# Using Social Media for Collaborative Species Identification and Occurrence: Issues, Methods, and Tools

Dong–Po Deng[*, §]
dongpo@itc.nl

Guan–Shuo Mai[†]
trashmai@gmail.com

Cheng–Hsin Hsu[†]
jimshsu@gate.sinica.edu.tw

Tyng–Ruey Chuang[*]
trc@iis.sinica.edu.tw

Te–En Lin[‡]
dnlin@tesri.gov.tw

Hsu–Hong Lin[‡]
shlin@tesri.gov.tw

Kwang–Tsao Shao[†]
zoskt@gate.sinica.edu.tw

Rob Lemmens[§]
lemmens@itc.nl

Menno–Jan Kraak[§]
kraak@itc.nl

[*] Institute of Information Science
[†] Biodiversity Research Center
Academia Sinica
128 Academia Road, Sec. 2
Nangang 115, Taipei City
Taiwan

[‡] Endemic Species Research Institute
Council of Agriculture
1 Mingsheng East Road
Jiji 552, Nantou County
Taiwan

[§] Faculty of Geo–Information Science and Earth Observation (ITC)
Twente University
PO Box 217
7500 AE Enschede
The Netherlands

## ABSTRACT

The emergence of social media enables people to interact with others on the web in ways that are media-rich ("updates" or "posts" can be text, photo, audio, video, etc), time-shifted (correspondence need not happen at once or within a pre-defined time frame), and social in nature. 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.

As social media is not designed for scientific data collection and analysis, there is a problem in transferring unstructured information items (e.g. free-form text, unidentified images, etc.) often found in social media to structured data records for scientific tasks. To help bridge this gap, we propose an approach comprised of three steps: (1) Information Extraction, (2) Information Formalization, and (3) Information Reuse. We apply this approach to processing posts and comments from two Facebook interest groups on species observations. Our study demonstrates that with principled methods and proper tools, crowdsourced social media contents such as those from Facebook interest groups can be used for collaborative species identification and occurrence.


*ACM SIGSPATIAL GEOCROWD'12,* November 6, 2012. Redondo Beach, CA, USA


## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial Database and GIS; H.3.5 4 [**Online Information System**]: Web-based services; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

## General Terms

Algorithms, Design, Experimentation, Human Factors.

## Keywords

Citizen Science, Facebook, Linking Open Data, Social Media, Volunteered Geographic Information.

## 1. INTRODUCTION

The power of social media is increasing its influence on the production of scientific works. A large number of social-media users often contribute in situ information on the Web. They are often considered as human sensors who actively report what are happening in their surroundings [9]. Voluntary participation has become an important part of citizen science. The emergence of social media offers new opportunities to recruit more participants to citizen science projects. Utilizing social media to engage with a large number of netizens can be a way to improve data collection over a large geographic region and a long time span. However, the transformation from citizen science to netizen science is a problem. There is no social media specifically designed for citizen science. Social media applications and services facilitate social interactions, but not scientific activities and data analyses. Besides, the crowdsourced information contributed by netizens through social media is often in unstructured data format such as text and image. It is a challenge to process unstructured data collections for scientific purposes. In

order to engage social media with citizen science, there is a need to develop novel methods to transfer unstructured data to structured data.

This paper describes an approach consisting of three main processing phases to facilitate the use of social media in citizen science projects. The first phase is information extraction, which applies Natural Language Processing (NLP) tools to extract useful information from social media contents. The next phase is information formalization, which uses Semantic Web techniques to formalize the extracted information. The final phase is information reuse and improvement, which utilizes the collected structured information to further exploit social media for citizen science projects. The aim of this approach is nothing less than facilitating social media services for citizen science projects. More specifically, this paper deals with the semantic gap between uncontrolled structures typically found in crowdsourced information media and well-structured datasets provided and required by scientific communities. That is, our work not only help the scientists obtain information from the crowd, but also empower netizens to contribute structured data. The paper is organized as follows. Section 2 reviews the relevant literature. Section 3 provides a use case to illustrate our motivations and the purposes of this study. In Section 4, we present our empirical study in developing methods and tools to deal with crowdsourced information from Facebook groups for citizen science projects. We conclude in Section 5 with a discussion about our experiences and some suggestions for future works.

