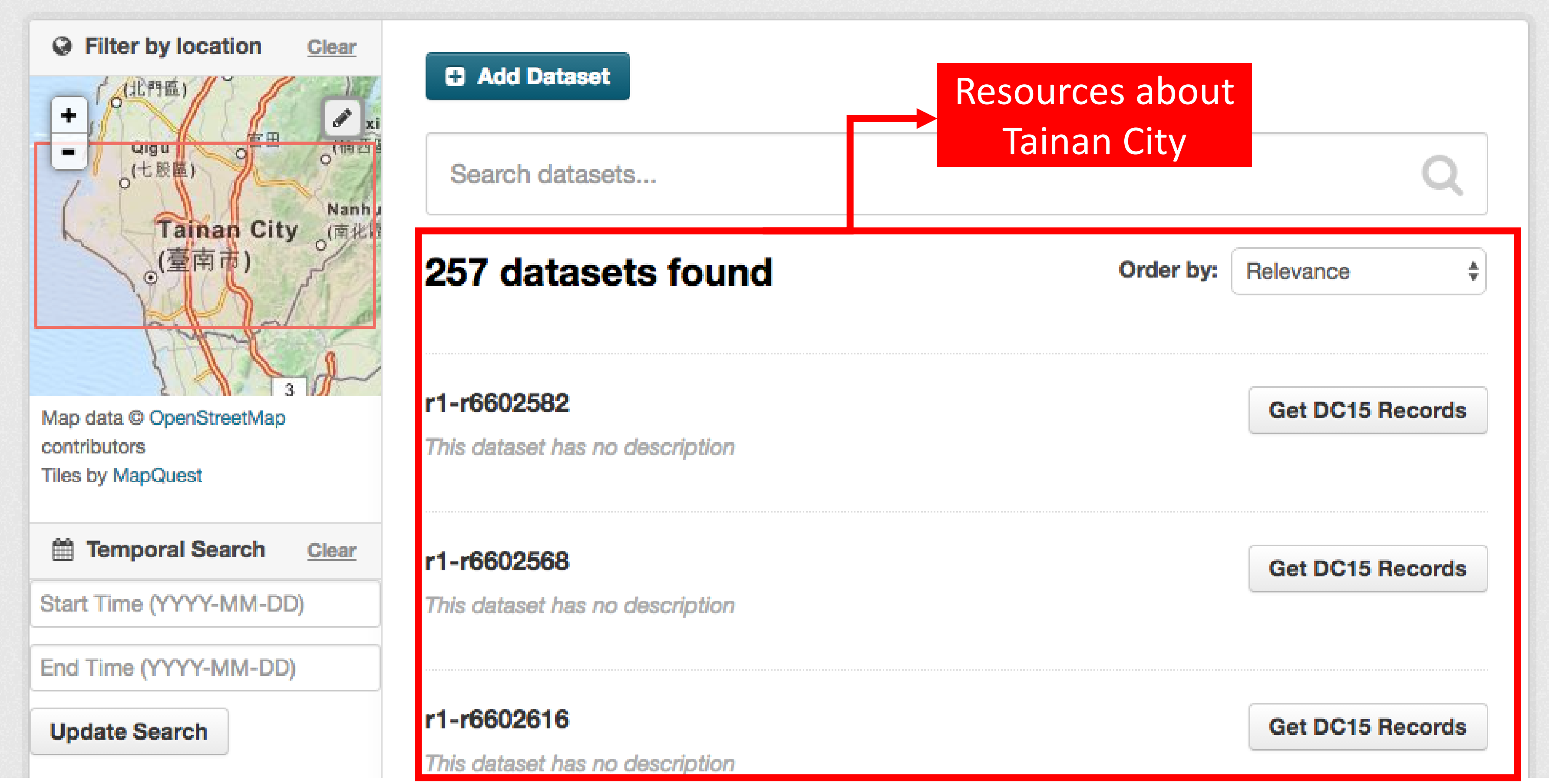
via the CKAN API

銅製沉思少女

METADATA

rdf:type	data:Reused, r4r:RRObject, dcat:Dataset
r4r:locateAt	http://data.odw.tw/record/d4502674
dcat:themeTaxonomy	data:Anthropology

