

“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/rJlcV6Op>

..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

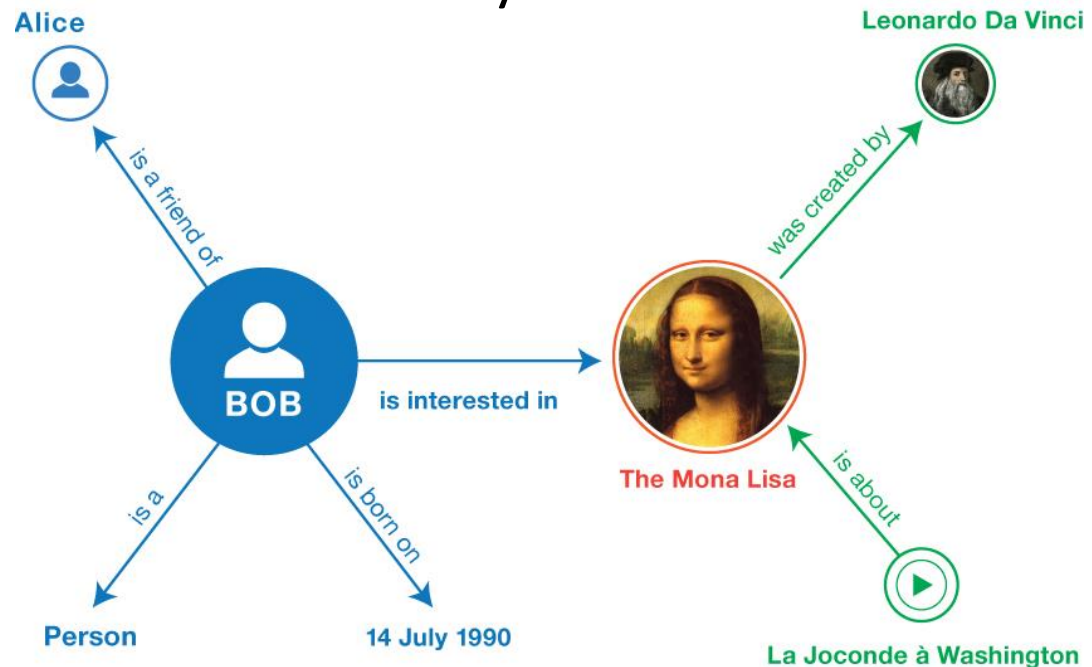
- Union Catalog of Digital Archives Taiwan
 - <http://catalog.digitalarchives.tw>
- Web catalog for digitized archives in 14 domains from many institutions.
- Part of the catalog is released under CC licenses
 - About 840,000 catalog records.
 - Free to copy and redistribute.
- Represent resources in 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>.



