Semantics and Linkage of Archive(d) Catalogs

Knowledge Organization and Cultural Heritage: Perspectives of Semantic Web Workshop
Taipei, June 2 2016

Tyng-Ruey Chuang, Andrea Wei-Ching Huang, Cheng-Jen Lee and Hsin-Ping Chen
Institute of Information Science, Academia Sinica, Taipei, Taiwan.

1. Archive(d) Catalogs

Archive(d) Catalogs

- Interface of Archive(d) Catalogs for General Public.
- Researchers need references from different institutions.
- Maintenance & Updates：Catalogs become Archived Objects！

- catalog.digitalarchives.tw：resources from 14 domains
- Part of catalogs adapt CC licenses：free to copy and distribute
  - Representations & Linked Data of the Archive(d) Catalogs
  - Semantic Query for Time, Place, People and Object
  - Implementation of the Linked Data

XML – RDF – CSV

- XML
  - Nested Hierarchy；Text & Markup in the same document
  - Markup language for document structure is customizable
- RDF
  - Node Network: ；individual-relation-individual（individual—attribute—value）
  - Semantic relation (attribute) is customizable
- CSV
  - Tabular Data；homogeneous grouping for the individual attribute of tables
  - Flexible column-aligned；fixed-column numbers

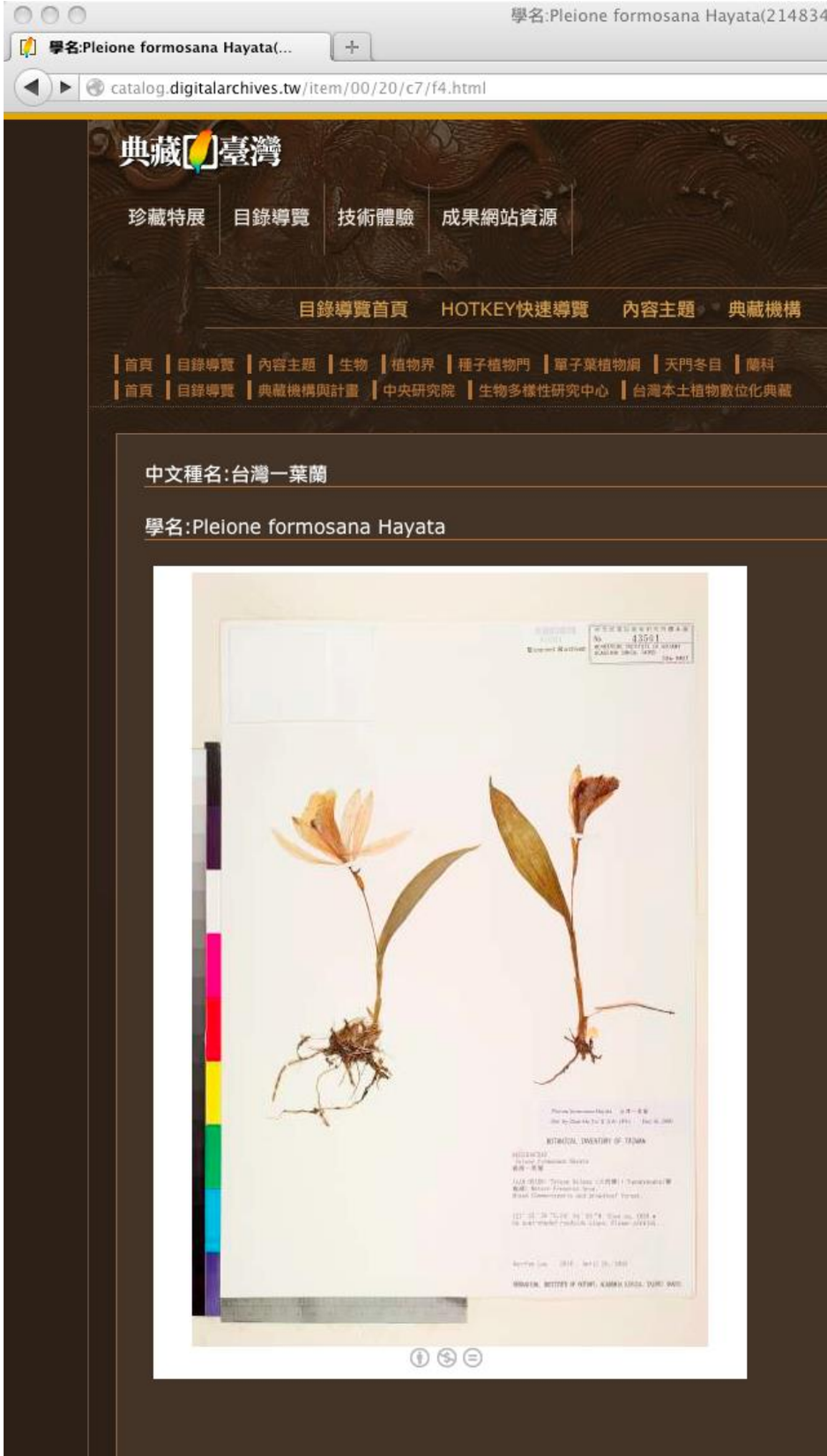
XML – eXtensible Markup Language
RDF – Resource Description Framework
CSV – Comma-Separated Values

From XML to RDF via CSV

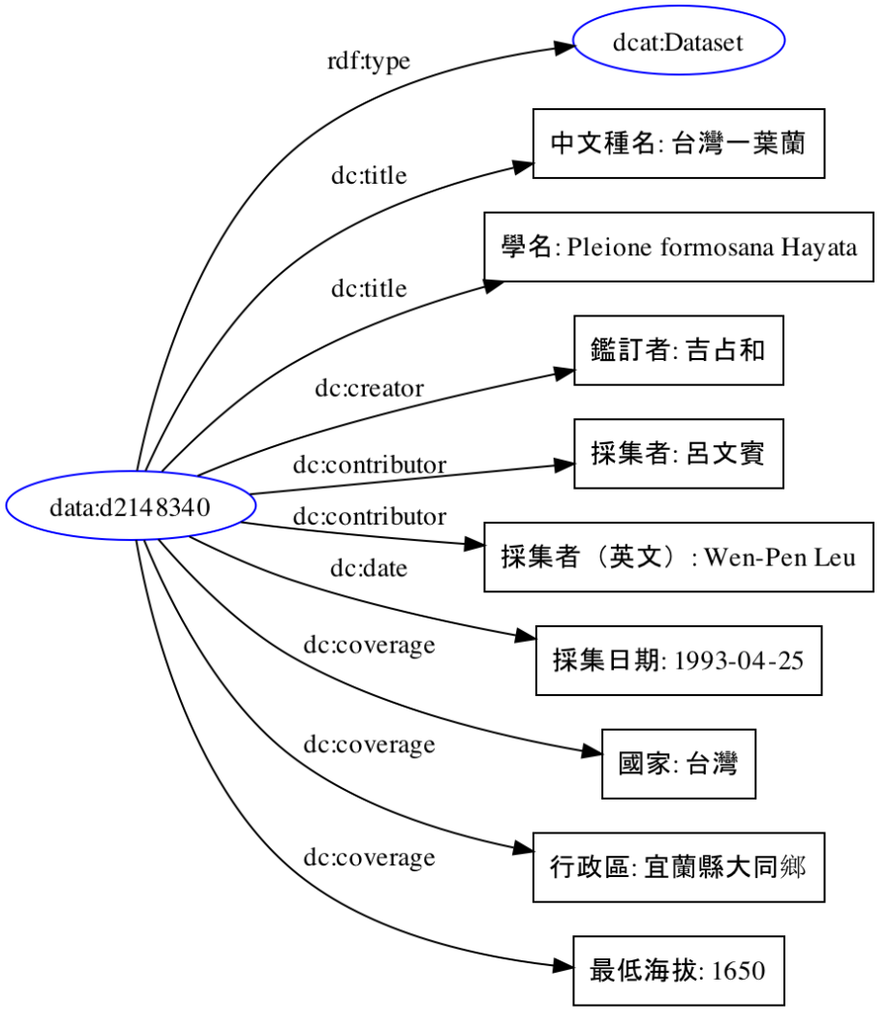
- Structure of Catalogs: XML
  - non-recursive、non-nested type document；fixed vocabulary；textual records
- Medium Tabular Data: CSV
  - mapping XML vocabulary to RDF vocabulary；catalogs in sequence of raw and column
- Linked Data: RDF
  - convert csv as linked data ；use domain vocabularies for data relation

- In focus:
  - CSV can refer to other CSV resources；CSV is human editable；Process of editing CSV is manageable；software tools are plenty available ；linked data generation is flexible.

```xml
<?xml version="1.0" encoding="BIG5"?>
<DACatalog>
  <AdminDesc>
    <Project Creator="中研院生物多樣性中心" GenDate="2011-05-13">台灣本土植物數位化典藏</Project>
    <Catalog>
      <Record>典藏機構與計畫:中央研究院:生物多樣性研究中心:台灣本土植物數位化典藏</Record>
      <Record>內容主題:生物:植物界:種子植物門:單子葉植物綱:天門冬目:蘭科</Record>
    </Catalog>
    <DigiArchiveID>43501</DigiArchiveID>
    <Hyperlink>http://www.hast.biodiv.tw/specimens/SpecimenDetailC.aspx?specimenOrderNum=43501</Hyperlink>
    <ICON license="CC2.5:BY-NC-ND">http://img.hast.biodiv.tw/specimenSmall/specimenSmall004/3/S_043501.jpg</ICON>
  </AdminDesc>
  <MetaDesc license="CC2.5:BY-NC-ND">
    <Title field="中文種名">台灣一葉蘭</Title>
    <Title field="學名">Pleione formosana Hayata</Title>
    <Creator field="鑑訂者">吉占和</Creator>
    <Contributor field="採集者">呂文賓</Contributor>
    <Contributor field="採集者(英文)">Wen-Pen Leu</Contributor>
    <Subject field="界">植物界</Subject>
    <Subject field="界(英文)">Plantae</Subject>
    <Subject field="門">種子植物門</Subject>
    <Subject field="門(英文)">Spermatophyta</Subject>
    <Subject field="綱">單子葉植物綱</Subject>
    <Subject field="綱(英文)">Monocotyledons</Subject>
    <Subject field="目">天門冬目</Subject>
    <Subject field="目(英文)">Asparagales</Subject>
    <Subject field="科">蘭科</Subject>
    <Subject field="科(英文)">ORCHIDACEAE</Subject>
    <Subject field="屬">一葉蘭屬</Subject>
    <Subject field="屬(英文)">Pleione</Subject>
    <Date field="採集日期">1993-04-25</Date>
    <Coverage field="國家">台灣</Coverage>
    <Coverage field="行政區">宜蘭縣大同鄉</Coverage>
    <Coverage field="最低海拔">1650</Coverage>
    <Identifier field="標本館號">43501</Identifier>
    <Identifier field="編目號">Wen-Pen Leu 2018</Identifier>
    <Publisher>中央研究院生物多樣性研究中心</Publisher>
    <Source>台灣本土植物資料庫(http://taiwanflora.sinica.edu.tw/)</Source>
    <Language>中文</Language>
    <Rights>中央研究院 生物多樣性研究中心 植物標本館 Herbarium, Research Center for Biodiversity, Academia Sinica, Taipei (HAST)</Rights>
  </MetaDesc>
</DACatalog>
```