Feature (2): Spatial Query



257 datasets found

r1-r6602582

This dataset has no description

r1-r6602568

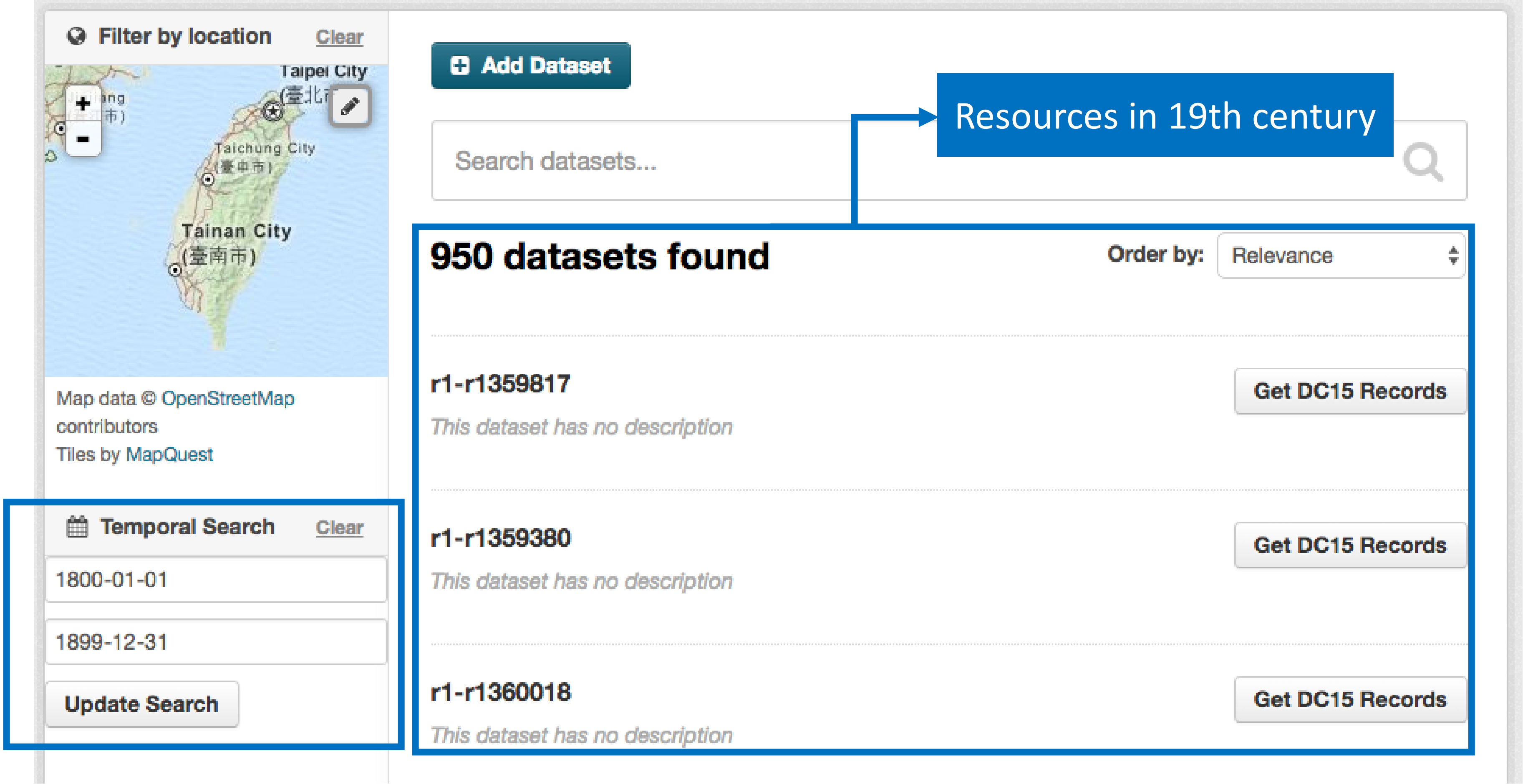
This dataset has no description

r1-r6602616

This dataset has no description

- Spatial indexing based on geo:lat and geo:long values.

