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

Dong–Po Deng[†]
Institute of
Information Science
Academia Sinica
Taipei, Taiwan

Guan–Shuo Mai
Biodiversity
Research Center
Academia Sinica
Taipei, Taiwan

Tyng–Ruey Chuang[‡]
Institute of
Information Science
Academia Sinica
Taipei, Taiwan

Rob Lemmens
Faculty of
Geo–Information
Science and Earth
Observation (ITC)
University of Twente
Enschede, Netherlands

Kwang–Tsao Shao
Biodiversity
Research Center
Academia Sinica
Taipei, Taiwan

ABSTRACT

The reach of dominating social media like Facebook and Twitter in the current population is enormous, and these media have long been leveraged for diverse applications. In particular, for some citizen science projects, existing social media increasingly become platforms on which participants interact and contribute. These user contributions, often termed User-Generated Content (UGC), can be a mix bag of posts, comments, images, and other media. We report in this paper a work-in-progress in formalizing user contributions from a large Facebook group (more than 4,000 users) established for biodiversity observation. A major part of our work is to extract structured datasets with well-defined semantics from unstructured UGC collections. We use common vocabularies from Darwin Core (DwC), Friend-of-a-friend (FOAF), Semantically-Interlinked Online Communities (SIOC), Semantic Sensor Network (SSN), among others, to formalize the extracted datasets, hence, make them readily linkable. A nice consequence of this approach is that a multi-faceted browser can be quickly built to explore biodiversity information in large collections of UGC.

Categories and Subject Descriptors

H.3.5 [**Online Information System**]: [Web-based services]; H.5.3 [**Group and Organization Interfaces**]: [Web-based Interaction]; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms

Management, Design, Human Factors.

Keywords

Citizen Science, Crowdsourcing, Facebook, GeoSPARQL, Linked Data, Sensor Network, User-Generated Content (UGC).

1. INTRODUCTION

Citizen science is a crowdsourcing mechanism that 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 [3, 20]. Incorporation with trained volunteers participating in scientific studies as field assistants has a long history [26]. However, the landscape of citizen science has been transformed by modern Web services and communications enabling people around the world to spread information. Social media is one of significant tools in changing the ways information is produced and used in citizen science projects. A social media site can offer participants of citizen science projects not only a virtual environment for social interactions but also a platform for sharing, discussing, and modifying data together. On one hand, social media potentially provide situational awareness and opportunities for assistance on an individual level [12]. The communication channels make possible for participants to share and manage their own

[*] This research is supported in part by the Ministry of Science and Technology (grant no. 102-2627-M-001-009) and by the Endemic Species Research Institute, Council of Agriculture, Taiwan. We are grateful to Te–En Lin and his group at the Endemic Species Research Institute for their help with the collected data.

[†] Dong–Po Deng is also a PhD candidate at the Faculty of Geo–Information Science and Earth Observation (ITC), University of Twente.

[‡] Tyng–Ruey Chuang is also affiliated with the Research Center for Information Technology Innovation and the Research Center for Humanities and Social Sciences (Center for Geographic Information Science), both at Academia Sinica.

Linked Data on the Web (LDOW2014), April 8, 2014. Seoul, Korea.

sightings on a globally accessible database [29]. That is, the citizens are locally acting as human sensors, and social media are acting as platforms connecting these human sensors. On the other hand, social media enable scientists to reach out a large number of people, over a large geographic region and over an extended time period, to introduce them to citizen science projects. Therefore, the use of social media has greatly increased citizen participation and improved data collection process in citizen science projects. Such crowdsourced approach often can reduce cost and effort in data management and exchange [26].

However, to utilize social media for citizen science projects, there is a need to bridge a knowledge gap between human and the machine. In using social media for collecting participants' observations, it is often hard in controlling the quality of the content. Social media applications and services facilitate social interactions, but not scientific activities and data exchanges. Valuable scientific content is mixed up with huge amounts of noisy, low-quality, unstructured text and media. Often a crowdsourcing effort only creates human-readable content but not machine-readable data. Moreover, often the lack of sufficient metadata for crowdsourced data makes it difficult to derive meaningful interpretations from the data. Correspondingly, data integration and sharing in different knowledge domains is hampered. To achieve semantic computing on crowdsourced data, it requires not only text mining for extracting valuable information from user-generated content but also semantic enrichment for interpreting the meaning of the extracted information.

An ontology, as a "shared conceptualization", plays an important role for the basis of connections between datasets [14]. It is because an ontology presents a formal modeling for knowledge representation geared towards resolving semantic ambiguity, and consequently it contributes to the achievement of semantic interoperability between information communities [17]. Linked Data refers to the publication of structured data on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets [2]. Technically, the Linked Data paradigm combines knowledge representation technologies, e.g. RDF and OWL, with traditional Web technologies, e.g. HTTP and REST, for publishing and interlinking data and information [28]. The technologies enable a process evolving transition from current document-oriented Web into a Web of interlinked data and, ultimately, into the Semantic Web [1].

This paper reports our experiences on processing crowdsourced data from social media into interlinked data for the Web. The process can be elaborated by the following:

- how the crowdsourced observation data can be transformed and represented by an ontology of citizens as sensors,

- how the crowdsourced observation data can be interlinked with other Linked Data resources such as biodiversity (TaiCOL) and geospatial information (Geonames),

- how the crowdsourced observation data can be accessible to machines by using the Linked Data paradigm and be readable for humans by means of a faceted browser.

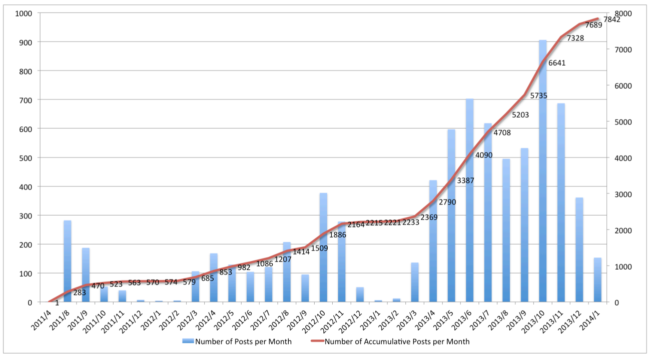


Figure 1: The growth of data in the Facebook group Reptile Road Mortality.