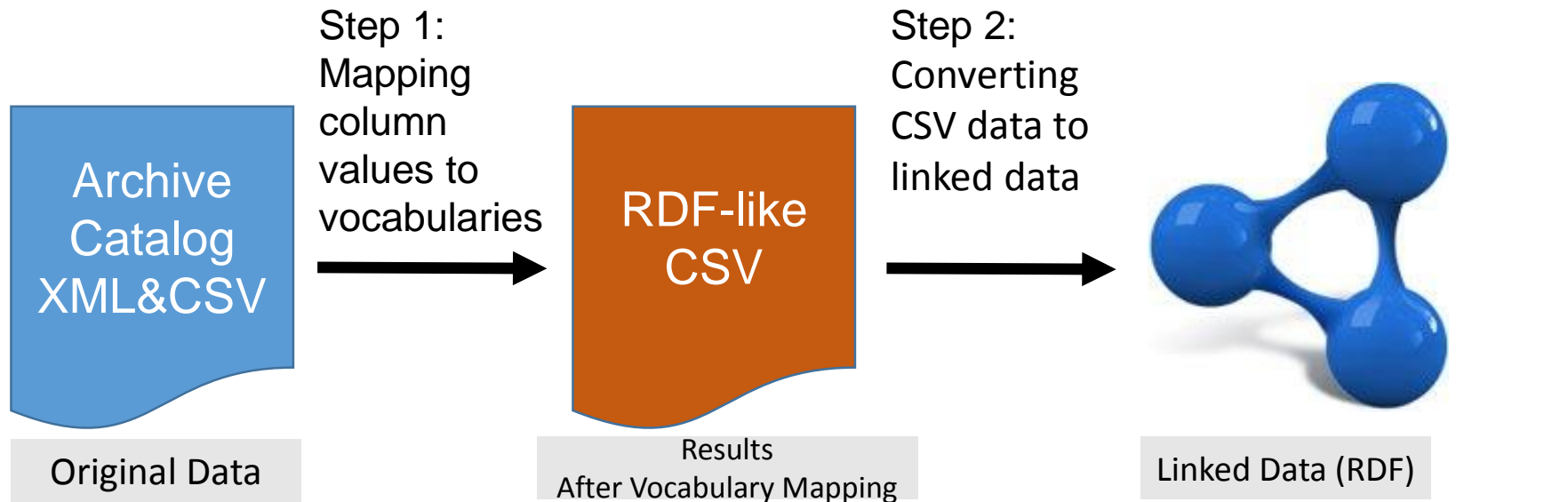
From Archive Catalog to Linked Data

- 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>



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)



Record

Refined

Resource

Sparql

Ontology

About

Search



Home / 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

Get Refined Records

學名: *Athyrium nakanoi* Makino

This dataset has no description

Get Refined Records

中文種名: 蕙 (水蕙、大水蕙)

10

Get Refined Records

Feature (1): Linked Data Browsing

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

The screenshot displays the website's interface. At the top, there is a navigation bar with links for Record, Refined, Resource, Sparql, Ontology, and About, along with a search box. Below the navigation bar, the page title is "/ Records". A red box labeled "Filters" points to a sidebar menu under the "Agent" category, listing various institutions and their record counts. A red box labeled "List of Resources" points to the main content area, which shows a search bar, the number of datasets found (475,013), and a list of individual dataset entries with buttons to refine the records.

Record Refined Resource Sparql Ontology About Search

Home / Records

Filters

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
Get Refined Records

學名: **Athyrium nakanoi Makino**
This dataset has no description
Get Refined Records

11

中文種名: 蕒 (水蕒、大水蕒)
Get Refined Records

Feature (1): Linked Data Browsing

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

The screenshot shows the website interface for data browsing. At the top, there is a navigation bar with links for Record, Refined, Resource, Sparql, Ontology, and About, along with a search box. Below the navigation bar, the page title is 'Records'. On the left side, there is a sidebar with a list of agents, including CBETA 協會, 中研院民族所, 台灣文獻館, 中研院生多中心, 政大廣電系, 銘傳商設系, 國家圖書館, 台灣大學, 台大人類所, and 暨南東南亞學系. The main content area displays a search result for '銅製沉思少女' (Copper Cast Thinking Girl). The search bar at the top of the main area contains the text 'Search datasets...'. Below the search bar, it shows '475,013 datasets found' and 'Order by: Relevance'. The search result for '銅製沉思少女' includes the text '保存狀況: 良好' and a blue callout box that says 'Get D or R version of the same resource'. To the right of the callout box is a button labeled 'Get Refined Records'. Below the search result, there is another entry for '學名: Athyrium nakanoi Makino' with the note 'This dataset has no description' and another 'Get Refined Records' button. At the bottom right, there is a page number '12' and a third 'Get Refined Records' button.

Record Refined Resource Sparql Ontology About Search

Home / 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

Get D or R version of the same resource

Get Refined Records

學名: **Athyrium nakanoi Makino**
This dataset has no description

Get Refined Records

12

中文種名: 蕙 (水蕙、大水蕙)

Get Refined Records

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

linked data
(triples)

METADATA

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

Example: “Girl Lost in Thought”

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:

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

Feature (2): Spatial Query

The screenshot displays a web interface for searching datasets. On the left, there is a map of Tainan City (臺南市) with a red bounding box around it. Below the map, there are controls for 'Filter by location' (with a 'Clear' button), 'Temporal Search' (with a 'Clear' button), and 'Update Search'. The main search area has a search bar with the text 'Search datasets...' and a magnifying glass icon. A red box highlights the search bar, and a red arrow points from a red text box 'Resources about Tainan City' to the search bar. Below the search bar, the results show '257 datasets found' and 'Order by: Relevance'. Three dataset entries are visible, each with a 'Get DC15 Records' button.