Feature (3): Temporal Query



Resources in 19th century

Map data © OpenStreetMap contributors
Tiles by MapQuest

Temporal Search

1800-01-01

1899-12-31

950 datasets found

r1-r1359817

This dataset has no description

r1-r1359380

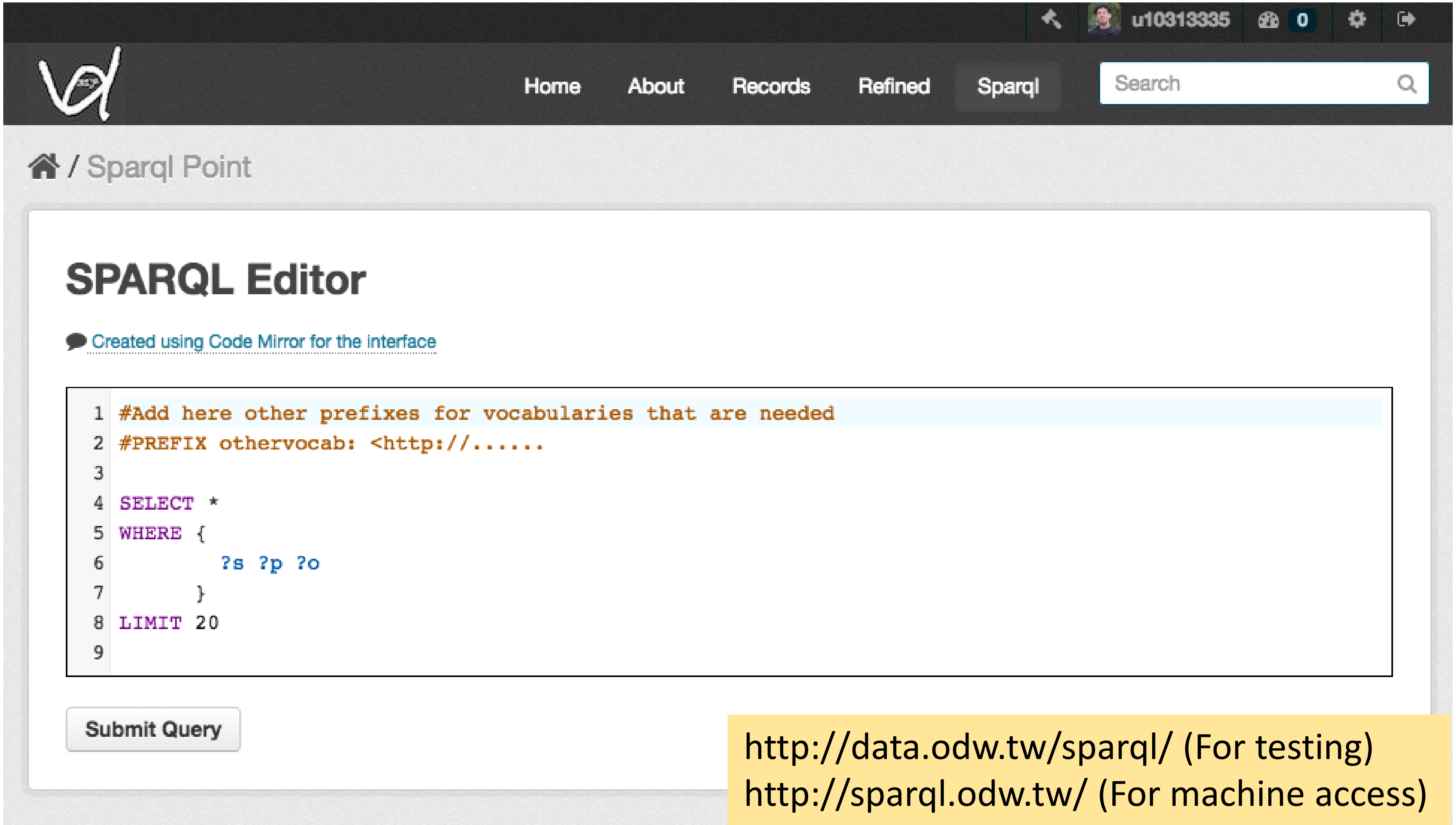
This dataset has no description

r1-r1360018

This dataset has no description

- Temporal indexing based on dct:W3CDTF, xsd:date, and xsd:gYear values.