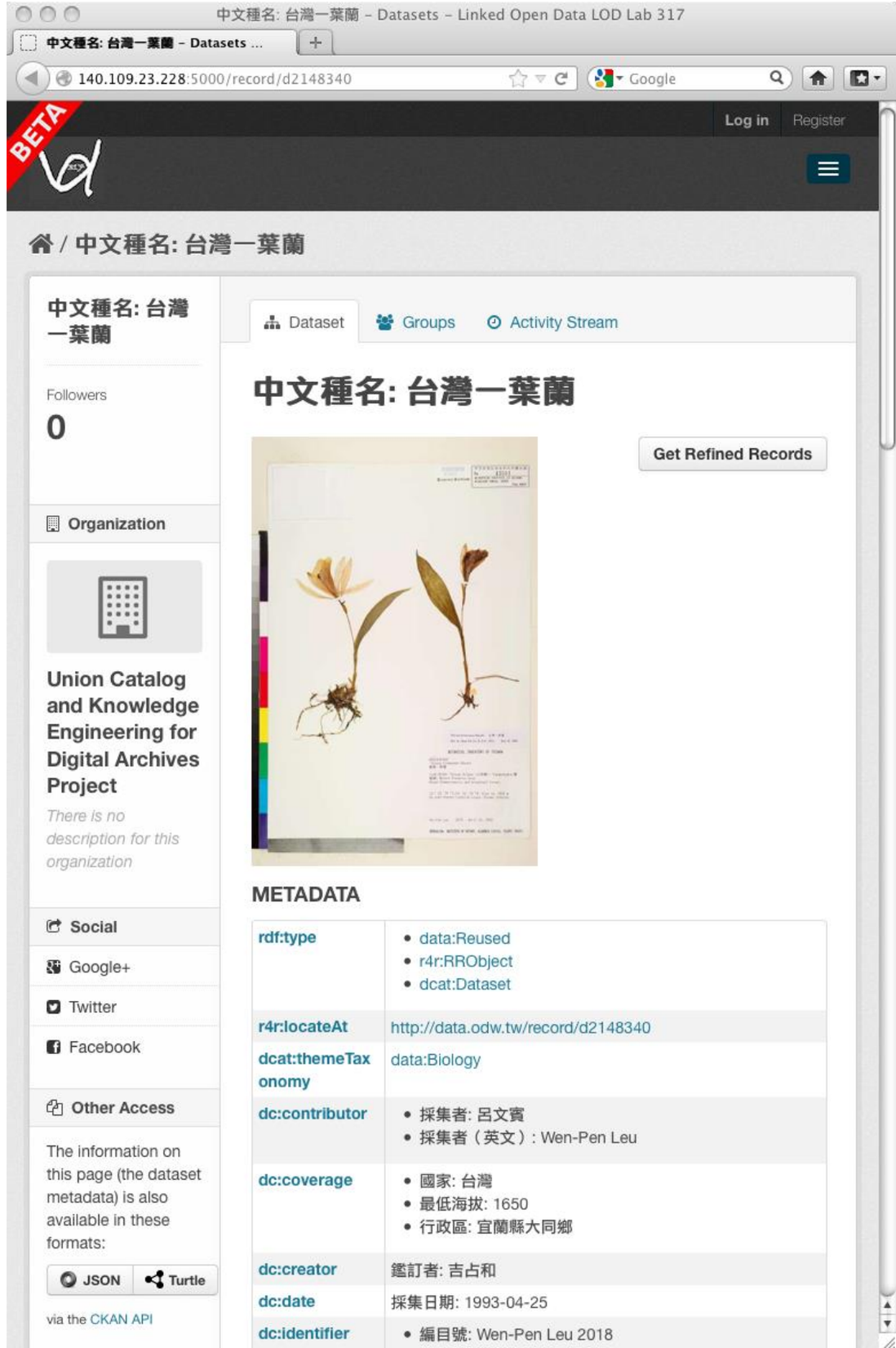


OID	Title::field	Title	Creator::field	Creator	Contributor::field	Contributor	Date::field	Date	Coverage::field	Coverage
2148340	中文種名	台灣一葉蘭								
2148340	學名	Pleione formosana Hayata								
2148340			鑑訂者	吉占和						
2148340					採集者	呂文賓				
2148340					採集者（英文）	Wen-Pen Leu				
2148340							採集日期	1993-04-25		
2148340									國家	台灣
2148340									行政區	宜蘭縣大同鄉
2148340									最低海拔	1650



Model:
(Unknown)

Namespaces:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
data: http://data.odw.tw/record/
dc: http://purl.org/dc/elements/1.1/
dcat: http://www.w3.org/ns/dcat#



2. A System for Linked Data

http://data.odw.tw/

From Archive(d) Catalogs to Linked Data

- Representation of Archive(d) Catalogs via Linked Data (RDF objects are pure text)
  - This is version D. (D is for DC15.)

- Mapping and linking external resources with domain vocabularies for enriched and refined semantics
  - This is version R. (R is for Refined.)
  - Extract place names from "Coverage" (dc:coverage), and map them to place IDs on geonames.org.
  - Normalize values in "Date" column (dc:date) using ISO 8601, or map time related terms to Wikidata IDs.
  - Map titles of biology items to entries on Encyclopedia of Life.

- First by automatic processes, then manual adjustments to the processes.

Vocabulary Mapping and Data Conversion:

Original Data

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

After Vocabulary Mapping

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

Linked Data (RDF)

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

- ❑ Use “profiles” to define mappings.
- ❑ Vocabularies are changeable through profiles.