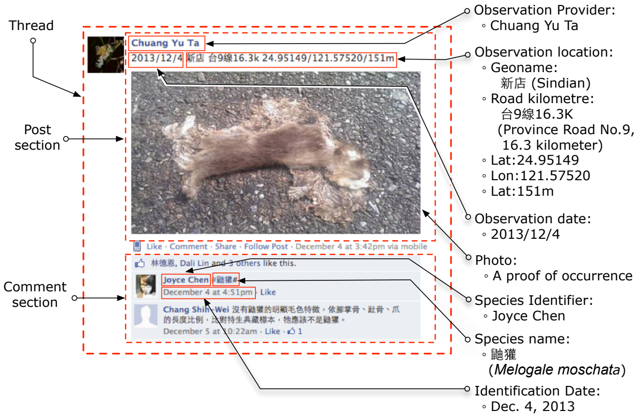
The paper is organized as follows. After introducing the citizen science project in Section 2, we describe how named-entities can be extracted from crowdsourced data, and the evaluation of the information extraction in Section 3. We explain the design of the synthesis ontology of citizens as sensors, and how crowdsourced data can be transferred to RDF data model in Section 4. In Section 5 we make spatiotemporal queries and present a faceted browser for the linked crowdsourced sensor data. Then we provide related work in Section 6. Finally we conclude in Section 7 with an outlook to future work.

2. REPTILE ROAD MORTALITY: A CITIZEN SCIENCE PROJECT

This section introduces the data collection in the citizen science project, *Reptile Road Mortality* (in Chinese, 路殺社). This citizen science project is hosted by the Endemic Species Research Institute, Council of Agriculture, Taiwan. The citizen science project aims to collect reports of dead animals that have been struck and/or killed by motor vehicles through the use of a Facebook group. The reason of using Facebook as a crowdsourced data collection platform is its high user base in the Taiwanese population. According to a statistic of Socialbakers[1], over half of Taiwanese population has a Facebook account. Facebook thus can be a good social place for recruiting participants. The number of participants in the Reptile Road Mortality is 4,187 at the end of year 2013, but only 618 persons ever posted at least one observation. The ratio of participants and contributive participants reveals the reality of mass collaboration, which is often said that 80% of the work is done by 20% of the people. Up to Jan. 4 2014, the group has assembled 7,842 posts as shown in Figure 1.

Any user possessing a Facebook account can join this citizen project and post his/her observations of roadkill animals. Figure 2 illustrates a roadkill observation posted in the Facebook group Reptile Road Mortality. Chuang Yu-Ta saw a killed animal on the road, so he took a photo and posted his observation with location and time description on the group. When Joyce read the post, she identified the species in Chuang Yu-Ta's photo and left the species name as comment. Thus, the roadkill observation was composed of photo, description of location and time, and identification

[1] http://www.socialbakers.com



Chuang Yu Ta

2013/12/4 新店 台9線16.3k 24.95149/121.57520/151m

Like · Comment · Share · Follow Post · December 4 at 3:42pm via mobile

林德恩, Dali Lin and 3 others like this.

Joyce Chen #鼬獾#

December 4 at 4:51pm · Like

Chang Shih-Wei 沒有鼬獾的明顯毛色特徵，依腳掌骨、趾骨、爪的長度比例，比對特生典藏標本，牠應該不是鼬獾。

December 5 at 10:22am · Like · 👍 1

The post is published on <http://www.facebook.com/238918712815615_694835510557264>.

Figure 2: A post on the Facebook group Reptile Road Mortality, as well as biodiversity observation information embedded in the post.

of species.

The participants of this citizen project would be asked to provide the location and time descriptions for the their observations. Because of privacy and security issues, Facebook strips metadata (EXIF) from the photos. Without EXIF data, a photo from Facebook is just an image; the photo cannot in itself indicates the date and location on which it was taken. The text messages accompanying the photos will be the main sources for extracting biodiversity information about the species in the photos.

Facebook posts can be retrieved using the Facebook Graph API[2] which enables developers to read from and write data to Facebook. This API offers a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections in between them (e.g., friend relationships, shared content, and photo tags).

3. INFORMATION EXTRACTION

3.1 Name-Entity Recognition

The data offered by the Facebook Graph API is structured around the Facebook social graph which is useful for processing social relationships. However, this citizen science project focuses on collecting occurrences of roadkill animals. The valuable information is in the photos for proving occurrences of roadkill animals, and in the texts for describing the time and location of occurrences of roadkill animals. To extract the information of occurrences of roadkill animals, we apply name-entity recognition to identify location, time, and species in Facebook posts and comments. Because the participants in the Facebook group use traditional Chinese as the communication language, our task of name-entity recognition actually aims at Chinese text processing. Chinese texts are character-based, not word-based. Moreover, there is often no space between characters in written Chinese sentences. This unique language feature leads to a challenge of word segmentation.

Several different algorithms have been proposed to deal with the challenge. Generally speaking, the algorithms can be classified into character-based and word-based approaches [33]. The character-based approaches ignore the concept of words, and use characters to extract word-level information in the construction of information extraction system. The word-based approaches apply lexicon to segment Chinese words. They often reply on a rich lexicon, sophisticated word segmentation, and/or syntactic analysis in extracting word-level information from documents [4]. However, existing Chinese lexicons are constructed for general applications. The lacks of domain-specific corpora often hamper the information extraction in specific domains such as geography and biodiversity. For example, the group Chinese Knowledge Information Processing (CKIP)[3] is continually building a Chinese word lexicon with rigid syntactic information. The lexicon now contains over 140,000 word entries, and is used in a corpus with over a million parsed sentences. This is a great research resource. Unfortunately using the CKIP lexicon for extracting location and species names is not efficient.

To efficiently extract species and location names from Facebook threads, it is necessary to constitute specific lexicons. We compiled a geo-name lexicon from the Taiwan Geographic Names database[4] and a species-name lexicon from the Taiwan Catalogue of Life databases (TaiCOL)[5] . Note that, however, species names and place names found in Facebook posts and comments are not always in these two specific lexicons. The name-entity recognition approach we use was elaborated in a paper we previously published [10].

3.2 Evaluation of Name-Entity Recognition

Precision and recall are the basic measures used in Natural Language Processing to evaluate information extraction methods [9, 18, 30]. Generally speaking, it needs a training dataset to assess the quality of information extraction. Our training dataset is generated by domain experts. While the training dataset is considered as a positive set, the names extracted by Name-Entity Recognition (NER) is a negative set. According to whether an identification is correct, four sets can be distinguished: true positive, false positive, true negative, and false negative. From the statistical point of view, false negative are Type I errors, and false positives are Type II errors. Precision is the ratio of the number of correct names identified by both NER and domain experts (True Positive) to the total number of incorrect and correct names identified by NER (True Positive + False Positive) (Eq. 1). Recall is the ratio of the number of correct names identified by both NER and domain experts (True Positive) to the total number of correct names identified by domain experts (True Positive + False Negative) (Eq. 2). The F-score is an overall metric that is calculated from both precision and recall, treating these two metrics as equally important (Eq. 3).

$$Recall = \frac{|name_{actual} \cap name_{predict}|}{|name_{actual}|} \tag{1}$$

[2] https://developers.facebook.com/docs/graph-api

[3] http://ckip.iis.sinica.edu.tw/CKIP/engversion/index.htm

[4] http://placesearch.moi.gov.tw

[5] http://col.taibif.tw

Table 1: Confusion matrix of information extraction assessment.