# 2. STATE OF THE ART IN CITIZEN SCIENCE PROJECTS

## 2.1 Trends in Crowdsourced Data Collection and Analysis

Citizen science has been used in many situations to refer to many scenarios in which citizens participate in the scientific process along with professionals [14]. Citizen science typically involves trained volunteers participating in scientific studies as field assistants who collect data [5], especially in ecological and environmental research. The use of Internet has greatly increased participation and improved data collection in citizen science projects. This eliminated the costs and efforts in data management and exchange [19]. Volunteer participation in survey of distribution and abundance has a long history and has made significant contributions. By using this tradition, eBird (`http://ebird.org`) is a citizen science project that uses the Internet to engage a global network of birders as citizen scientists reporting their observations to a centralized database [23]. In the past decade, the use of social media has enabled people to directly contribute their data over the Web. The Web makes it easy for people to share their comments or photos within an online community. Web 2.0 platforms and social media services are considered new ways to collect data for citizen science projects. Flickr, an online photo-sharing platform, can be used as a tool for collecting witness photos for citizen science projects. For example, photos from Flickr are collected to monitor the the distribution of bee [20].

## 2.2 Social Media as Sources of Geographic Information

As many social media platforms provide location-based services, these services are often considered as tools for creating and collecting geographic information [22]. However, geography is often not the subject of social media messages. Unlike Wikimapia or OpenStreetMap, social media services in general are not tools for citizens to create institutional geographic information. Nevertheless, social media messages have geographic components, which are mostly used to refer to locations where they originated [21]. Besides GPS coordinates, place names are often used as geographic references in social media contents. Formalizing places in the use of social media is considered as a key connection between crowdsourced geographic information (location-based social media) and institutional geographic information [22]. There are many interesting research works on transferring social media contents for scientific purposes. It has been argued that Flickr photos can be used for meteorological purposes [12]. For example, researchers compared the position of hail detected in the atmosphere by radar with positions of Flickr photos depicting hail on the ground. Moreover, social media is often used to report situations in disaster-affected areas. After the January 12, 2010 Haiti earthquake, an open-source crisis-mapping system, Ushahidi, has been set up to capture, organize, and share critical information from Haitians. The information was gathered through social media as well as text messages sent via mobile phones [11]. It has even been argued that tweets can be used to support crisis management [7]. Researchers analyzed publicly available tweets referring to a forest fire near Marseille, France in July 2009, and they considered citizens as potential disaster relief information providers. Web-enabled geo-visual analytics approach has been proposed to leverage tweets in support of crisis management [13].

## 2.3 Issues of Data Quality in Citizen Science Projects

The issue of data quality in crowdsourced information, especially in citizen science context, has attracted much attention. To ensure the quality of data contribution, training and educating volunteers by experts or experienced participants is a common method in citizen science [8]. 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 [17]. The growth of data quantity will be slow if the data contribution is restricted to experts or trained participants. Contrariwise, data volume often increases rapidly if data contribution is entirely open to participants (but data quality is hard to guarantee). There has been some validation of this assumption via OpenStreetMap (OSM) datasets for Volunteered Geographic Information (VGI) [10]. However, as most OSM mapping tools are so sophisticated, not everyone can use them easily. Also, the tools have many functions to assist participants in creating structured data. However in the context social media, user-generated geographic contents are normally comprised of textual data and multimedia files but not structured geographic data.

## 2.4 Reusing and Structuring Social Media Contents for Citizen Science Projects

Social media is a challenging new ground for developing knowledge discovery techniques [1]. Besides social network graph mining methods based on analyzing the links amongst

messages [18], there are studies tackling social media by using text mining methods. For example, an event notification system that monitors tweets has been proposed in which it also delivers semantically relevant tweets if they meet a user's information need [15]. To help structure contents from social media, the exploitation of external semantic resources to disambiguate contents is considered an efficient method. A folksonomy is comprised of a set of tags. Because a folksonomy is an informative resource from the crowd, it is often used to develop recommendation engines or semantic tagging tools. To enrich the semantics of folksonomies, researchers not only built up relations among tags via statistical analysis but also integrated the structured tags with the cloud of Linking Open Data (LOD) through the DBpedia [3, 4]. The effort of LOD has been a cornerstone in the realization of the Semantic Web vision [2]. The LOD is supported by a large community in the academia and the industry; many freely available technologies and tools have been developed. LOD also offers new opportunities for semantic data integration. Map4RDF is a mapping and faceted browsing tool for exploring and visualizing linked datasets enhanced with geometrical information [6]. Recently researchers are working to integrate crowdsourced information with LOD resources for disaster management [16]. This will increase the impact of crowdsourced data in disaster management, and it shall help humanitarian agencies make informed decisions.