Storage and Representation

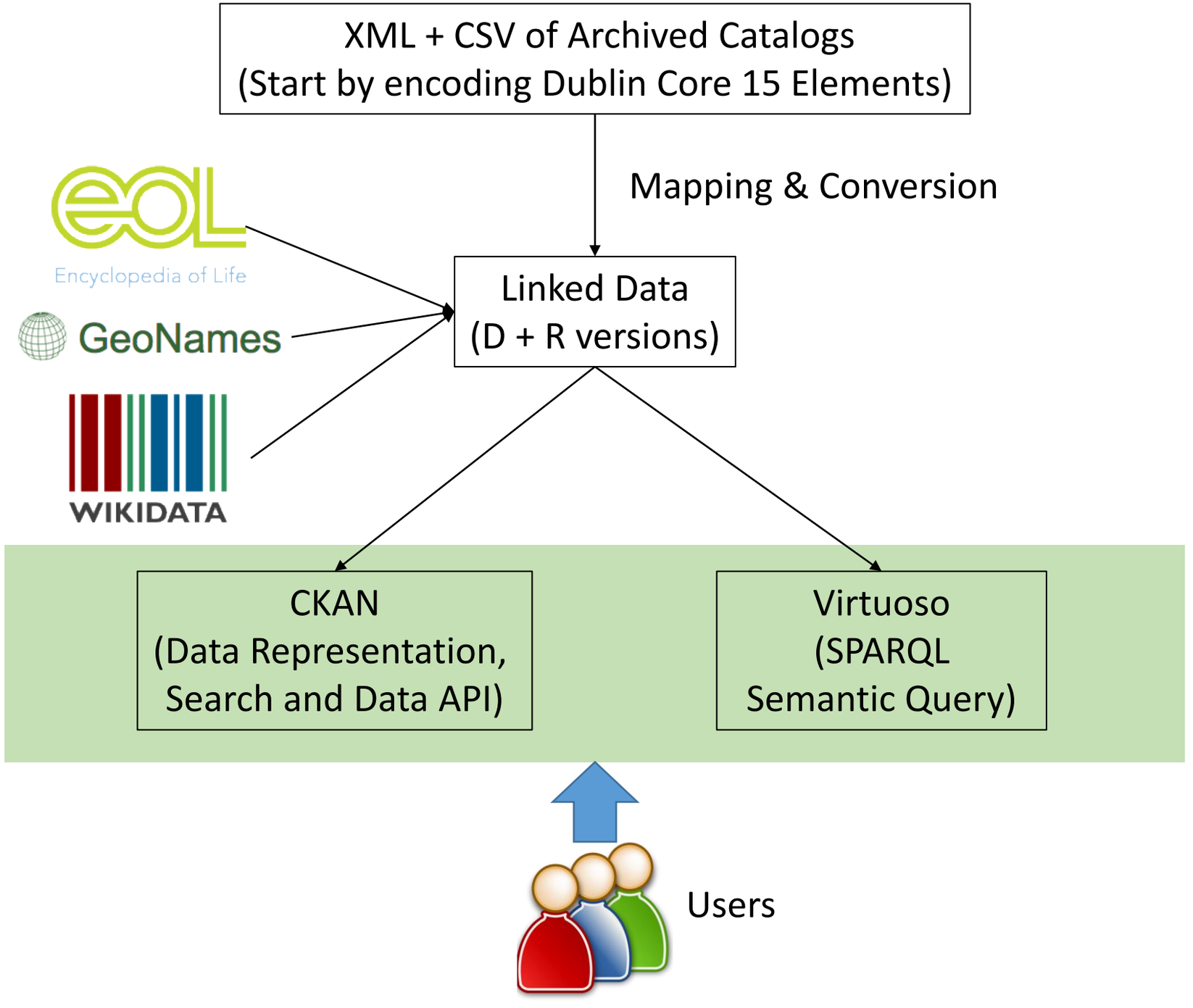
For Machine Access

SPARQL Endpoint
High Performance
High Reliability

For Linked Data Browser

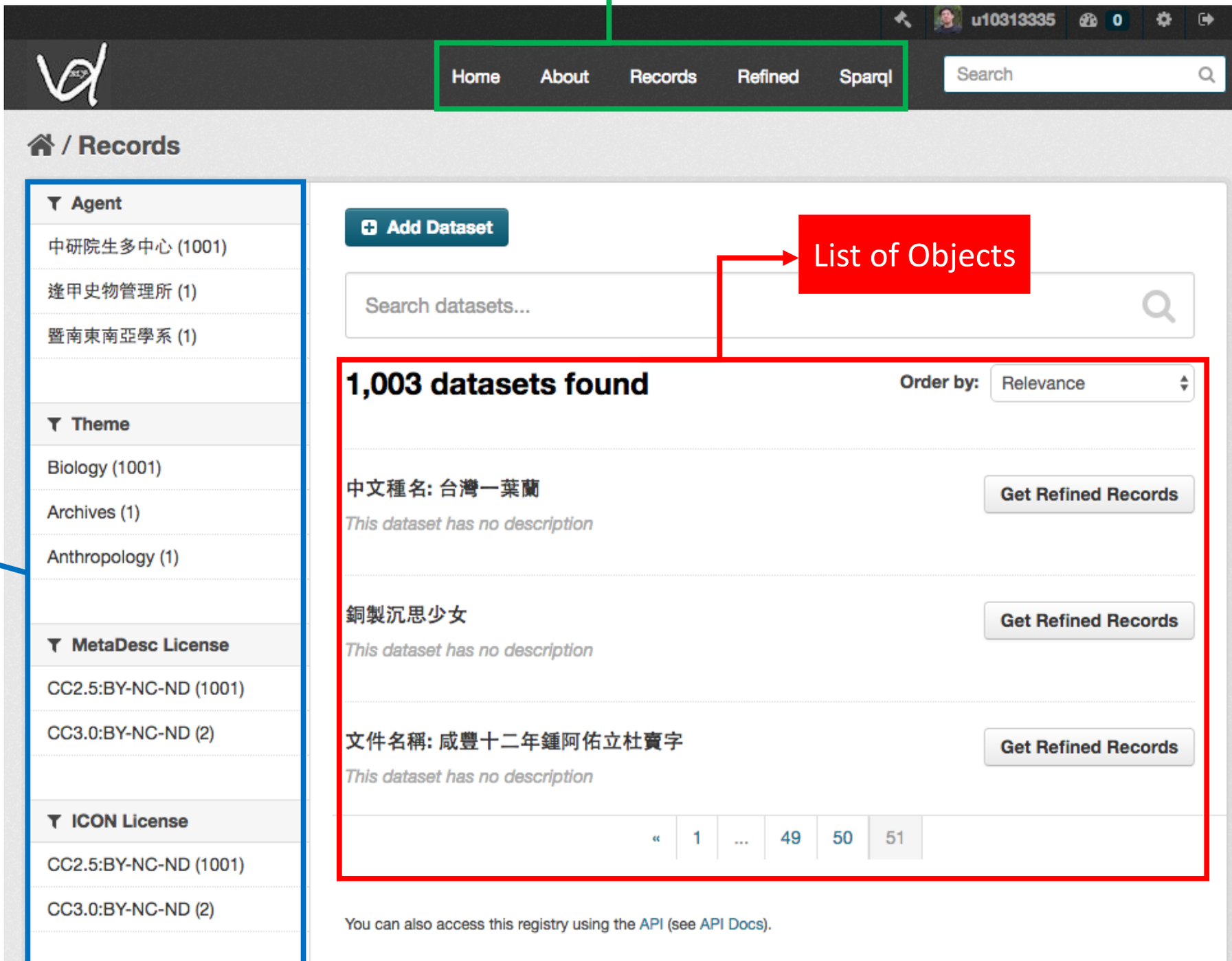
Popular Open Source Data Portal
Customizable UI
Import/Export Linked Data

System Diagram and Operation Overview



Function (1): Linked Data Browsing

Main Menu
Records: Original Data (D Version)
Refined: Semantics Enriched (R Version)



Filters

Home About Records Refined Sparql Search

/ 中文種名: 台灣一葉蘭

中文種名: 台灣一葉蘭

Dataset Groups Activity Stream Manage

Title and Image

中文種名: 台灣一葉蘭

Get Refined Records

Quick switch between D and R versions of the same record

Follow

Organization

Union Catalog and Knowledge Engineering for Digital Archives Project

There is no description for this organization

Social

Google+

Twitter

Facebook

Other Access

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

JSON Turtle

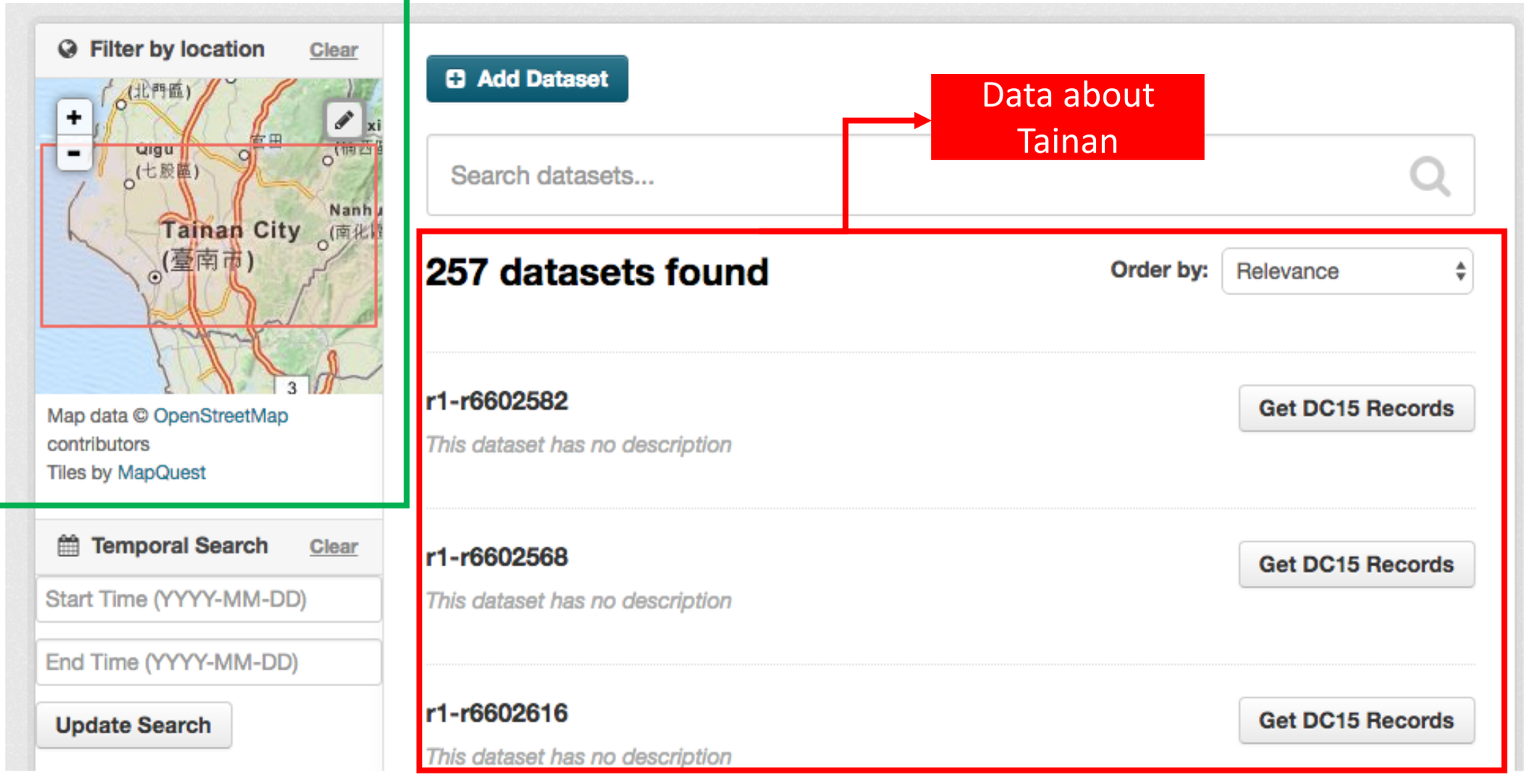
via the CKAN API

METADATA

rdf:type	• data:Reused • r4r:RRObject • dcat:Dataset
r4r:locateAt	http://data.odw.tw/record/d2148340
dcat:themeTaxonomy	data:Biology
dc:contributor	• 採集者: 呂文賓 • 採集者（英文）: Wen-Pen Leu
dc:coverage	• 國家: 台灣 • 最低海拔: 1650 • 行政區: 宜蘭縣大同鄉
dc:creator	鑑訂者: 吉占和
dc:date	採集日期: 1993-04-25

