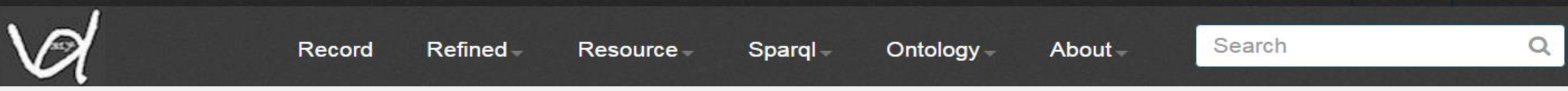
Metadata as Linked Data for Research Data Repositories

Cheng-Jen Lee, Andrea Wei-Ching Huang, Tyng-Ruey Chuang
Institute of Information Science, Academia Sinica, Taiwan

http://data.odw.tw



INTRODUCTION

Current research repositories can not meet the needs of innovative solutions providing feature-rich services for helping data publishing such as visualization, validation & reuse in different applications.

(Assante, Candela, Castelli & Tani, 2016)
Are scientific data repositories coping with research data publishing?. Data Science Journal, 15, p.6. DOI: http://doi.org/10.5334/dsj-2016-006

The metadata of research data increases the access to and reuse of the data.

(Willis, Greenberg and White, 2012)
Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology 63(8): 1505–1520, DOI: http://dx.doi.org/10.1002/asi.22683

*From 2014 to 2018, Stanford, Harvard, and Cornell are collaboratively working on **linked data**. The goal is to gather contextual information about research resources such as books, articles, serials, datasets, and multimedia into a semantic-web-based **Scholarly Resource Semantic Information Store (SRSIS)**.*

https://www.ld4l.org/ and https://wiki.duraspace.org/display/ld4l/Project+Rationale

Before

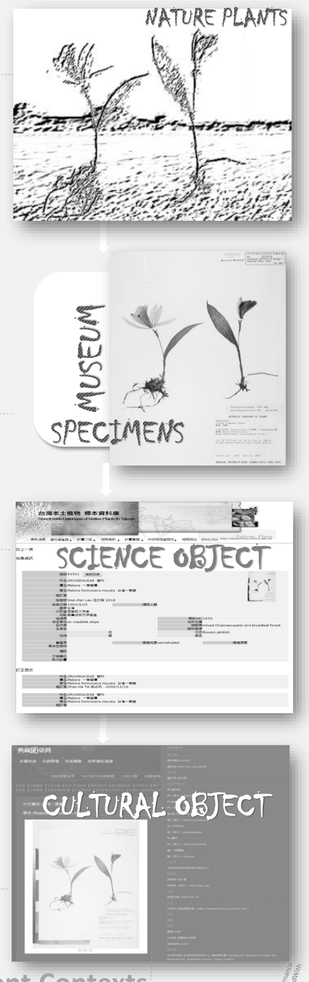
THE STORY OF "THE PLANT"

Before 1993-04-25, a natural plant may be recognized as *Pleione Formosana,* "the plant". "The plant" was in somewhere around Datong , Yilan, Taiwan.

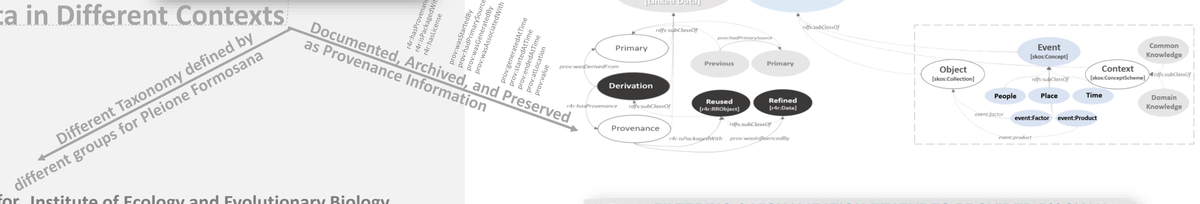
A researcher, *Wen-Pen Leu* , took a specimen collection activity on *1993-04-25*, and then made "the plant" as a specimen for science. "The plant" has been curated in the *Herbarium, Research Center For Biodiversity, Academia Sinica, Taipei (HAST).*

Since 2003, the HAST, have completed the "Database of Native Plants in Taiwan", and digitalized "the plant" with its image of herbarium specimen and metadata information. "The plant " becomes a *science object* served for natural scientists via internet and web. The HAST uses its own data schema to store "the plant" fitting their database and domain knowledge requirements.

Around 2011-05-13, the HAST collaboratively worked with *the Union Catalog of Digital Archives Taiwan* (CATDAT). "The plant " was then converted its own data schema and imported the metadata as a catalogue record and a cultural object to CATDAT based on Dublin Core 15 Elements.