## 3. A USE CASE ON USING SOCIAL MEDIA FOR CITIZEN SCIENCE PROJECTS

To illustrate the research issues involved in this study, we describe a use case as follows. In this use case, four characters are identified: Sarah is an ecologist who takes charge of a citizen science project. She exploits Facebook as a network to collect data for this project. James is an amateur naturalist, and Sinensis is an experienced naturalist. They both are engaged in this project by joining a Facebook interest group. Dave is an information engineer working for the project.

By using Facebook as a platform for citizen science projects, James can easily and quickly contribute his observations on the Facebook group. He need not worry about his observation posts would be rejected even though he may not be able to identify the species in the photo he posts. As shown on Figure 1, James posts an observation photo with a text message describing the location and date of his observation. But he does not provide the species name. Sinensis identifies the species in the photo, and she provides the species name and other references to James by commenting on James' post.

By exploiting the Facebook network, the number of participants and observations can rapidly increase. However, Sarah is not satisfied by just using Facebook groups. To obtain ecological observation data, it is necessary to transform a Facebook conversation thread into an ecological observation record. But the participants' observations are text messages and photos, which are unstructured information items. To manually extract species and place names from text messages is a time- and labor-intensive task. Sarah needs a tool to efficiently extract useful data from Facebook posts for ecological research such as the spatial distribution of a species. In particular, Sarah expects tools which can assist participants with standardized species names and place


**Figure 1: James' observation was posted to a Facebook group. ① is a Facebook post, which is comprised of a text message and a photo. ② is a Facebook comment. A Facebook thread is comprised of a Facebook post and several Facebook comments.**

names for use in their observation posts. Thus, mistakes and ambiguities about species and places can be reduced.

Semantic Web techniques provide information engineers new ways to meaningfully interlink data. Semantically rich data can be efficiently managed and utilized. Thus, Dave chooses to use Semantic Web techniques to meet Sarah's requirements. By exploring the Facebook posts, he identifies a gap between uncontrolled vocabulary used in flat text messages found in crowdsourced observations and conversations, and structured data items provided and required by Semantic Web resources. To deal with unstructured data, Dave considers using Natural Language Processing (NLP) tools to obtain location, date, and species information from texts.

To semantically structure the extracted information, Dave designs an ontology, and then uses Resource Description Framework (RDF) to compile the extracted information. The RDFized datasets not only can be published via Linking Open Data (LOD) principles, but also can be transferred to content management systems for managing the datasets. Moreover, the RDFized datasets and NLP tools can be used to develop a semantic annotation tool, which can be used to automatically check the text of Facebook posts and comments. If the tool detects vagueness in the description of place or species, an information box will be popped up asking data contributors to provide standardized place or species names. After they click on one of the candidate names in a pop-up box, an ID for the standardized name would be inserted into the text message. The ID points to a permanent URI referring to a formalized place or species record in RDFized datasets.

## 4. ISSUES, METHODS, AND TOOLS FOR COLLABORATIVE SPECIES IDENTIFICATION AND OCCURRENCE

This section describes the processing of two Facebook interest groups, *Reptile Road Mortality* (in Chinese, 路殺社) and *Enjoy Moths* (in Chinese, 暮光之城), which are both citizen science projects on species identification and occurrence

hosted by the Endemic Species Research Institute, Council of Agriculture, Taiwan. *Reptile Road Mortality* aims to collect reports of road-killed reptiles, and *Enjoy Moth* focuses on collecting observations of moths. Facebook makes it easy for users to contribute their observations to citizen science projects. However, transferring the participants' observations to structured data for scientific purposes is a challenge. The participants' observations are comprised of texts and photos. For privacy and security reasons, most EXIF data is stripped when photos are uploaded to Facebook. Without EXIF data, a photo from Facebook is just a graphic record; the photo cannot in itself indicate the date and location on which it was taken. The text messages accompanying the photos will be the main sources for extracting ecological information about the species in the photos.