Dataset ID	Description	Action
r1-r6602582	<i>This dataset has no description</i>	Get DC15 Records
r1-r6602568	<i>This dataset has no description</i>	Get DC15 Records
r1-r6602616	<i>This dataset has no description</i>	Get DC15 Records

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

Feature (3): Temporal Query

The screenshot displays a web interface for searching datasets. On the left, a map of Taiwan is shown with markers for Taipei City (台北市), Taichung City (臺中市), and Tainan City (臺南市). Below the map is a 'Temporal Search' section with a calendar icon, a 'Clear' link, and two date input fields containing '1800-01-01' and '1899-12-31'. An 'Update Search' button is located below these fields. In the top right, a search bar contains the text 'Resources in 19th century' with a magnifying glass icon. Below the search bar, a blue box highlights the search results, which are ordered by 'Relevance'. The results list three datasets, each with a unique ID, a description of 'This dataset has no description', and a 'Get DC15 Records' button.

Filter by location [Clear](#)

Map data © OpenStreetMap contributors
Tiles by MapQuest

Temporal Search [Clear](#)

1800-01-01

1899-12-31

Update Search

Add Dataset

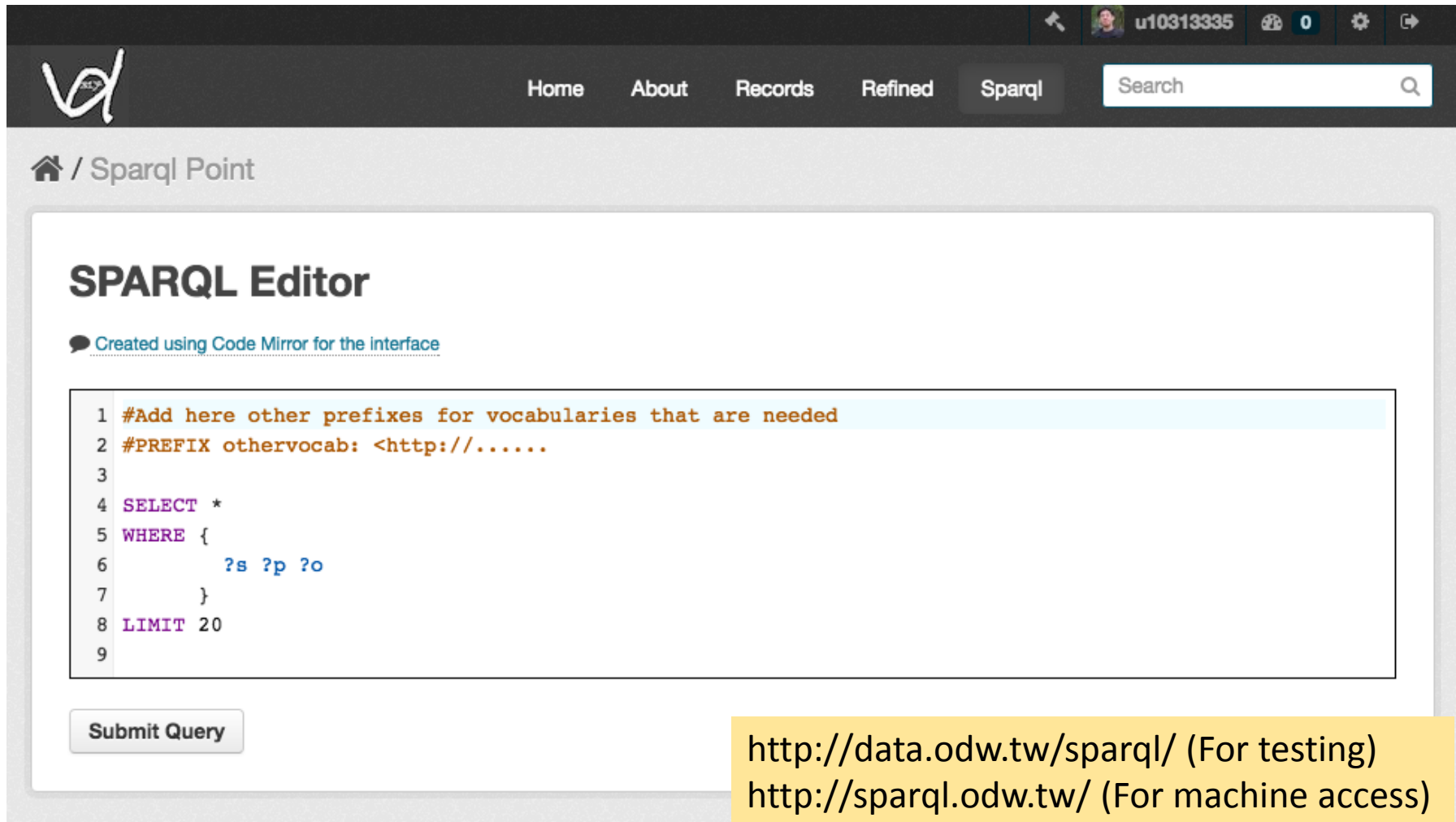
Search datasets... **Resources in 19th century**