Data in Different Contexts



Herbarium, Research Center for Biodiversity, Academia Sinica — **Institute of Ecology and Evolutionary Biology, College of Life Science, National Taiwan University**

界（英文）: Plantae	○	界（英文）:Plantae
界: 植物界	○	界:植物界
門（英文）: Spermatophyta	○	門（英文）:Spermatophyta
門: 種子植物門	○	門:胚胎植物門
綱（英文）: Monocotyledons	○	綱（英文）:Angiospermae
綱: 單子葉植物綱	○	綱:被子植物綱
目（英文）: Asparagales	○	目（英文）:Orchidales
目: 天門冬目	○	目:蘭目
科（英文）: Orchidaceae	○	科（英文）:Orchidaceae
科: 蘭科	○	科:蘭科
屬（英文）: Pleione	○	屬（英文）:Pleione
屬: 一葉蘭屬	○	屬:一葉蘭屬
	○	種小名:bulbocodioides

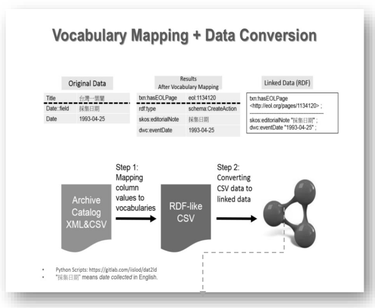
METHODS

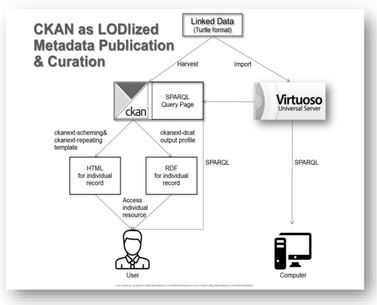
Semantics in LOD Context

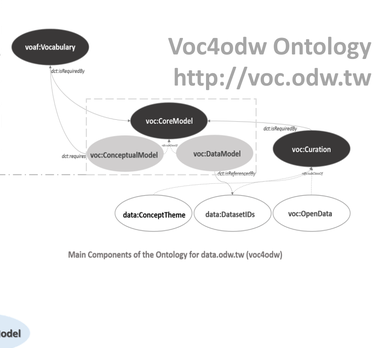
A Reused Data:
"The Plant" is a derivation data, now as a data:d2148340, (data:Reused) modelled from the science and cultural object of this *Pleione Formosana*. It shares the meaning of the r4r:RRObject in the R4R Ontology: that any resource served as a component for reuse is defined as a Reusing Related Object (RRObject).

A Semantically Refined Data:
A derivation data, r1:r1-r2148340 (data:Refined), is extracted something from the data:Reused as to enrich semantics of the resource. Semantic meanings of different interpretations are provided from the conceptual model. At the same time, different interpretations or derivation processes are curated in different refined versions (ex. r1, r2, r3...).

A dcat:Dataset:
The data:d2148340 at data.odw.tw is a collection of data (which includes different versions of refined data), published or curated by a single agent, and is available for access or download in one or more formats.

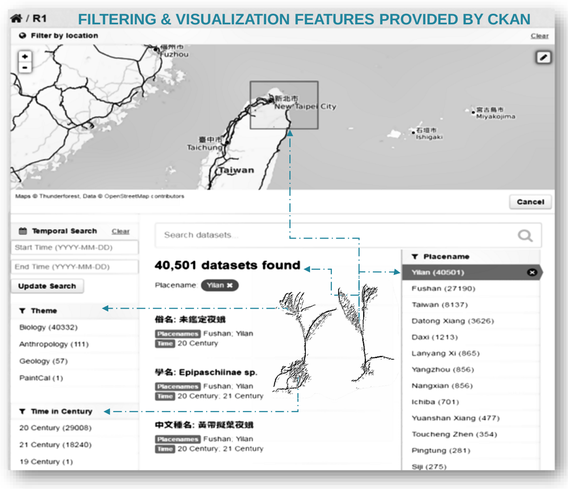






Main Components of the Ontology for data.odw.tw (voc4odw)