To deal with this crowdsourced information, we propose an approach which is comprised of three steps, as shown in Figure 2. The first step is information extraction. To extract useful information from Facebook groups, we applied NLP methods and tools to identify place and species names in Facebook posts and comments. The next step is information formalization. To structure the extracted information, we used RDF to encode the extracted information. To formalize the extracted information, we used LOD to connect the information to other knowledge resources, and to publish it on the Web. The final phase is information reuse, which utilizes the structured information to further improve the quality of data collected from social media for citizen science projects. The details of our approach are described in the following.

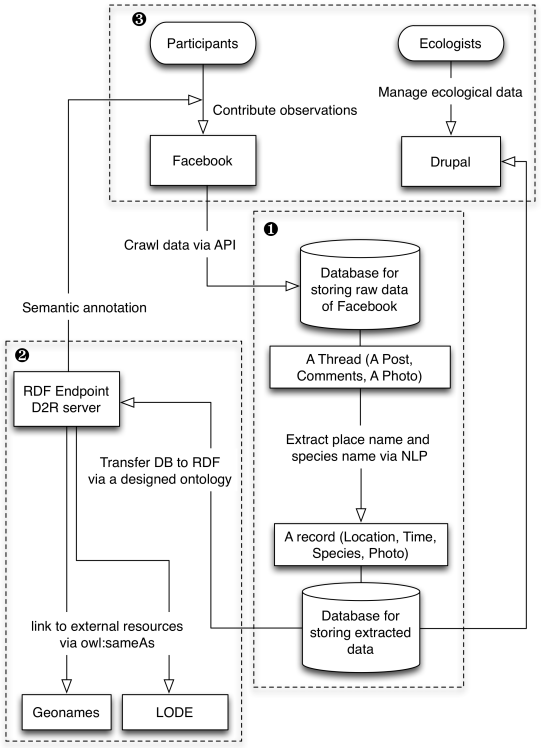



**Figure 2: Our approach for processing crowdsourced information is comprised of three steps: ❶ Information Extraction, ❷ Information Formalization, and ❸ Information Reuse.**

## 4.1 Information Extraction through Natural Language Processing

Since the participants in these two Facebook groups mainly use traditional Chinese as the communication language, Chinese Natural Language Processing is an important component in information extraction. Chinese texts are character-based, not word-based. Moreover, there is no space between characters in written Chinese sentences. This unique language feature leads to a challenge of word segmentation. To properly separate out words in sentences is an important step in Chinese NLP tasks. Using a lexicon as a resource to conduct the segmentation is simple and efficient. This approach basically compares a Chinese character string with the entries in a lexicon. If a substring in the text matches a lexicon entry, the substring is a word. That is, the quality of a word segmentation task depends on the lexicon it uses. The richer is the lexicon; the better is the word segmentation. However, most Chinese NLP tools are developed for general purposes. Their lexicons are rarely comprised of rich species and place names.

To efficiently extract species and place names from Facebook threads, it is necessary to constitute specific lexicons. We compiled a place-name lexicon from Taiwan Geographic Names database (`http://placesearch.moi.gov.tw`) and a species-name lexicon from Taiwan Catalogue of Life databases (TaiCOL) (`http://col.org.tw/`). The FudanNLP toolkit (`http://code.google.com/p/fudannlp/`) is used as it allows users to use extra lexicon resources for word segmentation tasks. Note that, however, species names and place names found in Facebook posts and comments are not always in the specific lexicons.

### *4.1.1 Extraction of Species Names*

A Chinese species name often has four or more characters. For convenience, people often select two or three characters from a full species name and use them as a shorthand. A Chinese species name often is comprised of an adjective as a prefix for describing a feature of the species, followed by a family or genus name. In casual conversations, often the adjective prefixes alone are used to indicate the species. For example, *Papilio Polytes* (玉帶鳳蝶), *Atrophaneura Horishana* (曙鳳蝶), and *Papilio Hermosanus* (琉璃紋鳳蝶) are three kinds of papilionidae butterflies. In conversations about butterflies, people may just use "玉帶" to stand for *Papilio Polytes* (玉帶鳳蝶), "曙鳳" to stand for *Atrophaneura Horishana* (曙鳳蝶), and "琉璃" stand for *Papilio Hermosanus* (琉璃紋鳳蝶). However, an adjective prefix may not uniquely specify a species. Take the prefix "細紋" (pronounced *Si-Wen*, meaning "fine veined") as an example. There are 15 species names with "細紋" as the prefix in the TaiCOL database. There is a need to disambiguate the short species names people use for full species names in Facebook threads.