	Expert determine	Expert not determine
NER predict	282	7
NER not predict	10	101

$$Precision = \frac{|name_{actual} \cap name_{predict}|}{|name_{predict}|} \quad (2)$$

$$F\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (3)$$

where $name_{actual}$ is the set of place names or species names that has been identified from Facebook messages by domain experts, and $name_{predict}$ is the set of place names or species names that has been identified from Facebook messages by the NER.

400 posts are randomly selected from the entire 7,842 posts for the evaluation. The confusion matrix of the information extraction assessment is shown in Table 1. Thus, the precision is $282/(282+7) = 0.9758$, the recall is $282/(282+10) = 0.9656$, and the F-score is 2.8973.

4. AN ONTOLOGY FOR CITIZENS AS SENSORS

4.1 A synthesis of social networks and sensor networks

Before we begin to transform the crowdsourced content to RDF, we first develop an ontology for not only expressing the notions of "Citizens as Sensors" but also formalizing the extracted name-entities, e.g. species and geospatial names. To make linked data interoperable, the ontology reuses suitable vocabularies from the existing ontologies as many as possible. Since the crowdsourced dataset is retrieved from Facebook, a social media site, its content can be mapped to RDF using existing social semantic web ontologies. The Semantically Interlinked Online Communities (SIOC)[6] is used for representing the content of the Facebook group *Reptile Road Mortality*, e.g. threads, posts, and images. The Friend of a Friend (FOAF)[7] can be used to describe content creators. Figure 3 shows the vocabularies of SIOC and FOAF used in our ontology.

In this study, "Citizens as Sensors" means that a Citizen voluntarily reporting his/her observations via social media for a citizen science project. The citizen acts as a Sensor which enables automatic measurement and/or recording of physical properties. To express the notion, the vocabularies of W3C Semantic Sensor Network (SSN) ontology are used to express the content from social networks. Conceptually, the action that a participant reports her/his roadkill observation matches the pattern of Stimulus-Sensor-Observation. The pattern describes a process that a sensor transforms a stimulus from the physical world into an observation and thereby it allows us to reason about the observed properties of particular features of interest [15]. A roadkill animal actually is the stimulus which triggers a citizen to a post her/his observations on the Facebook at specific time and location. Also, the species of the animal is the feature of interest. Figure 4 displays the use of the vocabularies of the SSN ontology in our ontology.

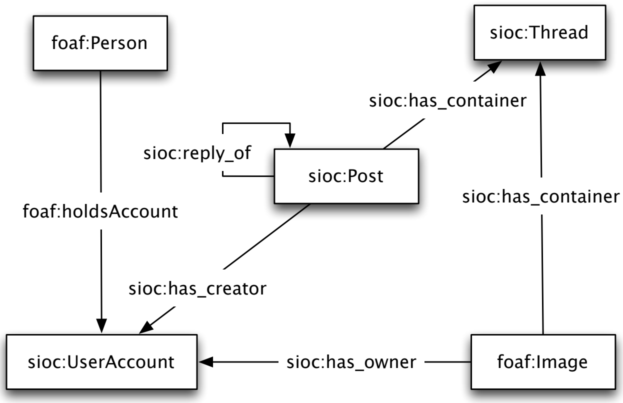


Figure 3: The vocabularies of SIOC and FOAF used in our ontology.

However, the citizen is a person and cannot exactly be regarded as a sensor. The persons can be expressed as `foaf:Person`, and the sensors can be defined to `ssn:Sensor`. All individuals of `foaf:Person` cannot be the same as all individuals of `ssn:Sensor`. Only some of these individuals can be expressed as not only `foaf:Person` but also `ssn:Sensor`. To clarify the concept, we create the class `Citizen_As_Sensor` which is a subclass of the intersection of the two classes. That is, an individual of the class `Citizen_As_Sensor` can be an instance of both classes. But the instances of `foaf:Person` or `ssn:Sensor` are not necessary to be the individuals of the class `Citizen_As_Sensor`. Moreover, the same situation occurs for `ssn:SesnorOutput`, as some instances are in `sioc:Post` or in `sioc:Image`. Therefore, we define the class `Post_As_SesnorOutput` to be in the intersection of `sioc:Post` and `ssn:SensorOutput`, and the class `Image_As_SesnorOutput` to be a subclass of both `sioc:Image` and `ssn:SensorOutput`.

4.2 Formalizations of the extracted name-entities

4.2.1 Geospatial information

In the process of information extraction, name entity recognition is used to identify the geospatial and species names. The extraction of geospatial information includes not only location names (such as names of populated places and point of interests) and road names with kilometers but also coordinates (longitude and latitude). If coordinates were not written in the texts of observation posts, the location names would be used to retrieve the longitude and latitude. To semantically encode geospatial data, we use the vocabularies of Open Geospatial Consortium (OGC) GeoSPARQL. The GeoSPARQL is one of OGC standards which provides three main components for semantically encoding geographic data: (1) The definitions of vocabularies for representing features, geometries, and their relationships; (2) A set of domain-specific, spatial functions for use in SPARQL queries; (3) A set of query transformation rules [21].

The ontology of the GeoSPARQL standard includes three

[6] http://sioc-project.org

[7] http://www.foaf-project.org

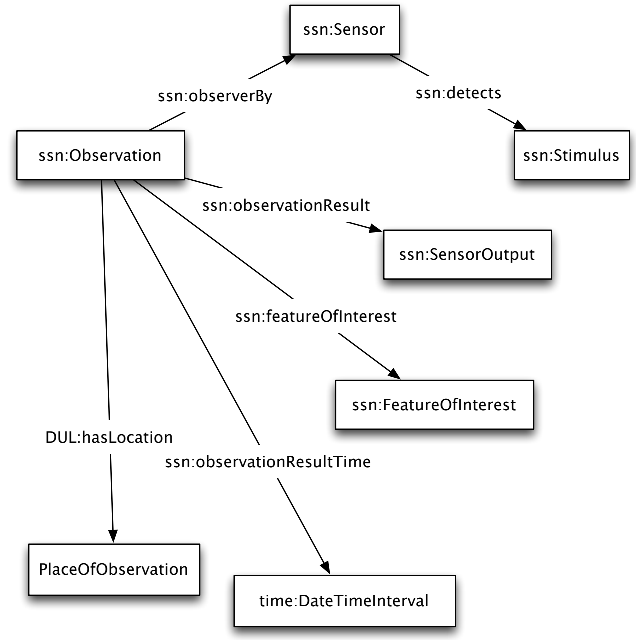


Figure 4: The vocabularies of W3C SSN used in our ontology.

main classes: geo:SpatialObject , geo:Features, and geo:Geometry . The geo:Feature and geo:Geometry are the subclass of geo:SpatialObject. The geo:Feature class represents features, which are abstractions of real world phenomena. The concept of feature is derived from ISO 19109 General Feature Model. The geo:Geometry, expressing spatial geometries of the features, has sixteen subclasses defining a hierarchy of geometry types such as point, polygon, curve, arc, and multi-curve. These geometry classes are derived from ISO 19107 Spatial Schema. RDF literals are used to store geometry values. There are two ways to store geometry values via RDF literals: Well Known Text (WKT) and Geography Markup Language (GML). The geo:asWKT and geo:asGML properties map between the geometry entities and the geometry literals. Geometry values for these two properties use the geo:WKTLiteral and geo:GMLLiteral data types respectively. Figure 5 shows the classes and properties of GeoSPARQL used in our ontology.