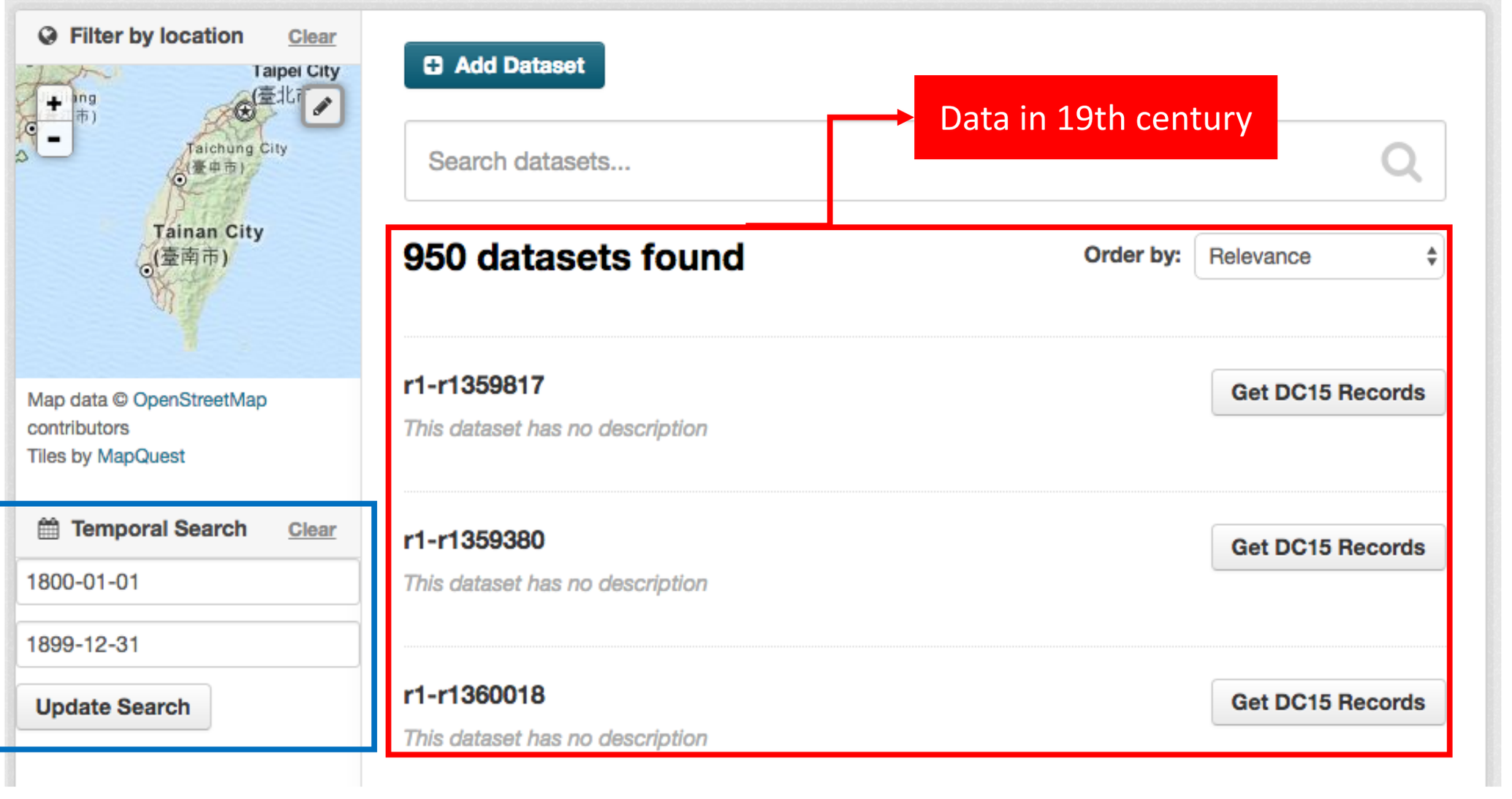

Content of Linked data
Social
Google+
Twitter
Facebook
Other Access
The information on this page (the dataset metadata) is also available in these formats:
JSON
Turtle
via the CKAN API
JSON and Turtle formats provided
METADATA
rdf:type
data:Reused
r4r:RRObject
dcat:Dataset
r4r:locateAt
http://data.odw.tw/record/d2148340
dcat:themeTaxonomy
data:Biology
dc:contributor
採集者: 呂文賓
採集者（英文）: Wen-Pen Leu
dc:coverage
國家: 台灣
最低海拔: 1650
行政區: 宜蘭縣大同鄉
dc:creator
鑑訂者: 吉占和
dc:date
採集日期: 1993-04-25
dc:identifier
編目號: Wen-Pen Leu 2018
標本館號: 43501
dc:language
中文
dc:publisher
中央研究院生物多樣性研究中心
dc:rights
中央研究院 生物多樣性研究中心 植物標本館 Herbarium, Research Center for Biodiversity, Academia Sinica, Taipei (HAST)
dc:source
台灣本土植物資料庫（http://taiwanflora.sinica.edu.tw/）
dc:subject
目（英文）: Asparagales
門（英文）: Spermatophyta
門: 種子植物門
目: 天門冬目
綱（英文）: Monocotyledons
綱: 單子葉植物綱
屬: 一葉蘭屬
科（英文）: ORCHIDACEAE
界: 植物界
屬（英文）: Pleione
科: 蘭科
界（英文）: Plantae
dc:title
中文種名: 台灣一葉蘭

Function (2): Spatial Query



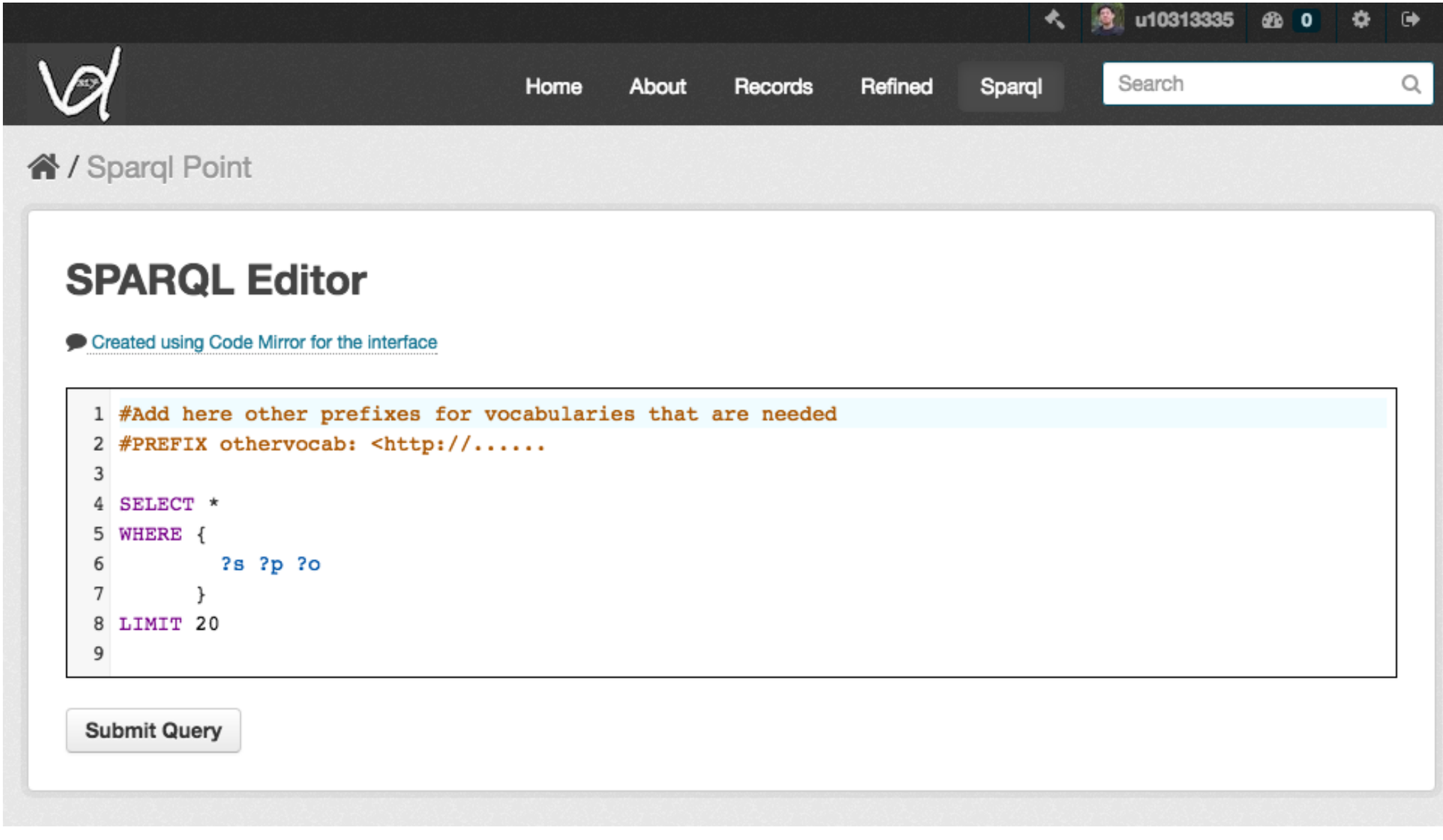
- Spatial indexing based on geo:lat and geo:long values.

Function (3): Temporal Query



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

Function (4): SPARQL Endpoint



Spatial Representation (in development)

Home About Records Refined Sparql Search

/ 中文種名: 台灣一葉蘭

中文種名: 台灣一葉蘭

Followers
0

Follow

Organization

Union Catalog and Knowledge Engineering for Digital Archives Project

There is no description for this organization

Social

Google+

Twitter

Facebook

Other Access

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

JSON Turtle

via the CKAN API

Dataset Groups Activity Stream Manage

中文種名: 台灣一葉蘭

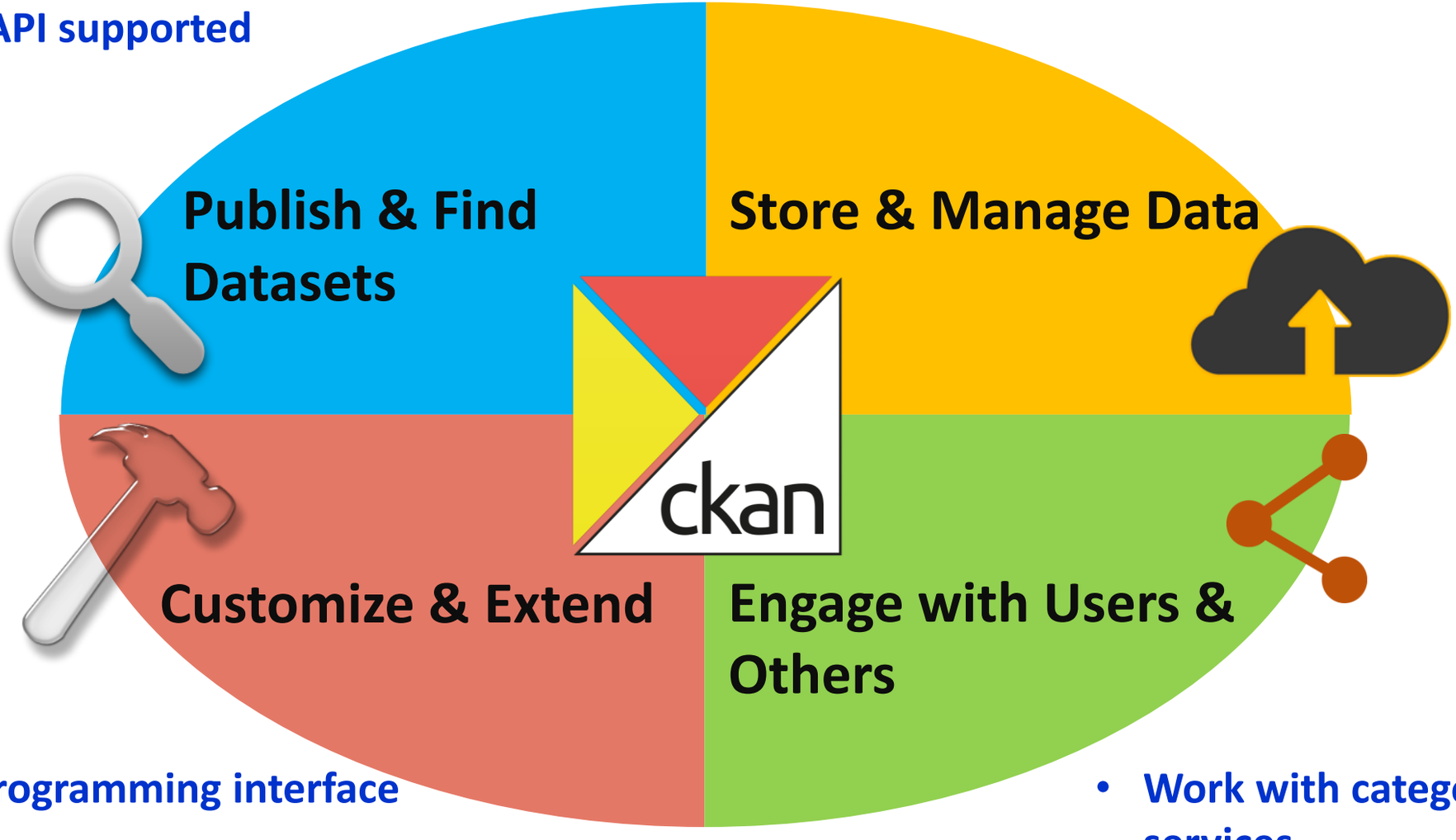
Get Refined Records

METADATA

rdf:type	• data:Reused • r4r:RRObject • dcat:Dataset
r4r:locateAt	http://data.odw.tw/record/d2148340
dcat:themeTaxonomy	data:Biology
dc:contributor	• 採集者: 呂文賓 • 採集者（英文）: Wen-Pen Leu
dc:coverage	• 國家: 台灣 • 最低海拔: 1650 • 行政區: 宜蘭縣大同鄉
dc:creator	鑑訂者: 吉占和
dc:date	採集日期: 1993-04-25