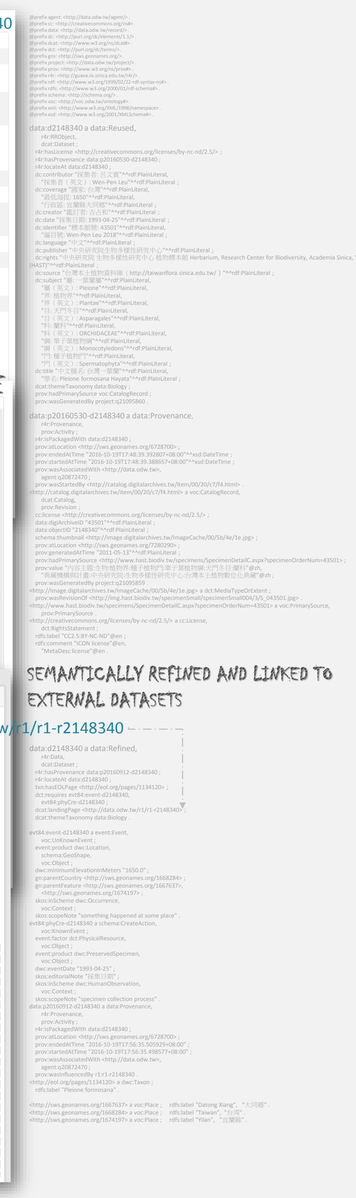
FILTERING & VISUALIZATION FEATURES PROVIDED BY CKAN



After

FOR HUMAN

FOR MACHINE

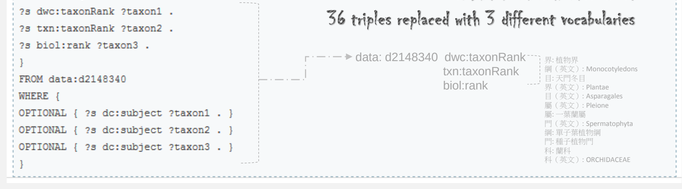



```
# Multiple Semantic Representations and Transforming Between Vocabularies
# Example: Convert DC Elements to Darwin Core Terms,
TaxonConcept Ontology or Biological Taxonomy Vocabulary .

PREFIX data: <http://data.odw.tw/record/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX biol: <http://purl.org/NET/biol/ns#>
PREFIX txn: <http://lod.taxonconcept.org/ontology/txn.owl#>

CONSTRUCT {
?s dwc:taxonRank ?taxon1 .
?s txn:taxonRank ?taxon2 .
?s biol:rank ?taxon3 .
}
FROM data:d2148340
WHERE {
OPTIONAL { ?s dc:subject ?taxon1 . }
OPTIONAL { ?s dc:subject ?taxon2 . }
OPTIONAL { ?s dc:subject ?taxon3 . }
}
```

SPARQL QUERY RESULTS:
36 triples replaced with 3 different vocabularies



RESULTS

1. An Use Case for Curation, Publication & Reuse of Metadata as Linked Data.

- 843,309 CC licensed metadata records of 14 domains reused from the Union Catalog of Digital Archives Taiwan.
- 44,806,400 triples (Linked Data) encoded with Dublin Core 15 Elements and Provenance Information.
- 25,913,304 triples from 832,803 records semantically refined with spatial & temporal normalization, mapping, and linking with domain knowledges (external vocabularies, ontologies. and knowledges bases).
- 14 domains include Archaeology, Architecture, Archives, Artifacts, Biology, Geology, Manuscript, Multimedia, NewsMedia, PaintCal , RareBook, ResearchReuse, StoneRub.
- 80 projects and 74 agents associated with metadata records are curated by their linked data formats and Wikidata ID: they have roles in NGO (2), Museum (5), Library (2), Government (9), Archive (1) and Academia (55).

2. A New Method to Manage Data for General-use & Discipline-specific Repositories.

- For Open Science: using the CKAN (Comprehensive Knowledge Archive Network) as a major solution that makes linked metadata available, citable, and validated.
- Availability: data shared with multiple formats, CSV, XML, Turtle, RDF/XML, JSON-LD, consumed both by human & machine.
- Validation and Reproducibility: each data encoded with provenance in details while at the same time a complete mechanism for publishing article, data and code is designed and implemented.
- A flexible and adaptable ontology for describing different data context (common knowledge or domain knowledge), event concepts (people, place, time) and objects collected by meaningful groups of different vocabularies is provided .
- Data Visualization is enhanced and integrated through spatial and temporal mapping, filtering and linking system design.

3. Data Semantically Enriched with Vocabularies and Knowledge Bases via Adaptable Mechanisms.

- 18 international vocabularies used for modeling common knowledge, and 5 domain specific vocabularies for place, time, art and humanity, or biology are applied. 3 knowledge bases like GeoNames, Wikidata, and Encyclopedia of Life are mapped and linked.
- The use of SPARQL language and endpoints provide data analytic semantic queries both in local and external. In addition, data from the RDF triplestore can be easily used in 3rd-party applications.
- Multiple DataClean Versions Mechanism: we treat data cleaning as a kind of interpretation. Refined Versions (R Versions i.e. r1, r2, r3…) provide different contexts to different needs of users.
- Multiple LinkedKnowledge Bases Mechanism: more knowledge bases like DBpedia, WordCat, or LinkedGeoData can be linked in future via different R Versions without sacrificing the integrality of the original Version, encoded with DC 15 .
- Multiple SemanticStructure Versions Mechanism: different interpretations results from the use of different vocabularies . Co-exists of multiple R Versions with different vocabularies or transforming vocabularies via SPARQL are solutions.