Although DUL:hasLocation is usually a predicate in between ssn:Observation and DUL:Entity in W3C SSN, it actually can be a property between any entities. To clarify the place of observation, we create a class PlaceOfObservation which is a subclass of both of DUL:Entity and geo:Feature. The class PlaceOfObservation not only keeps the DUL:hasLocation property but also inherits the formal geospatial concepts from geo:Feature. As for the time of an ssn:Observation event, ssn:observationResultTime can be a predicate in between the class ssn:Observation and the class time:DateTimeInterval.

4.2.2 Biodiversity information

Discovery and inventory of specimen data is a fundamental work in biodiversity informatics. With the development of Internet technologies, the aggregation and dissemination of biodiversity data has increased the scale from regional to global, and has broaden the scope beyond that of establishing species ranges [16]. To reach global biodiversity data coordination, standardized metadata vocabularies i.e. Darwin Core is used to develop data infrastructures for sharing biodiversity data. Darwin Core is a standard for sharing data about biodiversity — the occurrence of life on earth and its associations with the environment [32]. However, Darwin Core is comprised of technology-independent vocabularies. The classes in Darwin Core are categories and have no formal domain declarations for vocabularies [31]. To improve the knowledge representation of Darwin Core, Darwin-SW[8] designs the properties between classes and formalizes the classes including five existing core classes of Darwin Core (i.e. Taxon, Event, Identification, Location, Occurrence) and two new ones (i.e. Token and Individual Organism). Figure 6 shows the classes and properties of Darwin Core are used in our ontology.

Traditionally, a specimen collecting all or part of an organism serves as an evidence for the occurrence of the organism, and is a basis for identifying the organism to a taxon concept. However, the documentation process nowadays has many possible methods such as images, sound, or DNA sequences. The class dsw:Token is used to represent evidences from the classes dwctype:Occurrence and dwctype:Identification. To connect Darwin Core to W3C SSN, we create classes Token_As_FeatureOfInterest and Occurrence_As_Stimulus. Token_As_FeatureOfInterest is a subclass of the intersection of ssn:FeatureOfInterest and dwstype:Token. The class Occurrence_As_Stimulus is in the intersection of ssn:Stimulus and dwctype:Occurrence.

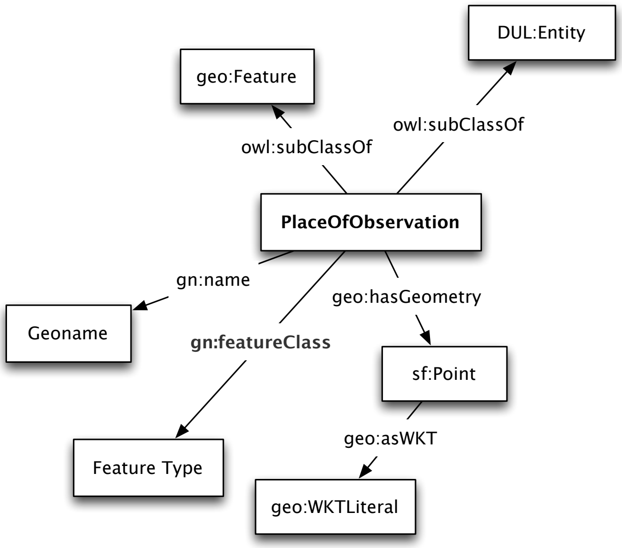


Figure 5: The vocabularies of GeoSPARQL used in our ontology.

4.3 Transformations from the extracted name-entities to the RDF model

[8] https://code.google.com/p/darwin-sw/

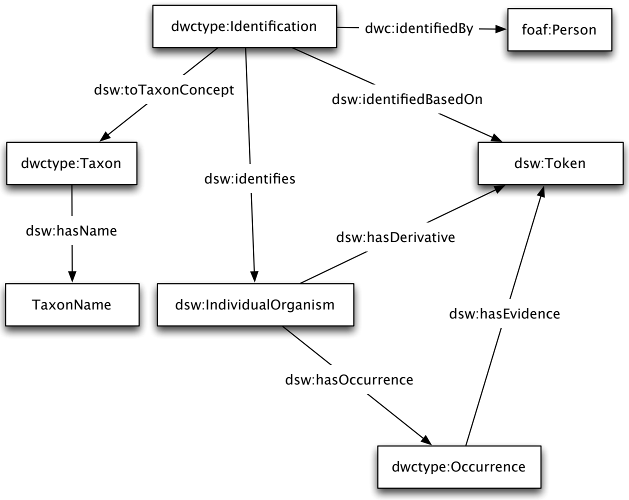
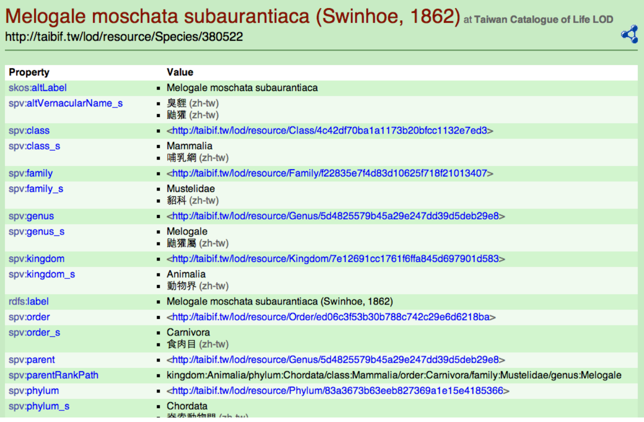


Figure 6: The vocabularies of Darwin-SW used in our ontology.