Feature (4): SPARQL Endpoint



SPARQL Editor

Created using Code Mirror for the interface

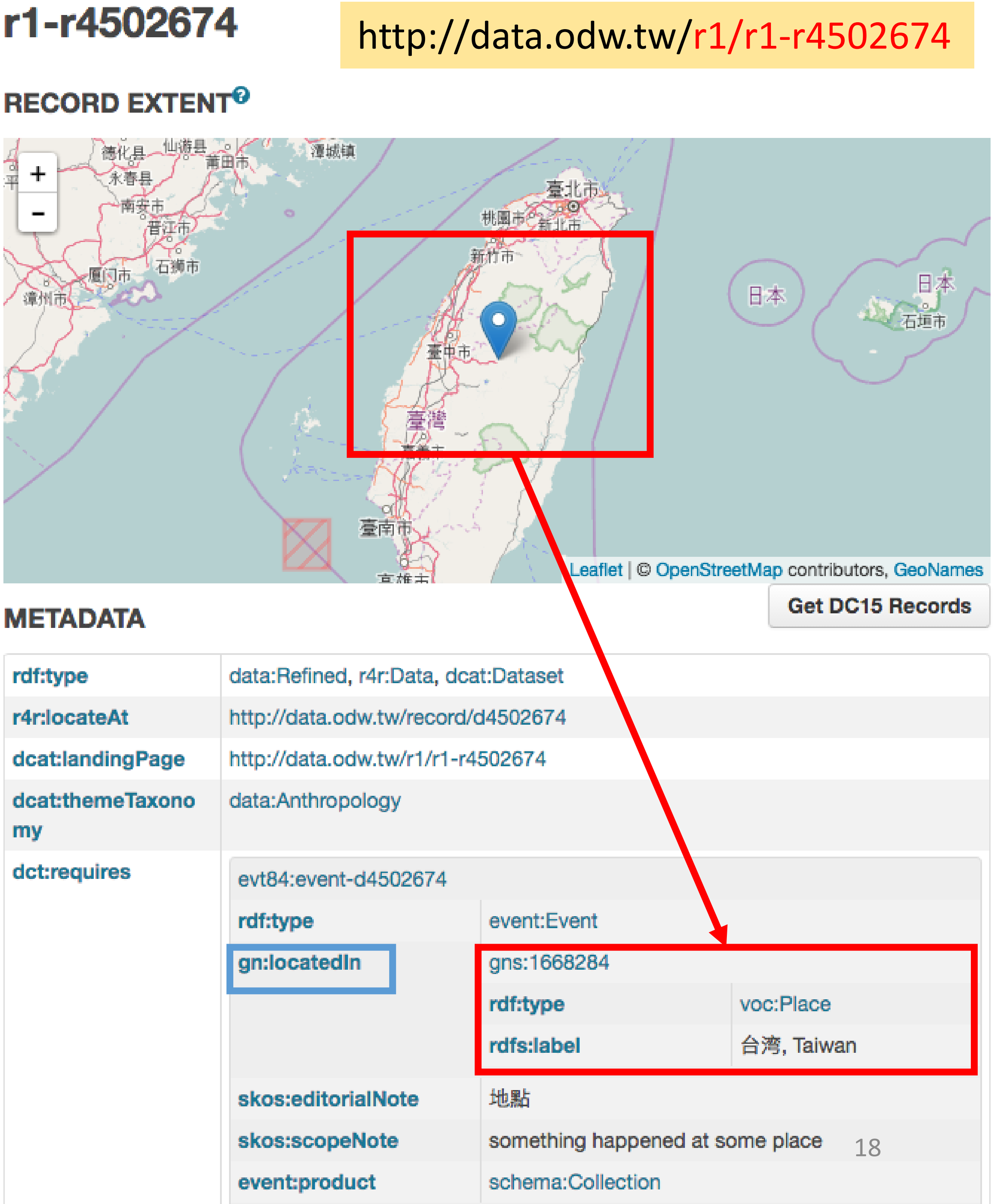
```
#Add here other prefixes for vocabularies that are needed
#PREFIX othervocab: <http://......

SELECT *
WHERE {
        ?s ?p ?o
      }
LIMIT 20

```

http://data.odw.tw/sparql/ (For testing)
http://sparql.odw.tw/ (For machine access)