We designed a process to identify species names in Facebook threads, as shown in Figure 3. If no full species name matches the substrings in the text, this process checks if any prefix of species names is used in the text. If a prefix is found, this process continues to find if a family or genus name exists in the thread. To enumerate all possible species-name prefixes, each species name forwardly generates a series of prefix terms (from $N$-gram prefix to 2-gram

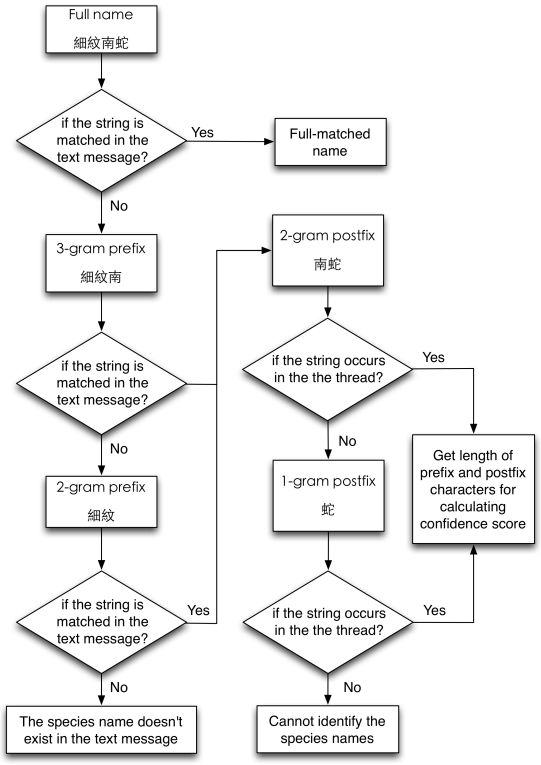



**Figure 3: Identifying shortened species names.**

prefix). To check for family or genus names, each species name backwardly generates a series of postfix terms (from 1-gram postfix to $N$-gram postfix). $N$ is the length of the species name minus one. The prefix and postfix terms are also used as part of the lexicon for word segmentation.

As shown in Figure 3, a species name "細紋南蛇" (pronounced *Si-Wen-Nan-She*, Ptyas Korros) can generate prefix terms "細蚊南"(*Si-Wen-Nan*) and "細蚊" (*Si-Wen*), as well as postfix terms "蛇" (*She*), "南蛇" (*Nan-She*), and "紋南蛇" (*Wen-Nan-She*). If the full name "細紋南蛇" does not match any substring in the text messages in a Facebook thread, the process start to find if any of the prefix terms exit in the thread. If a prefix matches some substrings in the text messages, the process will further find if the corresponding postfix term exists in the thread (for family or genus name). If a prefix term and a postfix term are both identified, the length of the two terms can be used to calculate a confidence score

$$\frac{L_{\text{prefix}} + L_{\text{postfix}}}{L_{\text{full}}}$$

where $L_{\text{full}}$ is the length of the fill species name, and $L_{\text{prefix}}$ and $L_{\text{postfix}}$ respectively the length of the prefix and postfix term found in the thread. For example, suppose the prefix term "細紋" and the postfix term "蛇" are both identified in a thread. The length of the prefix term "細紋" is 2, the length of the postfix term "蛇" is 1, and the length of the full species name "細紋南蛇" is 4. Therefore, the confidence score is $\frac{2+1}{4} = 0.75$. With this confidence score, we determine how likely we have identified a species name in a text message.

We find that species identification in Facebook groups is often a social interaction. When a participant provides a photo without species identification in the post, other participants would contribute their opinions for identifying the species. If the species in the photos are difficult to identify, there are often several different suggestions in the thread for the correct species name. This is especially the case in the road-kill interest group. Therefore, we specifically developed a method to determine the name of the species being discussed in a Facebook thread. This method considers three criteria.