- Friendly search interface
- Rich data format previewer
- Data API supported

- Complete data workflow
- Dataset access control



- Rich programming interface

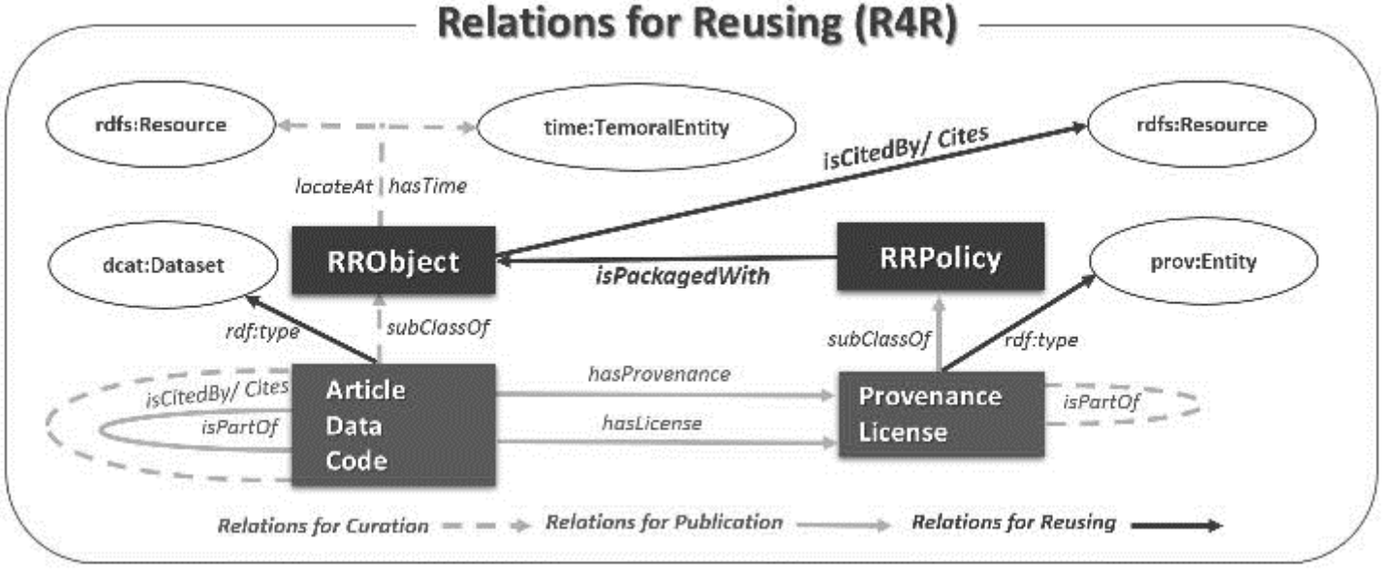
- Work with category services
- Data harvesting
- Social sharing

3. Semantics and Linkage :

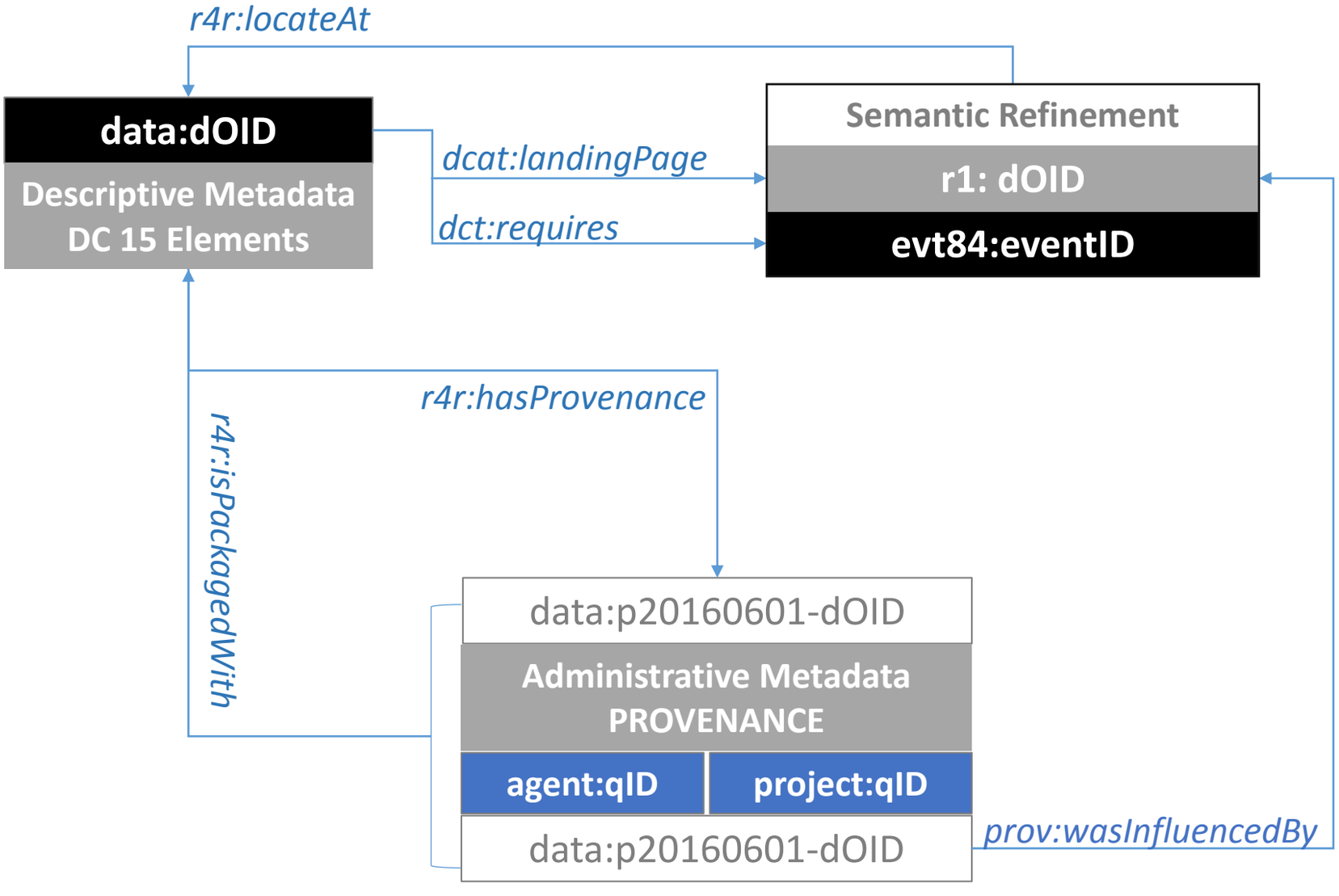
An Ontology for Open Data Web (voc4odw)