950 datasets found Order by: Relevance

r1-r1359817 <i>This dataset has no description</i>	Get DC15 Records
r1-r1359380 <i>This dataset has no description</i>	Get DC15 Records
r1-r1360018 <i>This dataset has no description</i>	Get DC15 Records

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

Feature (4): SPARQL Endpoint



The screenshot shows a web interface for a SPARQL endpoint. At the top, there is a navigation bar with a logo on the left and links for Home, About, Records, Refined, and Sparql. A search bar is located on the right side of the navigation bar. Below the navigation bar, the page title is "Sparql Point". The main content area is titled "SPARQL Editor" and includes a note: "Created using Code Mirror for the interface". A text area contains the following SPARQL query:

```
1 #Add here other prefixes for vocabularies that are needed
2 #PREFIX othervocab: <http://.....
3
4 SELECT *
5 WHERE {
6     ?s ?p ?o
7 }
8 LIMIT 20
9
```

Below the text area is a "Submit Query" button. At the bottom right of the interface, there is a yellow box containing the following URLs:

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

Feature (5): Spatial Representation

r1-r4502674

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

RECORD EXTENT



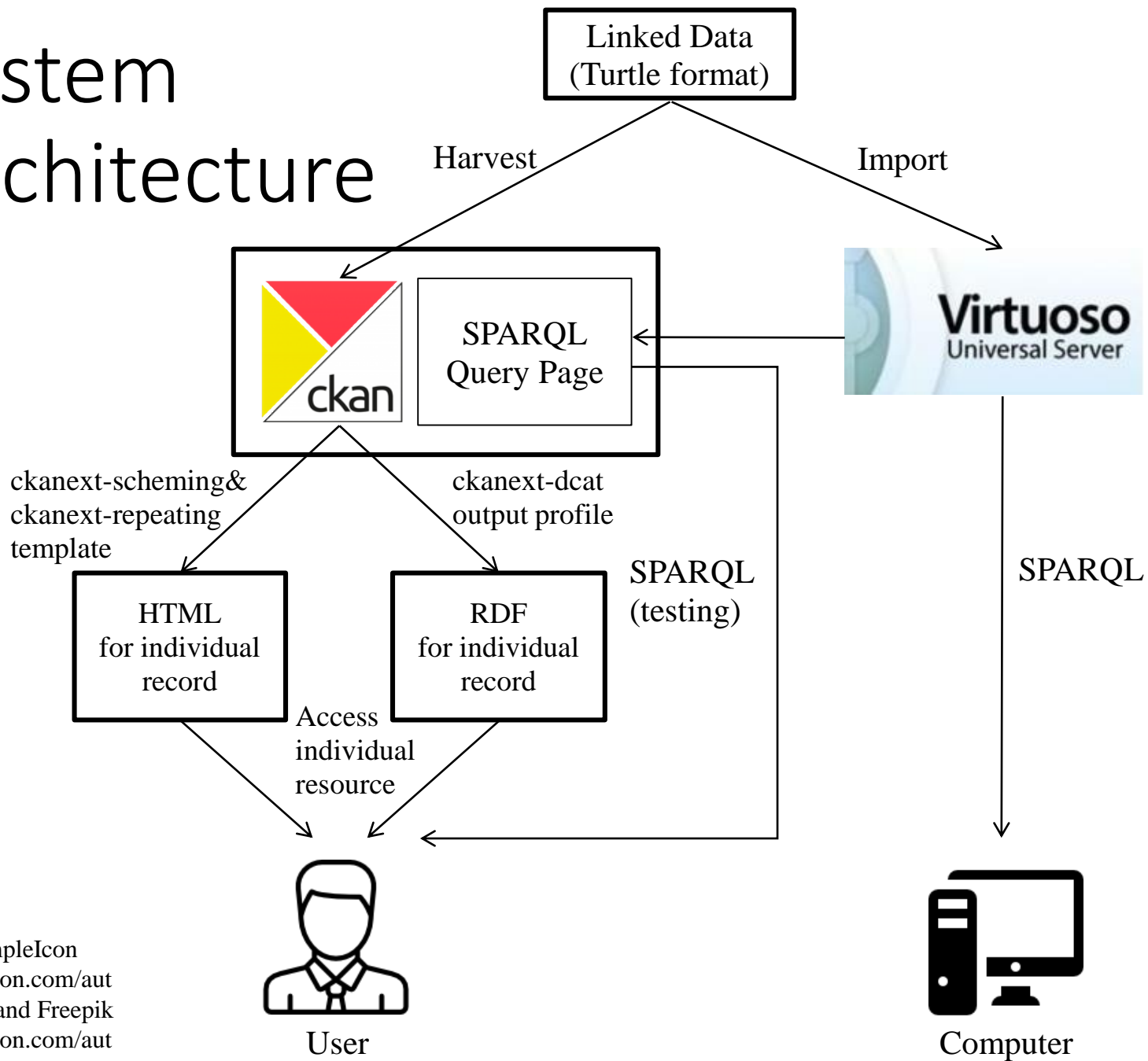
METADATA

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	
	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 18
	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/aut hors/simpleicon>) and Freepik