Feature (5): Spatial Representation



http://data.odw.tw/r1/r1-r4502674

r1-r4502674

RECORD EXTENT

METADATA

Get DC15 Records

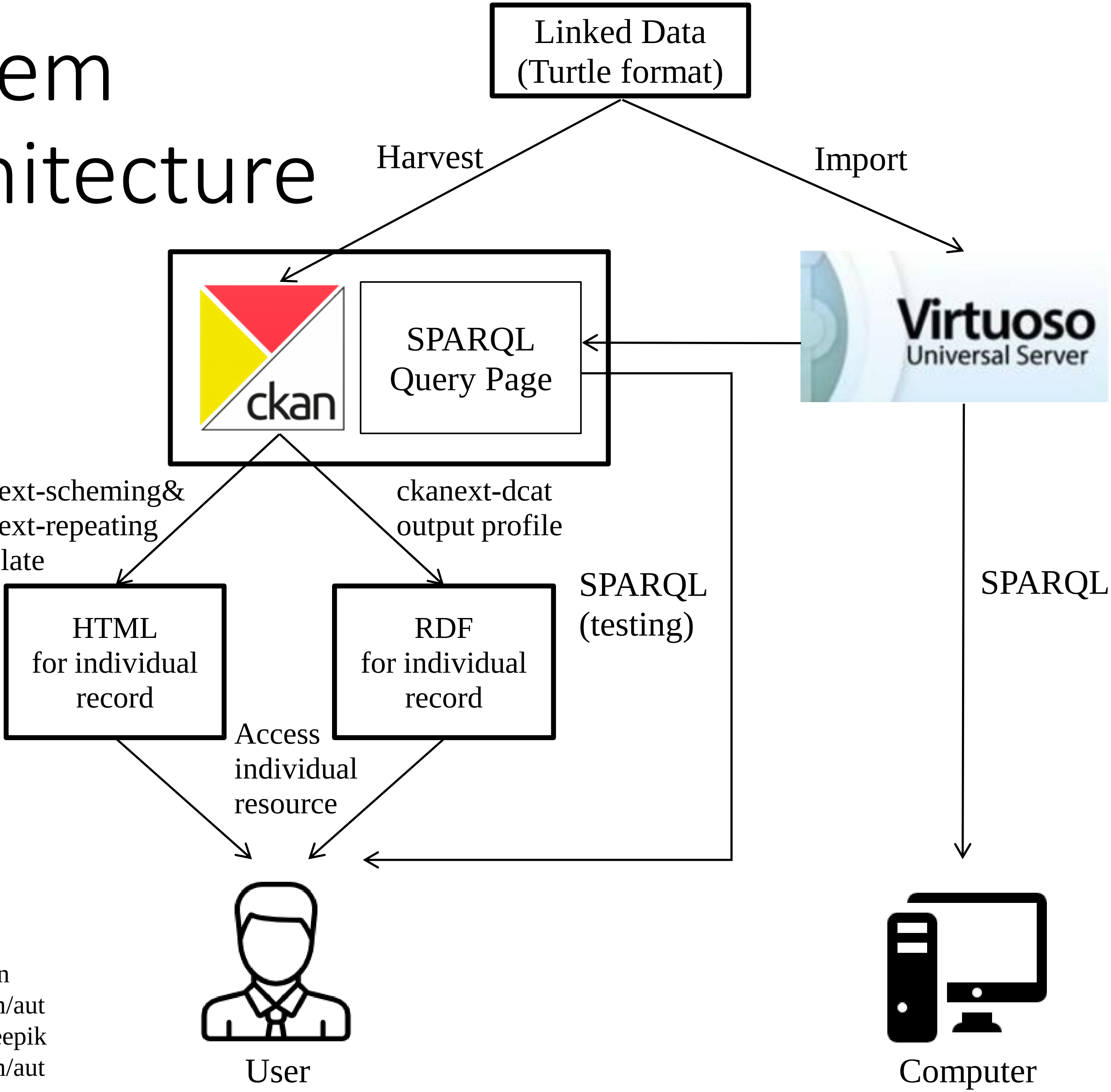
rdf:type	data:Refined, r4r:Data, dcat:Dataset
r4r:locateAt	http://data.odw.tw/record/d4502674
dcat:landingPage	http://data.odw.tw/r1/r1-r4502674
dcat:themeTaxonomy	data:Anthropology
dct:requires	evt84:event-d4502674

evt84:event-d4502674

rdf:type	event:Event	
gn:locatedIn	gns:1668284	
	rdf:type	voc:Place
	rdfs:label	台湾, Taiwan
skos:editorialNote	地點	
skos:scopeNote	something happened at some place	
event:product	schema:Collection	

- Only for R version (still uploading).
- Only shows geonames information in the gn:locatedIn property.