http://voc.odw.tw/

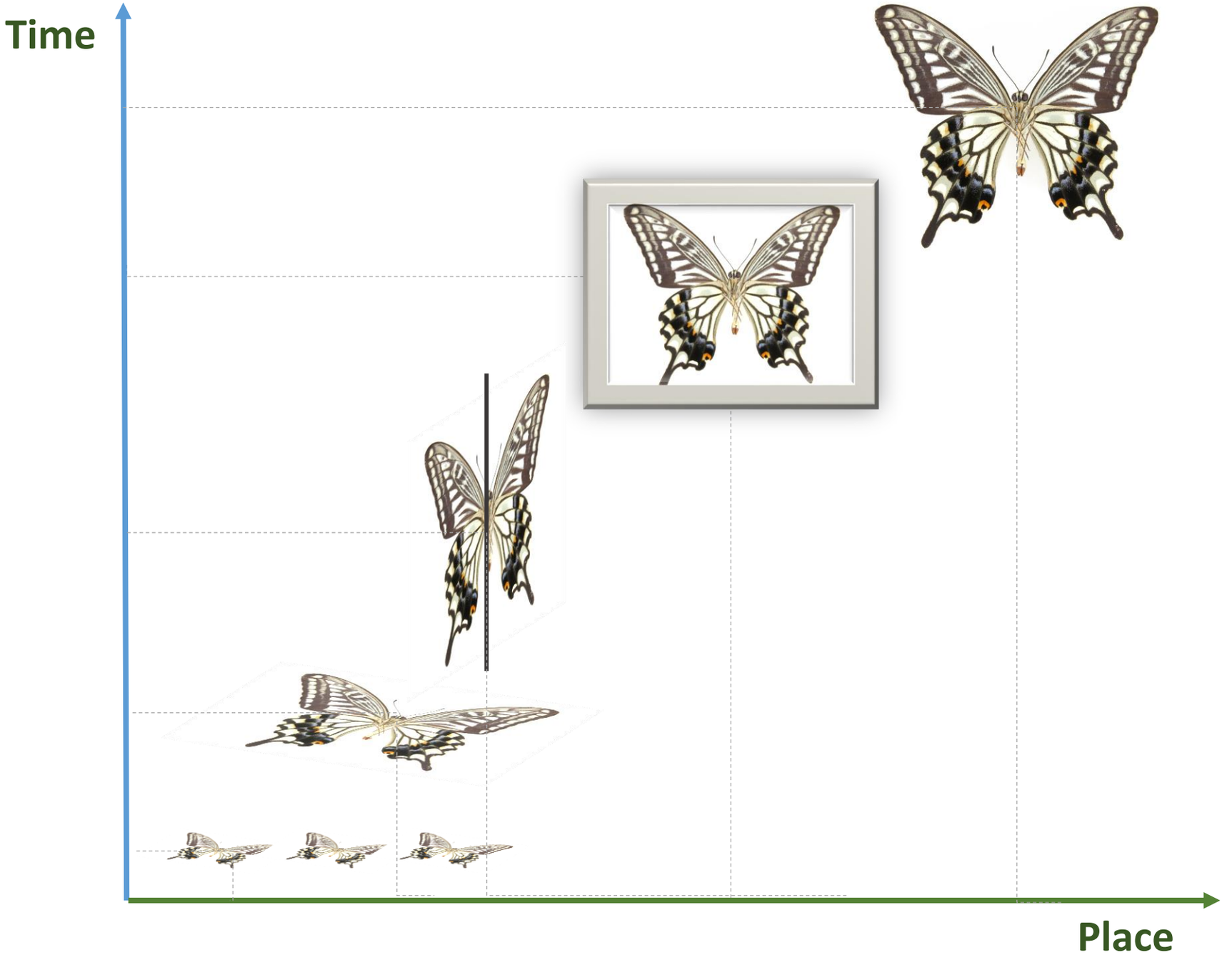
Start from the R4R ontology



Reusing 840,000 cc licensed objects



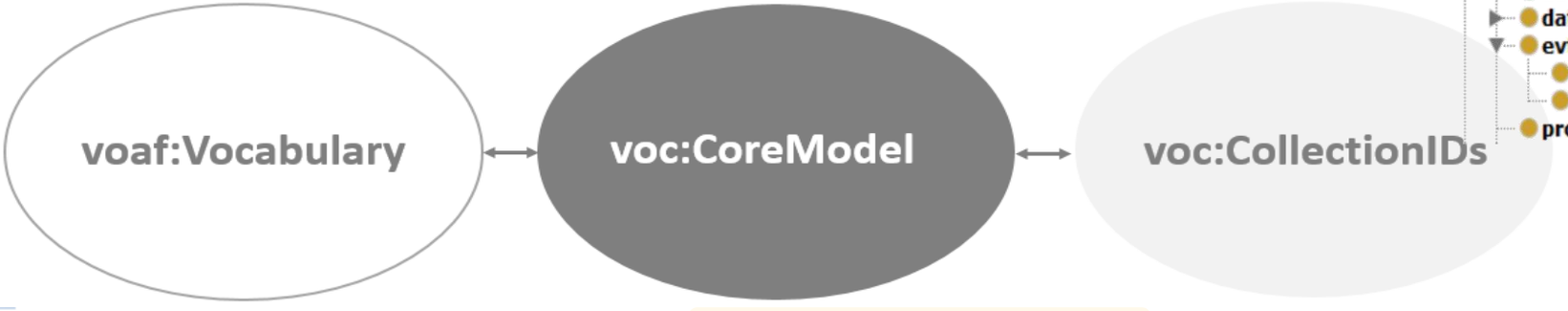
○ D version as the main RRObject (Reusing Related Object) ○ Each Object is packaged with Provenance ○ R version, a semantic enrichment is based on D version



Which object does the metadata refer to? When? Where?

- The used common vocabularies and their relations are independently grouped from the Core Model. This mechanism can simplify the structure of the Core.
- Without changing the Core Model, this mechanism assists the convenience of flexibility in adding or replacing vocabularies.

- To assist users to understand object ID structure when they reuse resources in data.odw.tw.

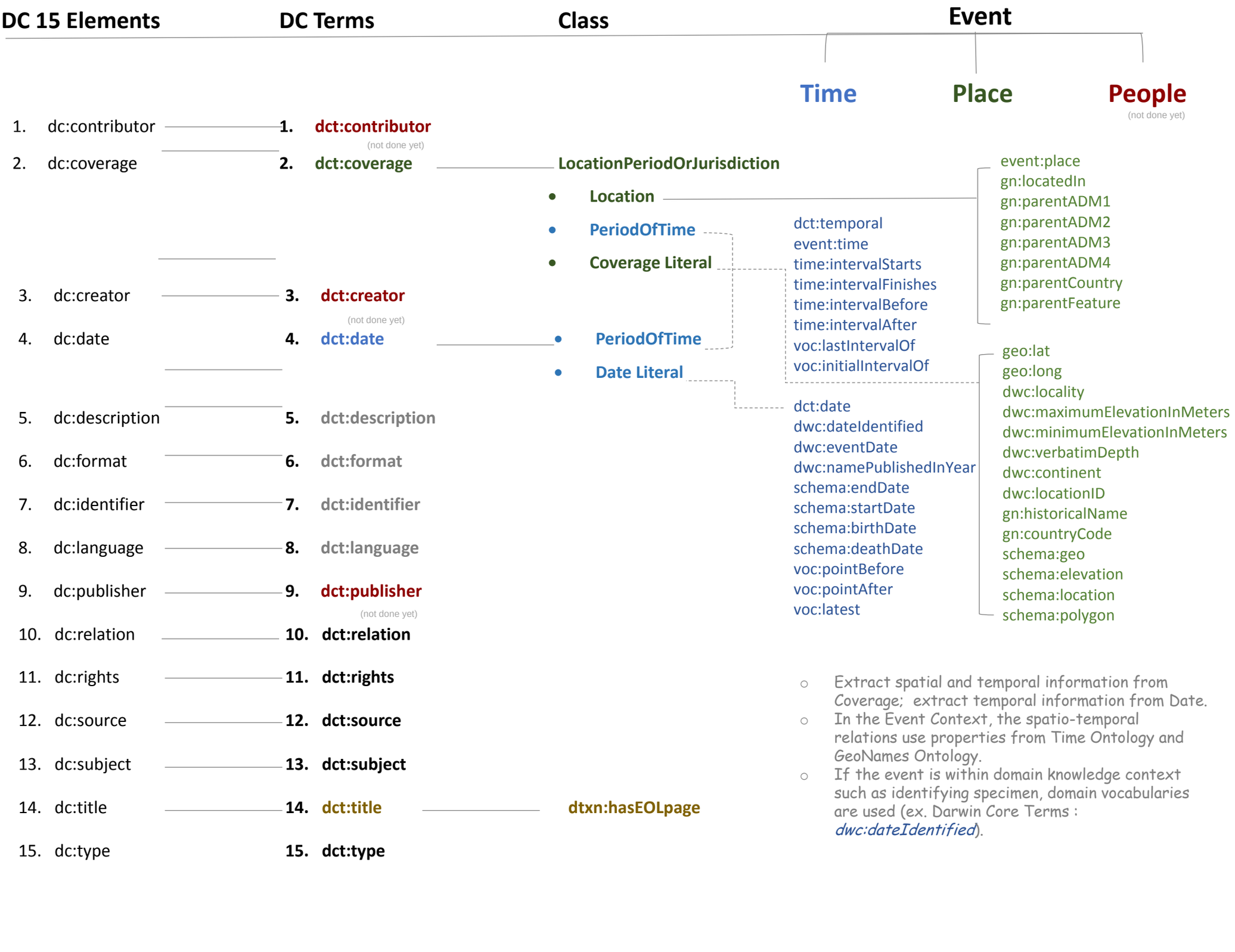


- voc:CollectionIDs
  - agent:qID
  - catdat:DAURL
  - data:dOID
  - data:mappedID
    - eol:ID
    - gns:ID
    - wde:qID
  - data:pOID
  - data:themeID
  - evt84:dOID
    - evt84:event-dOID
    - evt84:eventType-dOID
  - project:qID

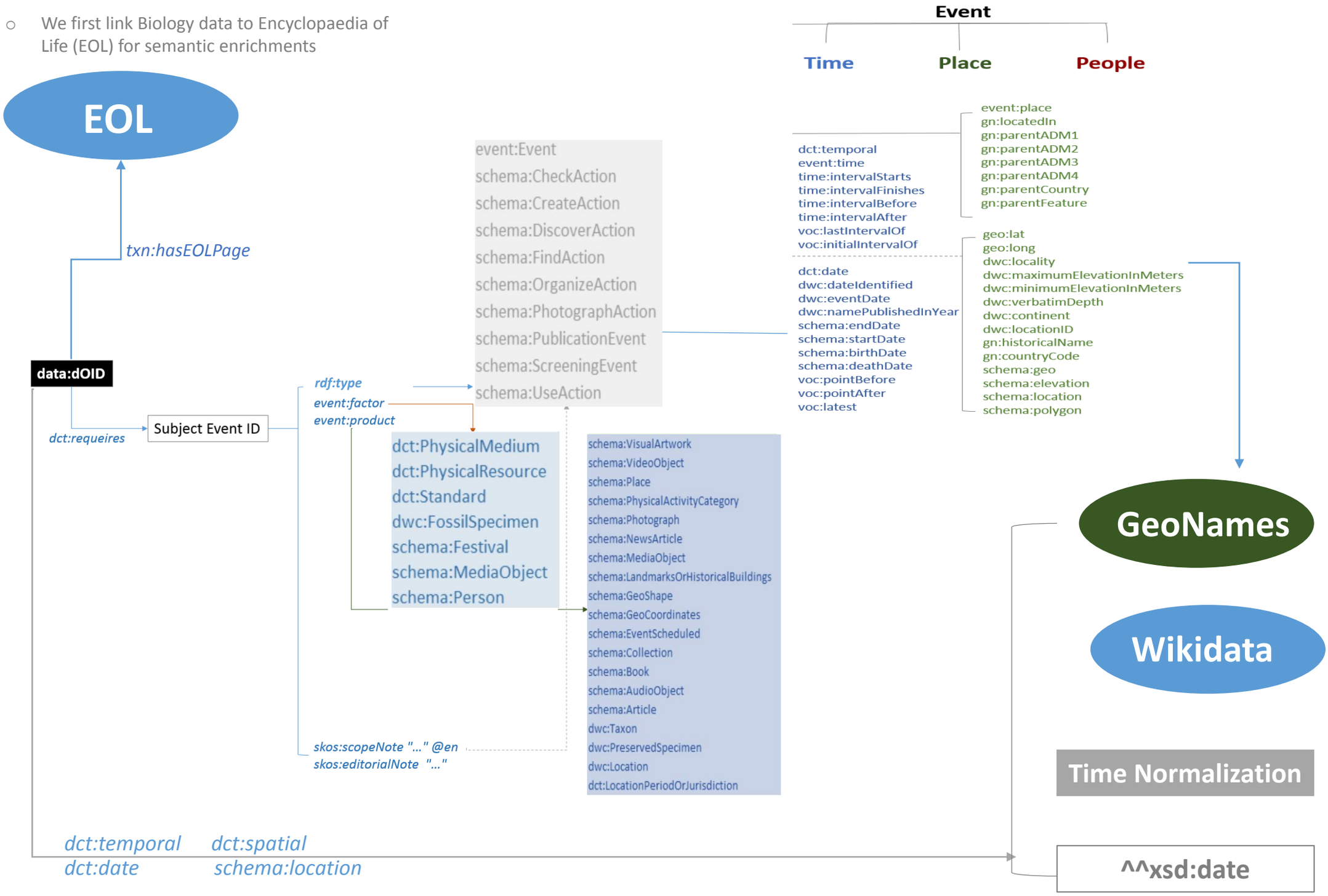
- voaf:Vocabulary
  - aat:AAThesaurus
    - aat:ActivitiesFacet
      - aat:Disciplines
        - aat:Anthropology
        - aat:Architecture
      - aat:Events
        - aat:ReligiousCeremonies
    - aat:AgentsFacet
      - aat:Tribes
    - aat:ObjectsFacet
      - aat:Archives
      - aat:CulturalArtifacts
      - aat:RareBooks
      - aat:RockCarvings
  - dcat:Namespace
  - dct:Terms
  - dwc:Terms
  - event:Ontology
  - foaf:Vocabulary
  - geo:SpatialThing
  - gn:Feature
  - org:Organization ≡ data:Agent ≡ foaf:Organization
  - prov:Namespaces
  - r4r:Ontology
  - rdfs:Resource
  - schema:Thing
  - skos:Namespace
  - time:TemporalEntity
  - txn:TaxonConcept