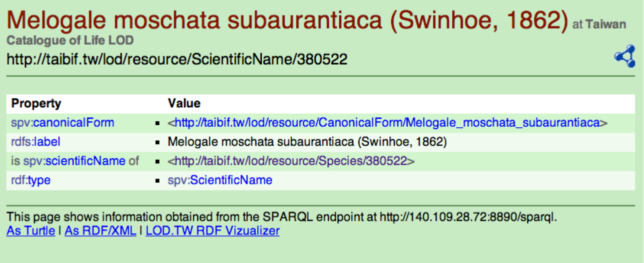


Melogale moschata subaurantiaca (Swinhoe, 1862) at Taiwan Catalogue of Life LOD
http://taibif.tw/lod/resource/Species/380522

Property	Value
skos:altLabel	▪ Melogale moschata subaurantiaca
spv:altVernacularName_s	▪ 臭貍 (zh-tw) ▪ 鼬獾 (zh-tw)
spv:class	▪ <http://taibif.tw/lod/resource/Class/4c42df70ba1a1173b20bfcc1132e7ed3>
spv:class_s	▪ Mammalia ▪ 哺乳綱 (zh-tw)
spv:family	▪ <http://taibif.tw/lod/resource/Family/f22835e7f4d83d10625f718f21013407>
spv:family_s	▪ Mustelidae ▪ 貂科 (zh-tw)
spv:genus	▪ <http://taibif.tw/lod/resource/Genus/5d4825579b45a29e247dd39d5deb29e8>
spv:genus_s	▪ Melogale ▪ 鼬獾屬 (zh-tw)
spv:kingdom	▪ <http://taibif.tw/lod/resource/Kingdom/7e12691cc1761f6ffa845d697901d583>
spv:kingdom_s	▪ Animalia ▪ 動物界 (zh-tw)
rdfs:label	▪ Melogale moschata subaurantiaca (Swinhoe, 1862)
spv:order	▪ <http://taibif.tw/lod/resource/Order/ed06c3f53b30b788c742c29e6d6218ba>
spv:order_s	▪ Carnivora ▪ 食肉目 (zh-tw)
spv:parent	▪ <http://taibif.tw/lod/resource/Genus/5d4825579b45a29e247dd39d5deb29e8>
spv:parentRankPath	▪ kingdom:Animalia/phylum:Chordata/class:Mammalia/order:Carnivora/family:Mustelidae/genus:Melogale
spv:phylum	▪ <http://taibif.tw/lod/resource/Phylum/83a3673b63eeb827369a1e15e4185366>
spv:phylum_s	▪ Chordata

Figure 8: The taxon concept of extract species name is linked to a URI in TaiBIF.

Assembling the above-mentioned vocabularies, we can create the ontology of "Citizen as Sensor", as shown in Figure 7. Such designed ontology plays as the schema for transforming crowdsourced content to linked sensor data. Take Figure 2 as example, we can correspondingly transform the user-generated content to RDF data, as shown in the Appendix. The extracted name entities of species and place names are pointed to by URLs. The word "鼬獾" (*M*elogale moschata subaurantiaca) is identified as a taxon <http://taibif.tw/lod/resource/Species/380522>, as shown in Figure 8, and mapped to the scientific name <http://taibif.tw/lod/resource/ScientificName/380522>, as shown in Figure 9. The extracted place name "新店" (Sindian) also is linked to a URI in Taiwan Geographic Name whose URIs are all mapped to Geonames.org, as shown in Figure 10.

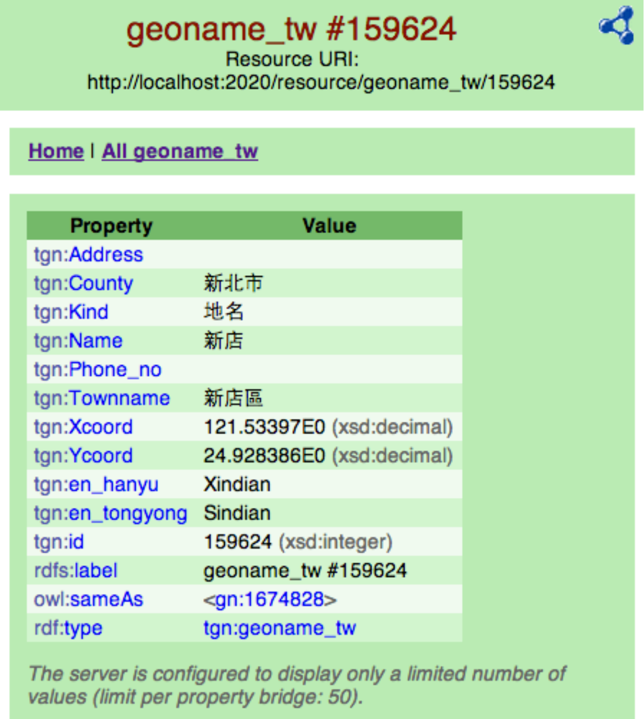


Melogale moschata subaurantiaca (Swinhoe, 1862) at Taiwan Catalogue of Life LOD
http://taibif.tw/lod/resource/ScientificName/380522

Property	Value
spv:canonicalForm	▪ <http://taibif.tw/lod/resource/CanonicalForm/Melogale_moschata_subaurantiaca>
rdfs:label	▪ Melogale moschata subaurantiaca (Swinhoe, 1862)
is spv:scientificName of	▪ <http://taibif.tw/lod/resource/Species/380522>
rdf:type	▪ spv:ScientificName

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

Figure 9: The taxon name of extract species name is linked to a URI in TaiBIF.



geoname_tw #159624
Resource URI:
http://localhost:2020/resource/geoname_tw/159624

Home | All geoname_tw

Property	Value
tgn:Address	
tgn:County	新北市
tgn:Kind	地名
tgn:Name	新店
tgn:Phone_no	
tgn:Townname	新店區
tgn:Xcoord	121.53397E0 (xsd:decimal)
tgn:Ycoord	24.928386E0 (xsd:decimal)
tgn:en_hanyu	Xindian
tgn:en_tongyong	Sindian
tgn:id	159624 (xsd:integer)
rdfs:label	geoname_tw #159624
owl:sameAs	<gn:1674828>
rdf:type	tgn:geoname_tw

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

Figure 10: The extract place name points to a URI in Taiwan Geographic Name.