1. How many *Like* does the post or the comments get? Generally speaking, giving *Like* to a post or a comment expresses positive feedback to the content of the post or the comment. The positive feedbacks often come from agreements to the content, or from connections to something that people care about. In this context, if a post or a comment containing a species name receives a relatively large number of *Like*, we can be fairly confident the species in the photo is correctly identified. In a thread, however, often a post gets more *Like* than any of its comments gets. This is even though the text content of the post is poor; people would give *Like* just because of the photo in the post. There is a need to discount the number of *Like* for a post when using the number for species identification. We set 10 *Likes* as a threshold to determine a post's weight. If the number is less than 10, the weight of the post is set to 1. If the number is more than 10, the weight is set to the (base 10) logarithm of the number. We also observe that the latest comment is more likely to provide the correct species name. However, the latest comment usually receives less *Like*. Thus we weight the numbers of *Like* in the sequence of comments. The later a comment is posted, the more weight it gets.

2. How prestigious are the people who post or make comments? The prestige value is determined by the numbers of *Like* one got from one's posts and comments. That more *Like* one got, the more prestigious one is.

3. How many times does a species name occur in a thread? The frequencies of the species names appearing in a thread are intuitive factors for determining the correct species name. The name that is mentioned most often may well be the correct name for the species.

### 4.1.2 Extraction of Geographic Names

Geographic references used in a Facebook post can be coordinates, road numbers with road miles, road names, or place names. There exist some pattern in the use of coordinates and road numbers with road miles, so these references can be extracted via regular expressions. But the extraction of road names and place names has to depend on natural language processing. For each Facebook post, we use the FudanNLP toolkit and a geographic name lexicon to extract geographic entities mentioned in the post's text. The geographic name lexicon is comprised of Taiwan Geographic Names database, as well as Taiwan road names obtained from Wikipedia. If a string within a text message matches an entry in the lexicon, the string would be identified as a geographic name. The identified geographic name, then, would be used to obtain the place's latitude, longitude, and hierarchical administrative data (e.g. the township and

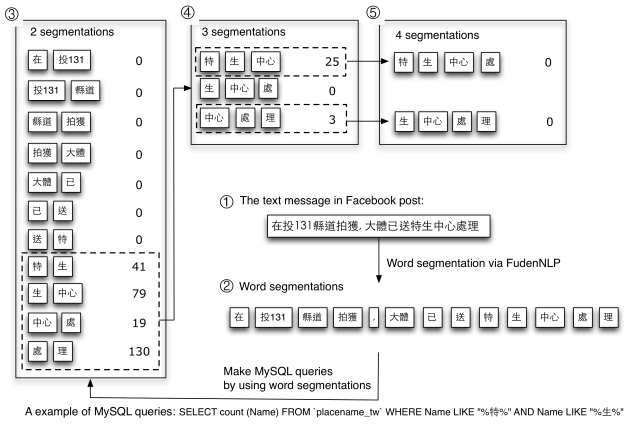



**Figure 4: Identifying shortened place names.**

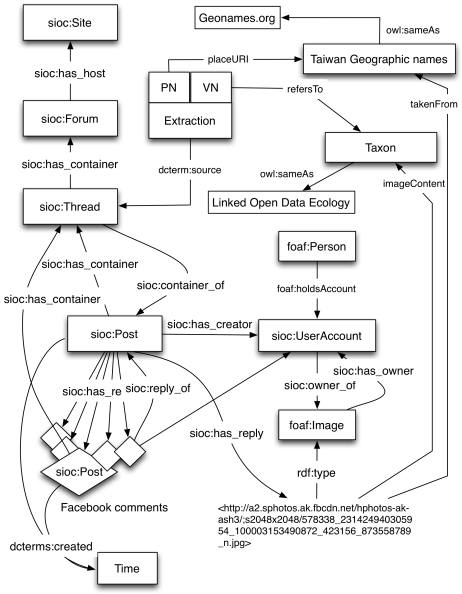



**Figure 5: An ontology for formalizing the extracted information from Facebook threads.**

county the place belongs to) from Taiwan Geographic Names database.

