Social Web Meets Sensor Web: From User-Generated Content to Linked Crowdsourced Observation Data

Dong-Po Deng\textsuperscript{1,3}, Guan-Shuo Mai\textsuperscript{2}, Tyng-Ruey Chuang\textsuperscript{1}, Rob Lemmens\textsuperscript{3}, Kwang-Tsao Shao\textsuperscript{2}

\textsuperscript{1}Institute of Information Science and \textsuperscript{2}Biodiversity Research Center, Academia Sinica, Taiwan
\textsuperscript{3}Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Netherlands

Linked Data in the Web 2014 @ WWW’14, April 8, 2014
Background & Motivation
Citizen Science

- Citizen science refers to a distributed, collaborative problem-solving model in which a crowd of undefined size is engaged to solve a complex or scientific problem through an open call.
  - Projects can happen at any scale—local, national, global
  - Projects generally take place over the long term tracking trends/changes over time
Galaxy Zoo

http://www.galaxyzoo.org/
Audubon

http://www.audubon.org/
iNaturalist

http://www.inaturalist.org/projects/asc-roadkill-observations
Reptile Road Mortality

https://www.facebook.com/groups/roadkilled/
Motivation

- By utilizing social media, citizen science projects can potentially engage many participants to contribute their observations covering a large geographic region and over a long time period.

- This is an improvement, for example, over traditional biodiversity surveys which typically involve relatively few people in confined regions and periods.
Motivation

- Linked Data aims on making interlinked data available on the Web, and to interconnect data with the aim of increasing its value for users.
A post on the Facebook group
Reptile Road Mortality

Observation Provider:
- Chuang Yu Ta

Observation date:
- 2013/12/4

Species Identifier:
- Joyce Chen

Species name:
- 鼬獾 (Melogale moschata)

Identification Date:
- Dec. 4, 2013

Post section

Comment section

Thread

Observation location:
- Geoname: 新店 (Sindian)
- Road kilometre: 台9線16.3K (Province Road No.9, 16.3 kilometer)
- Lat: 24.95149
- Lon: 121.57520
- Lat: 151m

Photo:
- A proof of occurrence
Aims

• to transform the crowdsourced observation data to RDF model via using an ontology of citizens as sensors;
• to interlink the crowdsourced observation data with other Linked Data resources such as biodiversity (TaiCOL) and geospatial information (Geonames);
• to make the crowdsourced observation data
  • accessible to machines by using the Linked Data paradigm and
  • readable for humans by means of a faceted browser.
Information extraction
Named-Entity Recognition

• Problems:
  • Chinese word segmentation
    • Chinese texts are character-based, not word-based
    • There is often no space between characters in written Chinese sentences
  • The lacks of domain-specific corpora such as geography and biodiversity.
  • Shorten names

• We compiled a geo-name lexicon from the Taiwan Geographic Names database and a species-name lexicon from the Taiwan Catalogue of Life databases (TaiCOL)

• The named-entity recognition approach we use was elaborated in a paper we previously published [10]
Ontology Engineering
The ontology
Transformations from crowdsourced data to RDF
Species names

eoe:iden_559070840853748_01  rdf:type dwctype:Identification ,
   owl:NamedIndividual ;
dwc:datIdentified  eoe:iden_time_559070840853748 ;
dsw:identifies  eoe:idv_238918712815615_694835510557264 ;
dsw:isBasedOn  eoe:token_559070840853748 ;
dsw:toTaxonConcept  taxon:380522 .

taxon:380522 rdf:type dwctype:Taxon ,
   owl:NamedIndividual ;
dsw:hasName  taibif:380522 ;
skos:preLabel "Melogale moschata subaurantiaca" ;
skos:altLabel "鼬獾" .

Observation Provider: Chuang Yu Ta
Observation location: Geoname: (Sindian)
Road kilometre: (Province Road No.9, 16.3 kilometer)
Lat: 24.95149
Lon: 121.57520
Lat: 151m
Observation date: 2013/12/4
Photo: A proof of occurrence
Species Identifier: Joyce Chen
Species name: (Melogale moschata)
Identification Date: Dec. 4, 2013

Observation: A proof of occurrence
Species Identifier: Joyce Chen
Species name: (Melogale moschata)
Identification Date: Dec. 4, 2013

Post section
Threading
Photo
Observation: A proof of occurrence
Species Identifier: Joyce Chen
Species name: (Melogale moschata)
Identification Date: Dec. 4, 2013

Observation: A proof of occurrence
Species Identifier: Joyce Chen
Species name: (Melogale moschata)
Identification Date: Dec. 4, 2013