System Architecture



Icon made by SimpleIcon (http://www.flaticon.com/authors/simpleicon) and Freepik (http://www.flaticon.com/authors/freepik)

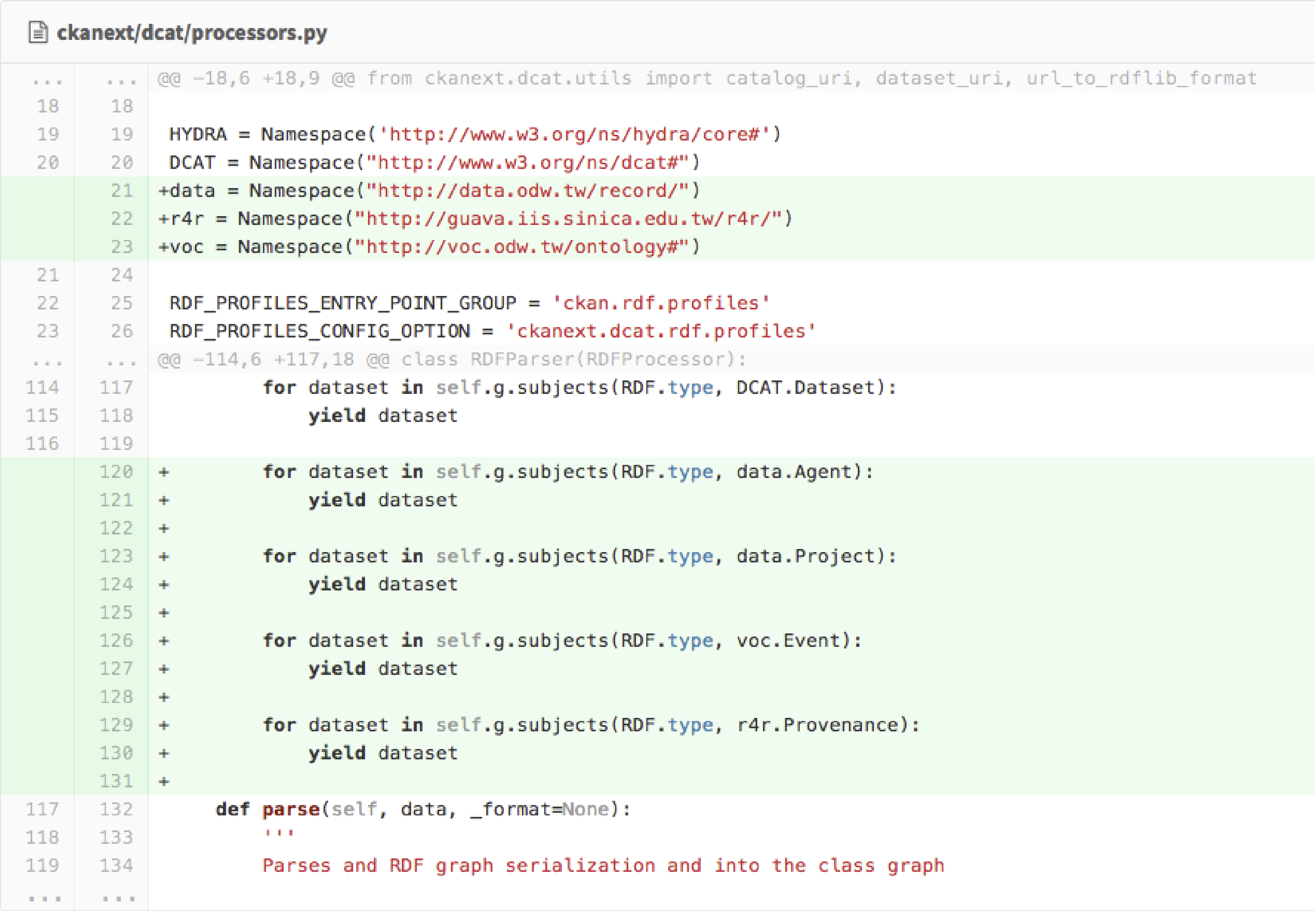
Implementation (1/3)