- voc:ConceptualModel
  - voc:Context ≡ skos:ConceptScheme
    - voc:CommonKnowledge
    - voc:DomainKnowledge
  - voc:Event ≡ event:Event ≡ skos:Concept
    - voc:People
      - data:Agent ≡ foaf:Organization ≡ org:Organization
      - data:Person
      - data:Project
    - voc:Place
    - voc:Time
  - voc:Object ≡ skos:Collection
- voc:DataModel
  - data:Provenance
  - voc:DerivationData
    - data:Refined
      - voc:KnownEvent
      - voc:UnKnownEvent
    - data:Reused ≡ r4r:RRObject
  - voc:PrimaryData
    - voc:CatalogRecord
    - voc:PrimarySource
- voc:OpenData

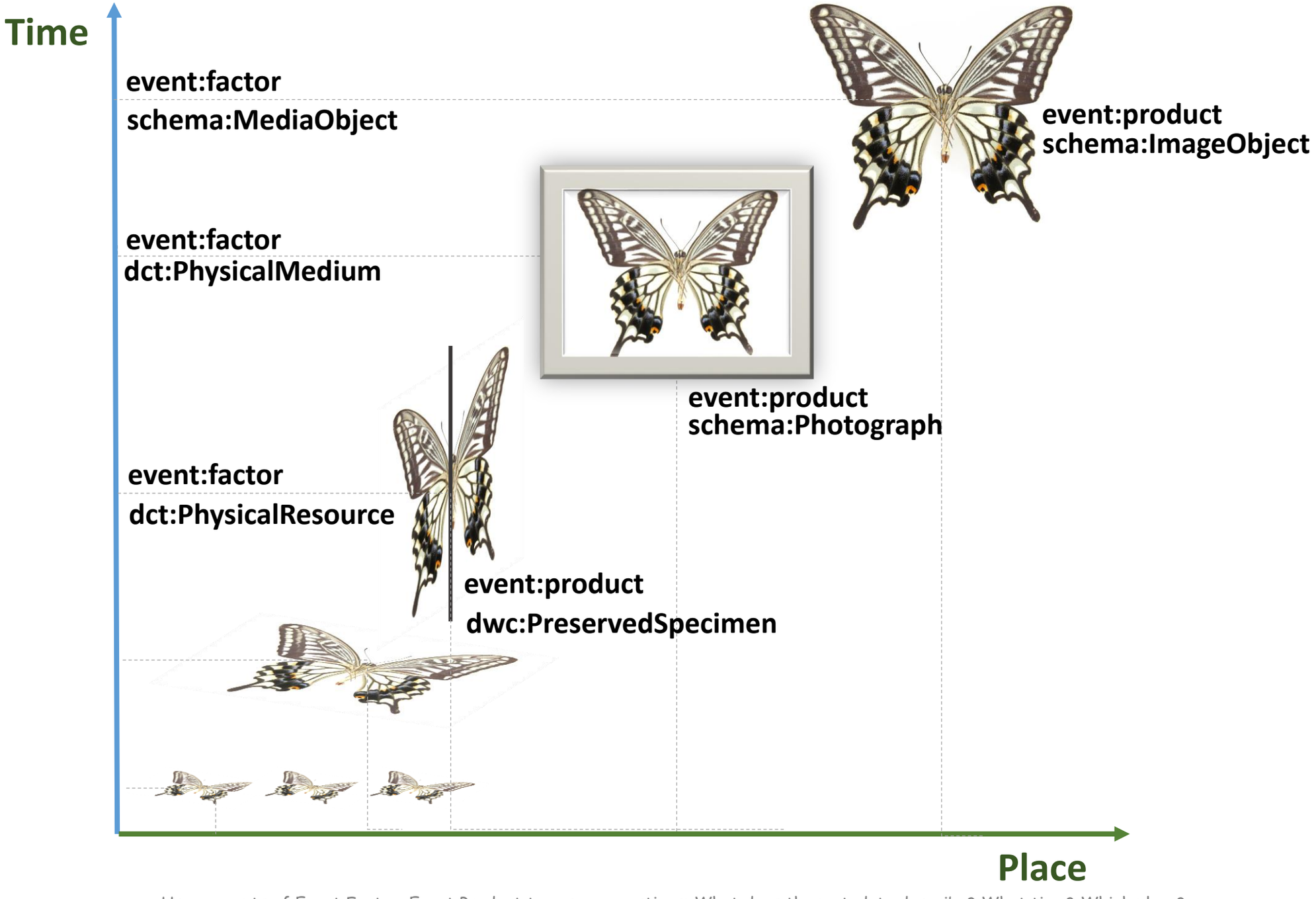
- Data Model depends on Conceptual Model to decide event concepts of People, Time and Place。
- Event Concept depends on Context which come from common or domain knowledge.
- Ex. A general photo event is described by Schema.org terms, and if this photography active is associated with domain knowledge, they will be further described through Art and Architecture Thesaurus (AAT) or Darwin Core Terms。


DC 15 Elements
DC Terms
Class
Event
Time
Place
People
(not done yet)
1. dc:contributor
1. dct:contributor
(not done yet)
2. dc:coverage
2. dct:coverage
LocationPeriodOrJurisdiction
Location
PeriodOfTime
Coverage Literal
event:place
gn:locatedIn
gn:parentADM1
gn:parentADM2
gn:parentADM3
gn:parentADM4
gn:parentCountry
gn:parentFeature
dct:temporal
event:time
time:intervalStarts
time:intervalFinishes
time:intervalBefore
time:intervalAfter
voc:lastIntervalOf
voc:initialIntervalOf
3. dc:creator
3. dct:creator
(not done yet)
4. dc:date
4. dct:date
PeriodOfTime
Date Literal
geo:lat
geo:long
dwc:locality
dwc:maximumElevationInMeters
dwc:minimumElevationInMeters
dwc:verbatimDepth
dwc:continent
dwc:locationID
gn:historicalName
gn:countryCode
schema:geo
schema:elevation
schema:location
schema:polygon
dct:date
dwc:dateIdentified
dwc:eventDate
dwc:namePublishedInYear
schema:endDate
schema:startDate
schema:birthDate
schema:deathDate
voc:pointBefore
voc:pointAfter
voc:latest
5. dc:description
5. dct:description
6. dc:format
6. dct:format
7. dc:identifier
7. dct:identifier
8. dc:language
8. dct:language
9. dc:publisher
9. dct:publisher
(not done yet)
10. dc:relation
10. dct:relation
11. dc:rights
11. dct:rights
12. dc:source
12. dct:source
13. dc:subject
13. dct:subject
14. dc:title
14. dct:title
dtxn:hasEOLpage
15. dc:type
15. dct:type
Extract spatial and temporal information from Coverage; extract temporal information from Date.
In the Event Context, the spatio-temporal relations use properties from Time Ontology and GeoNames Ontology.
If the event is within domain knowledge context such as identifying specimen, domain vocabularies are used (ex. Darwin Core Terms : dwc:dateIdentified).

- We first link Biology data to Encyclopaedia of Life (EOL) for semantic enrichments

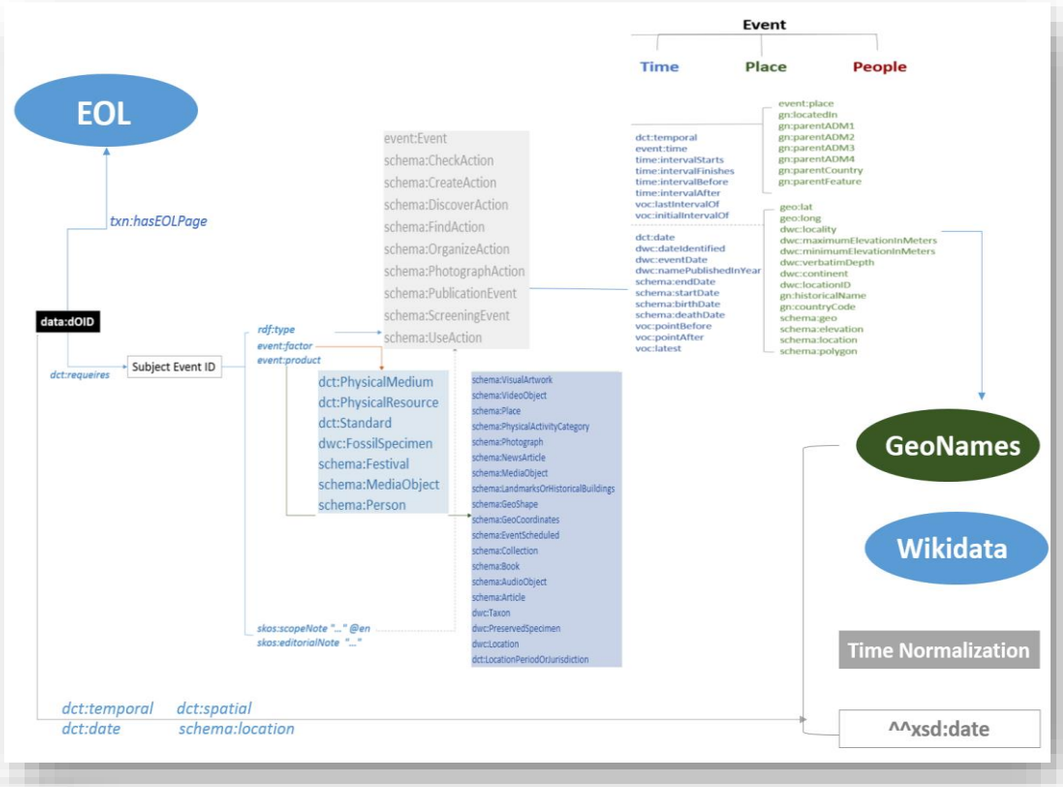


- Concepts for Event Type, Event Factor, Event Product are Class Individuals. This design assists users to increase, delete, or replace knowledge concepts based on their needs without affecting the structure of the Core Model.