Photo
Observation: A proof of occurrence
Species Identifier: Joyce Chen
Species name: (Melogale moschata)
Identification Date: Dec. 4, 2013
The taxon concept of extract species name is linked to a URI in TaiBIF
The taxon name of extract species name is linked to a URI in TaiBIF

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>spv:canonicalForm</td>
<td><a href="http://taibif.tw/lod/resource/CanonicalForm/Melogale_moschata_subaurantiaca">http://taibif.tw/lod/resource/CanonicalForm/Melogale_moschata_subaurantiaca</a></td>
</tr>
<tr>
<td>rdfs:label</td>
<td>Melogale moschata subaurantiaca (Swinhoe, 1862)</td>
</tr>
<tr>
<td>is spv:scientificName of</td>
<td><a href="http://taibif.tw/lod/resource/Species/380522">http://taibif.tw/lod/resource/Species/380522</a></td>
</tr>
<tr>
<td>rdf:type</td>
<td>spv:ScientificName</td>
</tr>
</tbody>
</table>

This page shows information obtained from the SPARQL endpoint at http://140.109.28.72:8890/sparql.
As Turtle | As RDF/XML | LOD.TW RDF Vizualizer
Geographic names

eoe:placeOfOb_559070840853748 rdf:type eoe:PlaceOfObservation, owl:NamedIndividual;
geo:hasGeometry eoe:point_559070840853748; gn:name "新店";
owl:sameAs http://lod.tw/placenames/159624.

eoe:point_559070840853748 rdf:type geo:Point, owl:NamedIndividual;
w3c_geo:long "121.575200";
w3c_geo:lat "24.951490";
geo:asWKT "Point(121.575200 24.951490)"^^sf:wktLiteral.
The extract place name points to a URI in Taiwan Geographic Name Database

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>tgn:Address</td>
<td></td>
</tr>
<tr>
<td>tgn:County</td>
<td>新北市</td>
</tr>
<tr>
<td>tgn:Kind</td>
<td>地名</td>
</tr>
<tr>
<td>tgn:Name</td>
<td>新店</td>
</tr>
<tr>
<td>tgn:Phone_no</td>
<td></td>
</tr>
<tr>
<td>tgn:Townname</td>
<td>新店區</td>
</tr>
<tr>
<td>tgn:Xcoord</td>
<td>121.53397E0 (xsd:decimal)</td>
</tr>
<tr>
<td>tgn:Ycoord</td>
<td>24.928386E0 (xsd:decimal)</td>
</tr>
<tr>
<td>tgn:en_hanyu</td>
<td>Xindian</td>
</tr>
<tr>
<td>tgn:en_tongyong</td>
<td>Sindian</td>
</tr>
<tr>
<td>tgn:id</td>
<td>159624 (xsd:integer)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>geoname_tw #159624</td>
</tr>
<tr>
<td>owl:sameAs</td>
<td><a href="">gn:1674826</a></td>
</tr>
<tr>
<td>rdf:type</td>
<td>tgn:geoname_tw</td>
</tr>
</tbody>
</table>

The server is configured to display only a limited number of values (limit per property bridge: 50).
Triplestore (BBN Parliament)

http://lod.tw/parliament/
Spatiotemporal SPARQL query

```sparql
SELECT Distinct ?Obs ?POO_geo ?POO_wkt
WHERE{
  ?Obs a ssn:Observation;
  DUL:hasLocation ?POO ;
  ssn:observationResultTime ?Int .
  ?Int time:xsdDateTime ?Time_xsd .
FILTER (geof:sfWithin(?POO_wkt,"POLYGON((
    121.756555 24.488236, 121.207238 24.488236,
    121.207238 25.141394, 121.756555 25.141394,
    121.756555 24.488236))"^^sf:wktLiteral)
Filter (?Time_xsd > "2013-12-19T16:00:00Z"^^xsd:dateTime )
```
The query result of SPARQL query

<table>
<thead>
<tr>
<th>Obs</th>
<th>POO_geo</th>
<th>POO_wkt</th>
</tr>
</thead>
</table>
The faceted browser

Conclusions

- This study explored the issues involved in the use of social media in citizen science projects, as well as reported our experiences in transferring unstructured collaborative information to structured data for scientific purposes.

- We shared our experiences in tackling the data collection from social process to scientific process.

- The successful implementation of this approach can further facilitate the development of social-media based citizen science projects.

- We believe it also has broader applications in user-generated content management, and promises to be a practical solution to an important design problem in citizen science projects on the Web.
Future work

- This study uses several tools for storing and visualizing the RDF triples. To make the browser more usable, a task to integrate the tools into a knowledge-based browser remains to be done in the future.

- Moreover, the triplified dataset should be considered for linkage to larger linked datasets such as DBPedia and other resources.
Thank you for your attentions!

Question?

dongpo@iis.sinica.edu.tw