- Custom fields
  - **ckanext-scheming** and **ckanext-repeating** extension
  - Define CKAN custom fields for a data type in a JSON file
    - Each data type has its own directory.
    - Ex. record.json is for D ver. (http://data.odw.tw/record/)
    - A field is defined by a JSON object, for example:

```
{
  "field_name": "dc:format",
  "label": "dc:format",
  "display_property": "dc:format",
  "preset": "repeating_text_modified”
},
```

Implementation (2/3)

- Import linked data
  - **ckanext-dcat** extension for linked data import/export
  - CKAN harvesting mechanism by ckanext-harvest extension
    - Extend **DCATRDFHarvester** in **ckanext.dcat.harvesters.rdf**
  - Extend **RDFProfile** in **ckanext.dcat.profiles**
    - def parse_dataset(self, dataset_dict, dataset_ref):
      - (Import) Parse *dataset_ref* from loaded linked data to CKAN's *dataset_dict*
    - def graph_from_dataset(self, dataset_dict, dataset_ref):
      - (Export) Generate a linked data graph *dataset_ref* from CKAN's *dataset_dict*
  - Modify **ckanext-dcat** itself
    - To support more namespace (ckanext-dcat is originally designed for DCAT vocabularies.)



ckanext/dcat/processors.py

```diff
@@ -18,6 +18,9 @@ from ckanext.dcat.utils import catalog_uri, dataset_uri, url_to_rdflib_format

 HYDRA = Namespace('http://www.w3.org/ns/hydra/core#')
 DCAT = Namespace("http://www.w3.org/ns/dcat#")
+data = Namespace("http://data.odw.tw/record/")
+r4r = Namespace("http://guava.iis.sinica.edu.tw/r4r/")
+voc = Namespace("http://voc.odw.tw/ontology#")

 RDF_PROFILES_ENTRY_POINT_GROUP = 'ckan.rdf.profiles'
 RDF_PROFILES_CONFIG_OPTION = 'ckanext.dcat.rdf.profiles'
@@ -114,6 +117,18 @@ class RDFParser(RDFProcessor):
         for dataset in self.g.subjects(RDF.type, DCAT.Dataset):
             yield dataset

+        for dataset in self.g.subjects(RDF.type, data.Agent):
+            yield dataset
+
+        for dataset in self.g.subjects(RDF.type, data.Project):
+            yield dataset
+
+        for dataset in self.g.subjects(RDF.type, voc.Event):
+            yield dataset
+
+        for dataset in self.g.subjects(RDF.type, r4r.Provenance):
+            yield dataset
+
     def parse(self, data, _format=None):
         '''
         Parses and RDF graph serialization and into the class graph
```

Implementation (3/3)

- Virtuoso SPARQL endpoint integration
  - **ckanext-sparql** extension
- Spatial indexing and searching
  - **ckanext-spatial** extension
- Time indexing and searching
  - We developed the **ckanext-tempsearch** extension.
- Source code available on GitLab.
  - https://gitlab.com/iislod/

Limitations

- Maintaining two triple stores (CKAN & Virtuoso).
  - They may be inconsistent since we do not sync them for now.
- Slow harvesting speed on CKAN.
  - 4 hrs+ for harvesting 20,000 records on a Core i7-2600 3.4 GHz machine (still uploading now).

Future Work

- Provide **native** SPARQL queries in CKAN.
  - Then we do not need Virtuoso anymore.
- Harvest multiple resources as a CKAN dataset
  - To improve import speed.
- Time and place names mappings to third-party datasets
  - Still need further verifications.

Open Data Web (http://data.odw.tw)

E-mail: ask AT odw.tw

We welcome your valuable

comments & suggestions!

Find me at @u10313335, http://about.me/SolLee, cjlee AT iis.sinica.edu.tw

Acknowledgement: Hsin-Ping Chen (k26021409 AT gmail.com)
for processing geonames data.