5. SPATIOTEMPORAL QUERIES

Since the geospatial information is formalized by the vocabularies of OGC GeoSPARQL, information in our RDF dataset can be retrieved via spatiotemporal queries. This study uses BBN Parliament, which is an open source triple store developed by Raytheon BBN Technologies. The BBN Parliament is compliant with OGC GeoSPARQL standard, and supports spatial and non-spatial SPARQL queries. Using BBN parliament, we build a GeoSPARQL endpoint[9]. for the linked crowdsourced sensor dataset. The following lists a GeoSPARQL query, and Figure 11 is the result of the query.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX sf: <http://www.opengis.net/ont/sf#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX eoe: <http://lod.tw/ontologies/eoe.owl#>
PREFIX DUL: <http://www.loa-cnr.it/ontologies/DUL.owl#>
PREFIX ssn: <http://purl.oclc.org/NET/ssnx/ssn#>
```

[9] http://lod.tw/parliament/

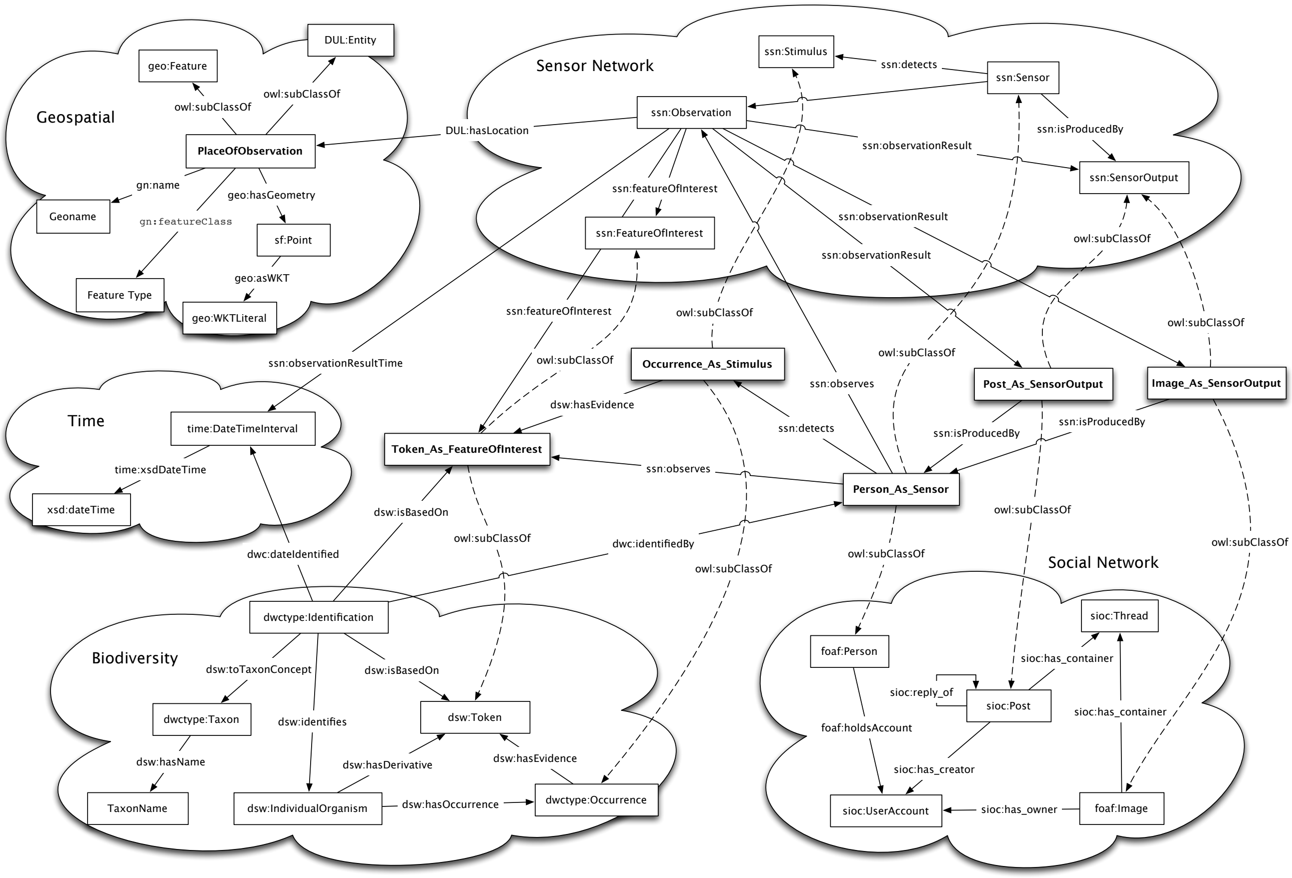
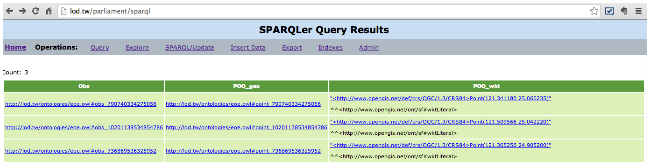


Figure 7: The ontology of "Citizen as Sensor".



lod.tw/parliament/sparql

SPARQLer Query Results

Home Operations: Query Explore SPARQL/Update Insert Data Export Indexes Admin

Count: 3

Obs	POO_geo	POO_wkt
http://lod.tw/ontologies/eoe.owl#obs_790740334275056	http://lod.tw/ontologies/eoe.owl#point_790740334275056	"<http://www.opengis.net/def/crs/OGC/1.3/CRS84>Point(121.341180 25.060235)" ^^<http://www.opengis.net/ont/sf#wktLiteral>
http://lod.tw/ontologies/eoe.owl#obs_10201138534854786	http://lod.tw/ontologies/eoe.owl#point_10201138534854786	"<http://www.opengis.net/def/crs/OGC/1.3/CRS84>Point(121.509566 25.042220)" ^^<http://www.opengis.net/ont/sf#wktLiteral>
http://lod.tw/ontologies/eoe.owl#obs_736869536325952	http://lod.tw/ontologies/eoe.owl#point_736869536325952	"<http://www.opengis.net/def/crs/OGC/1.3/CRS84>Point(121.365256 24.905200)" ^^<http://www.opengis.net/ont/sf#wktLiteral>

Figure 11: The result of a spatiotemporal query.

```
SELECT Distinct ?Obs ?POO_geo ?POO_wkt
WHERE{
        ?Obs a ssn:Observation;
             DUL:hasLocation ?POO ;
             ssn:observationResultTime ?Int .
        ?POO geo:hasGeometry ?POO_geo .
        ?POO_geo geo:asWKT  ?POO_wkt .
        ?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 )