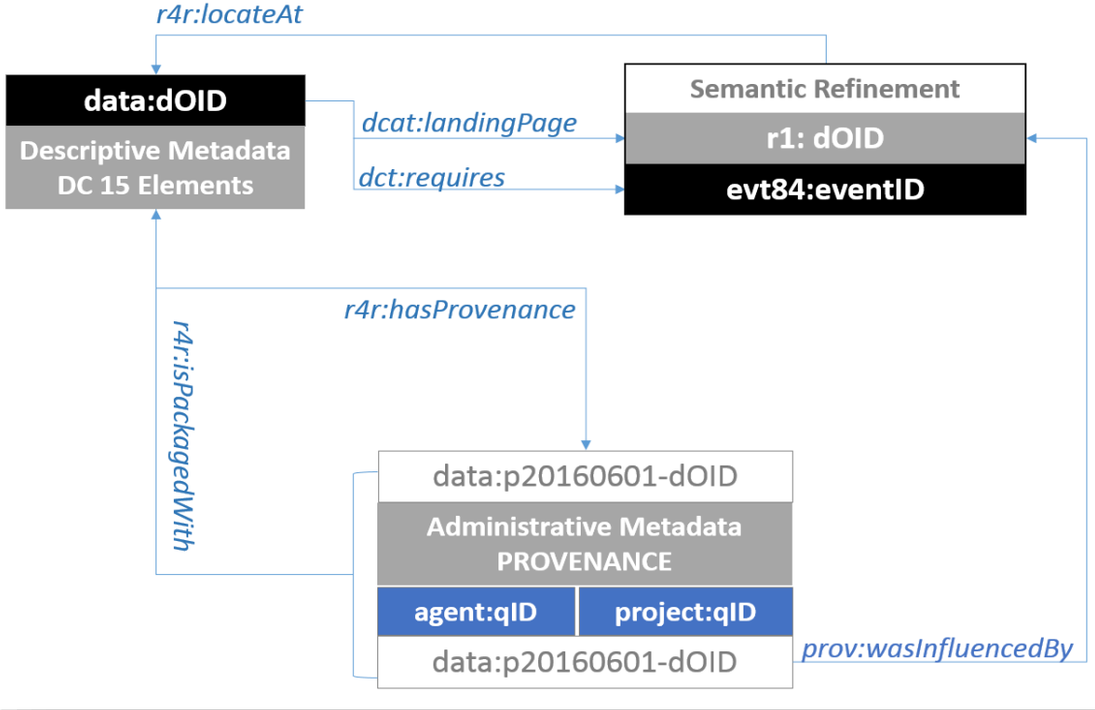


- Use concepts of Event Factor, Event Product to answer questions: What does the metadata describe? What time? Which place?
- Clarify the subject of the event in different digitalization processes: A physical resource, a physical medium, or a media object?

voc:CoreModel



Semantics and Linkage of Archive(d) Catalogs



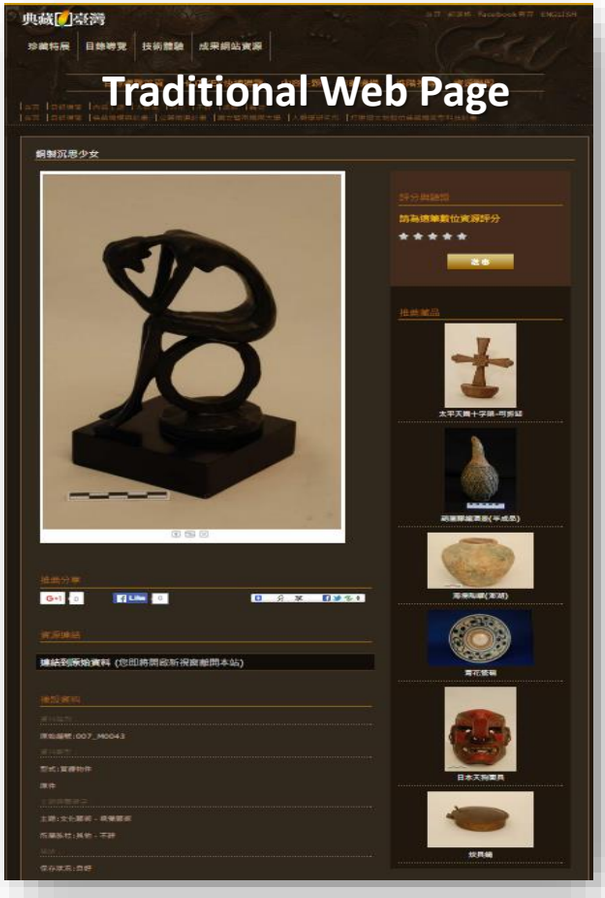
Provenance of Archive(d) Catalogs

Girl Lost in Thought :When is my birthday?

source: http://image.digitalarchives.tw/ImageCache/00/46/73/00.jpg

People understand

“Time period in Taiwan under Japanese rule”

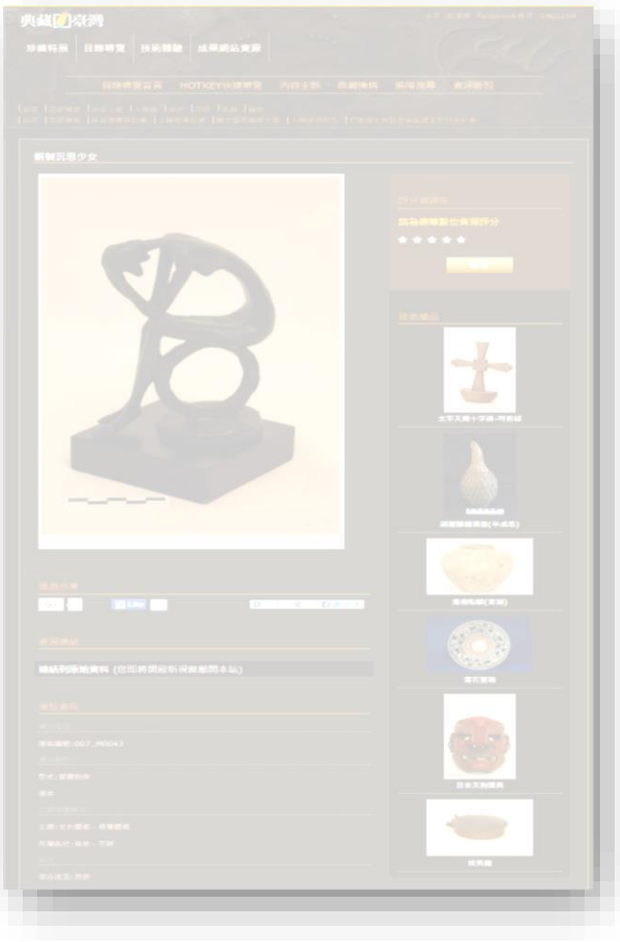


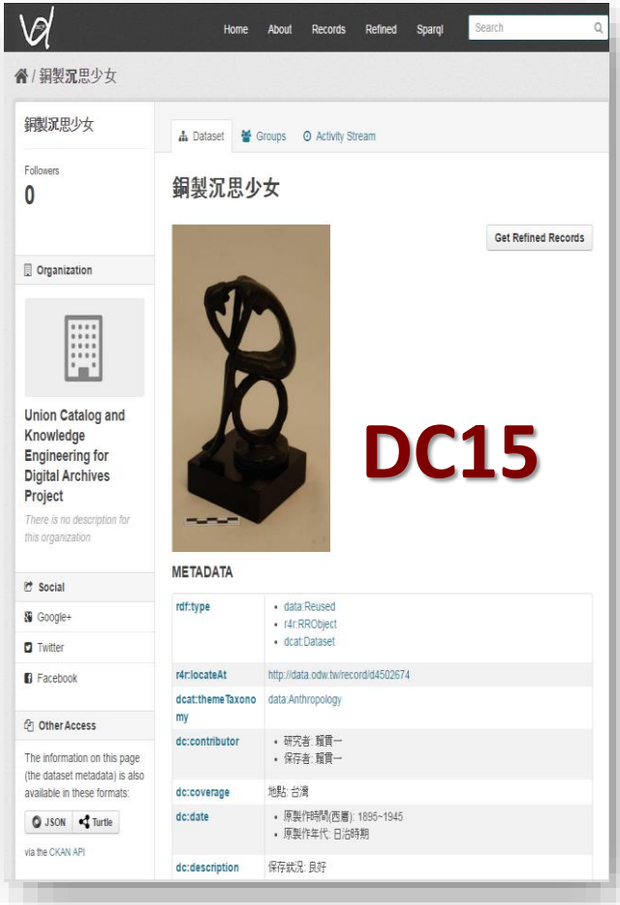
Pure text from metadata description

But machines cannot

People understand

“Time period in Taiwan under Japanese rule”





DC15

D version : Provenance

Machines understand a little

- “Time period in Taiwan under Japanese rule” is about Time
- The Girl’s Genealogy.

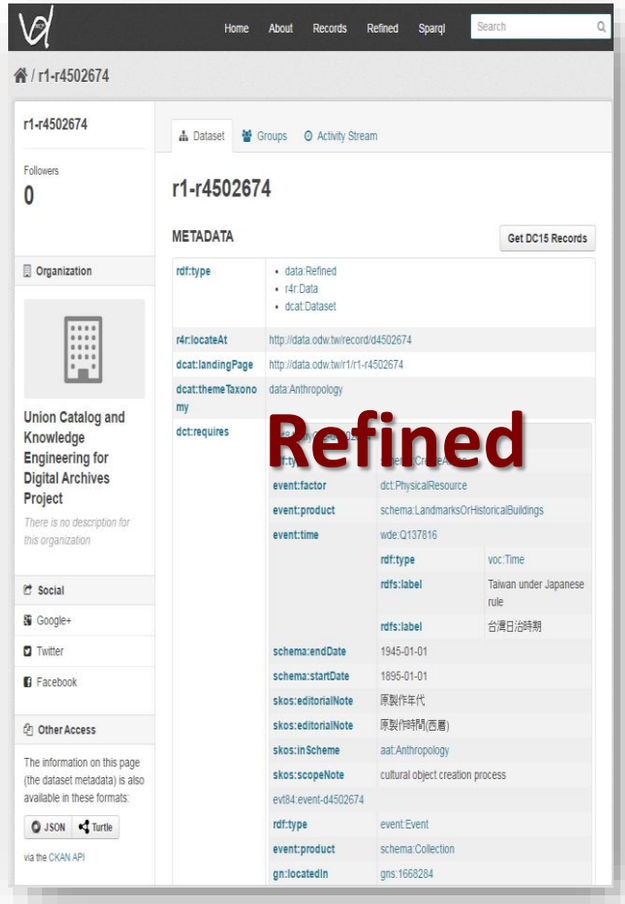
prov:wasStartedBy

British, German, French, Japanese, Spanish …can understand

wde:Q137816

More People understand

“Time period in Taiwan under Japanese rule”



Machines can make inferences too…

- New representations of time intervals, and their temporal relations.
- Allow multiple interpretations to co-exist for one resource.
- When external resources are updated and enriched, semantics of this resource is updated and enriched automatically as well.

data.odw.tw & **voc.odw.tw**

We welcome your valuable comments & suggestions!