(<http://www.flaticon.com/aut hors/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 **DCAT RDF Harvester** 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

```
...     ... @@ -18,6 +18,9 @@ from ckanext.dcat.utils import catalog_uri, dataset_uri, url_to_rdfli_format
18     18
19     19     HYDRA = Namespace('http://www.w3.org/ns/hydra/core#')
20     20     DCAT = Namespace("http://www.w3.org/ns/dcat#")
21     21     +data = Namespace("http://data.odw.tw/record/")
22     22     +r4r = Namespace("http://guava.iis.sinica.edu.tw/r4r/")
23     23     +voc = Namespace("http://voc.odw.tw/ontology#")
21     24
22     25     RDF_PROFILES_ENTRY_POINT_GROUP = 'ckan.rdf.profiles'
23     26     RDF_PROFILES_CONFIG_OPTION = 'ckanext.dcat.rdf.profiles'
...     ... @@ -114,6 +117,18 @@ class RDFParser(RDFProcessor):
114    117         for dataset in self.g.subjects(RDF.type, DCAT.Dataset):
115    118             yield dataset
116    119
120    +         for dataset in self.g.subjects(RDF.type, data.Agent):
121    +             yield dataset
122    +
123    +         for dataset in self.g.subjects(RDF.type, data.Project):
124    +             yield dataset
125    +
126    +         for dataset in self.g.subjects(RDF.type, voc.Event):
127    +             yield dataset
128    +
129    +         for dataset in self.g.subjects(RDF.type, r4r.Provenance):
130    +             yield dataset
131    +
117    132     def parse(self, data, _format=None):
118    133         ...
119    134         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.