}
```

To efficiently browse the RDF triples, we develop a faceted viewer[10] including a taxon tree, a social relation graph, and an observation map, as shown on Figure 12. The taxon tree can visualize the identified species names via their taxon concepts such as kingdom, phylum, class, order, family, and genus. The social relation graph shows the connections in between the participants in the citizen science project. It can be used to view who observes what species, and where the species occurs. To display locations of species occurrences, the coordinates are used to pin the species on the map. Also a timeline is used to show the times of the species occurrences.

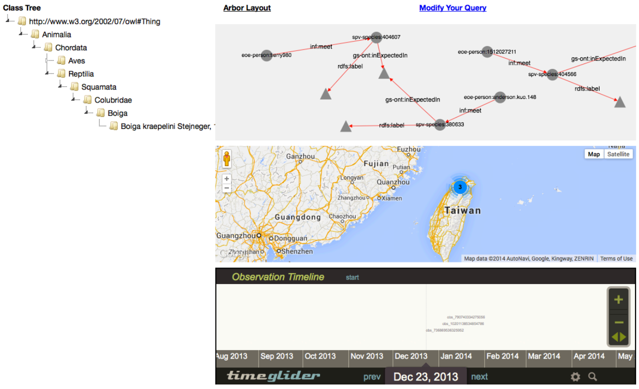


Figure 12: A faceted viewer.

[10] http://taibif.tw/vgd/ldow2014/viewer.php

6. RELATED WORK

Traditionally, in order to ensure the quality of data collections, training and educating volunteers by experts or experienced participants is a common method in citizen science [11]. The volunteers, thus, are capable to fill designated forms, to use well-defined terms, and/or to follow default steps on the web for reporting their observations. The user-contributed data, thus, can be fitted to a default data model. However, this method is difficult to apply when citizen science projects depend on Web applications and services. It is argued there exists an inherent trade-off between data quality and data quantity [23]. The growth of data quantity will be slow if the data contribution is restricted to experts or trained volunteers. On the contrary, data volume often increases rapidly if data contribution is entirely open to volunteers. But data quality is hard to guarantee. Such volunteered contributions can easily be imperfect (e.g. erroneous, incomplete, or fraudulent) and unstructured (e.g. in the form of texts and/or images) [6, 10]. Crowdsourcing is the first step of data collection in citizen science. After preprocessing and cleaning up the noise in crowdsourced data, it can provide more valuable information to scientists than what raw data can do. The role of semantic web technologies is increasingly important for tackling crowdsourced data. To enable semantic computing to process crowdsourced data, Sheth proposed semantics-empowered social computing architecture for dealing with crowdsourced data [25]. The architecture emphasized the use of domain-specific or spatial-temporal-thematic ontologies for extracting meaning in the data.

The idea of citizen sensing is not new. Goodchild coined the term "Volunteered Geographic Information" (VGI) to describe a contemporary trend where Web technologies empower a network of human sensors voluntarily reporting and interpreting in-situ information [13]. Sheth also described Internet users or Web-enabled social community as citizens. The ability to interact with Web 2.0 services can augment these citizens into citizen sensors [24]. He further explained the advantages of "human-in-the-loop sensing", emphasizing the background knowledge and past experiences from human in citizen sensing. Janowicz and Compton developed the Stimulus-Sensor-Observation ontology pattern which forms the Semantic Sensor Network (SSN) ontology as developed by the W3C SSN Incubator Group [15]. The design pattern provides a knowledge representation for integration of social web and sensor web. Some studies not only transformed the crowdsourced data to a standard format such as RDF but also leverage the power of the SSN ontology to describe the sensors on mobile devices for passenger information system and in emergency reporting applications on microblogging platforms [6, 7].

Linked Data has established itself as the de facto means for the publication of structured data over the Web. More and more ICT ventures offer innovative data management services on the top of Linked Open Data (LOD) [27]. Ortmann et al. described an approach based on LOD to alleviating the integration problems of crowdsourced data, and to improving the exploitation of crowdsourced data in disaster management [22]. To solve the problem of structural and semantic interoperability, they also suggested engage people in processing unstructured observations into structured RDF-triples according to Linked Open Data principles. The process would increase the impact of crowdsourced data in disaster management, and it shall help humanitarian agencies make informed decisions. The exploitation of external semantic resources to disambiguate contents is often said to be an effective method. To enrich the semantics of folksonomies, Choudhury et al. not only built up relations among tags via statistical analysis but also integrated the structured tags with the linked data cloud through the DBpedia [5]. Mendes et al. proposed a Linked Open Social Signals architecture for collection, semantic annotation, and analysis of real-time social signals from microblogging data [19]. The design of Linked Data management often aim to "reach a high level of automation with respect to the processing of an open and decentralized data space bringing together data sources published by different parties, of varying quality and using heterogeneous conceptual schemas and vocabularies" [27]. Crowley et al. proposed a generic framework for aggregating and linking heterogeneous data from various sources and transforming them to Linked Data [8]. The framework allows reuse and integration of the produced data with other data resources (including social media and sensors) enabling spatial business intelligence for various domain-specific applications.

7. CONCLUSION AND FUTURE WORK

Social media creates new opportunities for citizen science. The information created from social media is considered a new resource for scientific works. Meanwhile, the use of social media in citizen science projects also brings new issues to research data. 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.

This study deals with crowdsourced content from a citizen science project via a "Citizen as Sensor" ontology. The processed data is formalized by inheriting the concepts from the ontology. Thus, the extracted name entities can be mapped to the existing resources and linked to domain-specific concepts. With clarified domain-specific semantics, the triplified data can be applied in faceted exploration for new knowledge. 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.

8. REFERENCES

[1] S. Auer, J. Lehmann, and A.-C. N. Ngomo. Introduction to linked data and its lifecycle on the web. In *Proceedings of the 7th International Conference on Reasoning Web: Semantic Technologies for the Web of Data*, RW'11, pages 1–75, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[3] G. Chatzimilioudis, A. Konstantinidis, C. Laoudias, and D. Zeinalipour-Yazti. Crowdsourcing with smartphones. *Internet Computing, IEEE*, 16(5):36–44, Sept 2012.

[4] L.-F. Chien. Pat-tree-based keyword extraction for chinese information retrieval. In *ACM SIGIR Forum*, volume 31, pages 50–58. ACM, 1997.

[5] S. Choudhury, J. G. Breslin, and A. Passant. Enrichment and ranking of the YouTube tag space and integration with the linked data cloud. In *International Semantic Web Conference*, volume 5823 of *LNCS*, pages 747–762. Springer, 2009.

[6] D. Corsar, P. Edwards, N. Velaga, J. Nelson, and J. Z. Pan. Short paper: Addressing the challenges of semantic citizen-sensing. In *Proceedings of the 4th International Workshop on Semantic Sensor Networks(SSN'11)*, pages 101–106, 2011.

[7] D. Crowley, A. Passant, and J. G. Breslin. Short paper: Annotating microblog posts with sensor data for emergency reporting applications. In *Proceedings of the 4th International Workshop on Semantic Sensor Networks (SSN'11)*, pages 95–100, 2011.

[8] D. N. Crowley, M. Dabrowski, and J. G. Breslin. Decision support using linked, social, and sensor data. In *Proceedings of the Nineteenth Americas Conference on Information Systems*, 2013.

[9] K. Crowston, E. E. Allen, and R. Heckman. Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, 15(6):523–543, 2012.

[10] D.-P. Deng, G.-S. Mai, C.-H. Hsu, T.-R. Chuang, T.-E. Lin, H.-H. Lin, K.-T. Shao, R. Lemmens, and M.-J. Kraak. Using social media for collaborative species identification and occurrence: Issues, methods, and tools. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '12, pages 22–29, New York, NY, USA, 2012. ACM.

[11] A. Flanagin and M. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 72:137–148, 2008.

[12] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3):10–14, May 2011.

[13] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.

[14] T. R. Gruber. A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5:199–220, 1993.

[15] K. Janowicz and M. Compton. The stimulus-sensor-observation ontology design pattern and its integration into the semantic sensor network ontology. In *Proceedings of The 3rd International workshop on Semantic Sensor Networks 2010 (SSN10) in conjunction with the 9th International Semantic Web Conference (ISWC 2010)*, ISWC'10, 2010.

[16] S. Kelling, J. Gerbracht, D. Fink, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, and C. P. Gomes. A human/computer learning network to improve biodiversity conservation and research. *AI Magazine*, 34(1):10–20, 2013.

[17] R. Lemmens and D. Deng. Web 2.0 and semantic web: Clarifying the meaning of spatial features. *Semantic Web meets Geopatial Applications, AGILE*, 2008.

[18] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[19] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked open social signals. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 224–231. IEEE Computer Society, 2010.

[20] G. Newman, D. Zimmerman, A. Crall, M. Laituri, J. Graham, and L. Stapel. User-friendly web mapping: lessons from a citizen science website. *Int. J. Geogr. Inf. Sci.*, 24(12):1851–1869, Dec. 2010.

[21] OGC. GeoSPARQL - A Geographic Query Language for RDF Data. Technical report, http://www.opengeospatial.org/standards/geosparql, 2011.

[22] J. Ortmann, M. Linbu, W. Dong, and T. Kauppinen. Crowdsourcing linked open data for disaster management. In W. W. Cohen and S. Gosling, editors, *Terra Cognita*, pages 11–22, 2011.

[23] J. Parsons, R. Lukyanenko, and Y. Wiersma. Easier citizen science is better. *Nature*, 471(7336):37, Mar. 2011.

[24] A. Sheth. Citizen sensing, social signals, and enriching human experience. *Internet Computing, IEEE*, 13(4):87–92, July 2009.

[25] A. Sheth. Computing for human experience: Semantics-empowered sensors, services, and social computing on the ubiquitous web. *Internet Computing, IEEE*, 14(1):88–91, 2010.

[26] J. Silvertown. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467 – 471, 2009.

[27] E. Simperl. Crowdsourcing semantic data management: Challenges and opportunities. In *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 1:1–1:3, New York, NY, USA, 2012. ACM.

[28] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.

[29] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282 – 2292, 2009.

[30] K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. L. Johnson, C. Roeder, J. D. Choi, C. Funk, Y. Malenkiy, M. Eckert, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics*, 13(1):207, 2012.

[31] C. Webb and S. Baskauf. Darwin-sw: Darwin core data for the semantic web. *TDWG Annual Meeting;*

2011-10-18, 2011.

[32] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, 7(1):e29715, 2012.

[33] K.-F. Wong, W. Li, R. Xu, and Z.-s. Zhang. *Introduction to Chinese Natural Language Processing*. Morgan & Claypool Publishers, 2010.

APPENDIX

A. FROM UGC TO ENRICHED RDF DATA

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix DUL: <http://www.loa-cnr.it/ontologies/DUL.owl#> .
@prefix dwc: <http://rs.tdwg.org/dwc/terms/> .
@prefix dsw: <http://purl.org/dsw/> .
@prefix taibif: <http://taibif.tw/lod/resource/ScientificName/> .
@prefix ssn: <http://purl.oclc.org/NET/ssnx/ssn#> .
@prefix sf: <http://www.opengis.net/ont/sf#> .
@prefix w3c_geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix schema: <http://schema.org/> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dwctype: <http://rs.tdwg.org/dwc/dwctype/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix eoe: <http://lod.tw/ontologies/eoe.owl#> .
@prefix fb: <https://www.facebook.com/> .
@prefix tgn: <http://lod.tw/placenames/> .
@prefix taxon: <http://taibif.tw/lod/resource/Species/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix gn: <http://www.geonames.org/ontology#> .

eoe:img_559070840853748 rdf:type eoe:Image_As_SensorOutput ,
                                 owl:NamedIndividual ;
                        sioc:has_container eoe:thread_559070840853748 ;
                        sioc:has_owner fb:100002525111203 ;
                        ssn:isProducedBy eoe:person_100002525111203 .

fb:238918712815615_694835510557264 rdf:type eoe:Post_As_SensorOutput ,
                                            owl:NamedIndividual ;
                                   sioc:has_container eoe:thread_559070840853748 ;
                                   sioc:has_creator fb:100002525111203 ;
                                   ssn:isProducedBy eoe:person_100002525111203 .

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

eoe:token_559070840853748 rdf:type eoe:Token_As_FeatureOfInterest ,
                                   owl:NamedIndividual .

eoe:idv_238918712815615_694835510557264 rdf:type dsw:IndividualOrganism ,
                                                 owl:NamedIndividual .

eoe:obs_559070840853748 rdf:type ssn:Observation ,
                                 owl:NamedIndividual ;
                        ssn:observationResultTime eoe:obs_time_559070840853748 ;
                        DUL:hasLocation eoe:placeOfOb_559070840853748 ;
                        ssn:observationResult eoe:img_559070840853748 ,
                                              fb:238918712815615_694835510557264 ;
                        ssn:featureOfInterest eoe:token_559070840853748 ;
                        ssn:observedBy eoe:person_100002525111203 .

eoe:obs_time_559070840853748 rdf:type time:DateTimeInterval ,
                                      owl:NamedIndividual ;
                             time:xsdDateTime "2013-12-04T07:42:15"^^xsd:dateTime .

eoe:iden_time_559070840853748 rdf:type time:DateTimeInterval ,
                                       owl:NamedIndividual ;
                              time:xsdDateTime "2013-12-11T07:42:15"^^xsd:dateTime .

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 .

eoe:thread_559070840853748 rdf:type sioc:Thread ,
                                    owl:NamedIndividual ;
                           sioc:has_container fb:groups/roadkilled .

eoe:occr_559070840853748 rdf:type eoe:Occurrence_As_Stimulus ,
                                  owl:NamedIndividual ;
                         dsw:hasEvidence eoe:token_559070840853748 .

eoe:person_100002525111203 rdf:type eoe:Person_As_Sensor ,
                                    owl:NamedIndividual ;
                           rdfs:label "Chuang Yu Ta" ;
                           ssn:detects eoe:occr_559070840853748 ;
                           ssn:observes eoe:token_559070840853748 ;
                           foaf:account fb:100002525111203 .

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