As in the use of species names, people often shorten a long geographic name in their conversations if it is more than 4 characters long. But the way for shortening place names is different from the way used for shortening species name. For example, "特有生物研究保育中心" (pronounced *Te-You-Sheng-Wu-Yan-Jiou-Bao-Yu-Jhong-Sin*, meaning "The Endemic Species Research Institute") is often shortened to "特生中心" (*Te-Sheng-Jhong-Sin*). The four characters are extracted from the ten-character sequence for their respective meaning; the four characters do not form a prefix nor a postfix. Therefore, we designed an approach to identifying shortened geographic names in text messages. This approach is illustrated in Figure 4.

Take the identification of "特生中心" as an example. As shown in ① of Figure 4, the text "在投131縣道拍獲, 大體已送特生中心處理" (meaning "Photo was taken at Nantou County Road No. 131; the body had been sent to the Endemic Species Research Institute for preservation") is posted with a photo of road-kill. By using FudenNLP, a word segmentation of the text is obtained as ② of Figure 4. We then use pair-wise combinations of the segments to query the geographic name lexicon to see if they exist as (subsequences of) geographic name entities. For example, the pair "特" and "生" appears, as a subsequence, 41 times in the lexicon, as shown in ③ of Figure 4. We then progressively check all 3-segment combinations and more, as shown in ④ of Figure 4, until we stop at a combination that is long and appears often as a subsequence in the lexicon. As show in ⑤ of Figure 4, in this example we stop at the 3-segment combination "特生 中心" as it appears more often than the other 3-segment combination "中心 處 理", and none of the 4-segment combinations produce a match. The 4-character sequence "特生中心" in the post is now identified as a shortened geographic name for the corresponding entries in the geographic name lexicon.

## 4.2 Information Formalization Using Semantic Web Technologies and Tools

A Facebook thread is an entity comprised of social media contents involving peoples, places, time periods, photos, and links to other contents. There exist several ontologies that can be used to construct a domain-specific ontology for Facebook threads. For example, Semantically-Interlinked Online communities (SIOC) can be used to represent the structure of Facebook posts, comments, and threads. Friend of a Friend (FOAF) can be used to describe content creators, and Dublin Core for the interlinked contents they created. We put together these ontologies to design an ontology for our purpose, as shown on Figure 5.

Having this ontology, we can transfer the extracted data, originally stored in relational databases, to Resource Description Framework (RDF) format. The transformation task can be done by D2R Server which is an open source tool for publishing relational databases as RDF links on the Web. This tool also enables in-browser exploration of the databases as collections of RDF links, and allows queries to be made to the databases using the SPARQL query language. Figure 6 illustrates the case where the Facebook thread listed in Figure 1 is published through a D2R server using the designed ontology. Figure 7 displays the extracted species name and geographic name from the Facebook thread published in Figure 6. The names are linked to Taxon and Taiwan Geographic Names respectively. They both use the vocabulary *owl:sameAs* to link to external resources expressed in other ontologies. The species names are linked to Linked Open Data Ecology (LODE), and the geographic names are linked to Geonames.org. Figure 8 illustrates a taxon of the extracted species name. Besides a link to LODE, the taxon provides a map comparing the habitats drawn from official ecological records with the occurrences drawn from the crowdsourced ecological records. In this figure, the convex boundary is drawn from official records in TaiBIF (Taiwan Biodiversity Information Facility) while the placemark is the occurrence location drawn from Facebook threads. The comparison shows that crowdsourced ecological records can be complementary resources that enrich official ecological records.

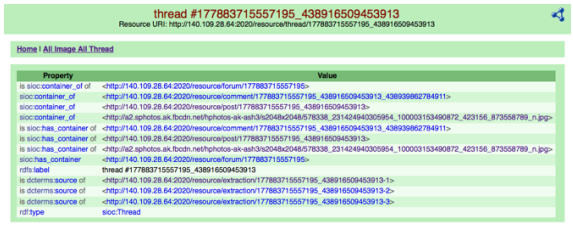
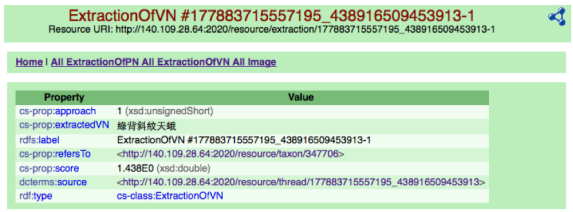



**Figure 6: The thread listed in Figure 1 is published via a D2R server using the designed ontology.**



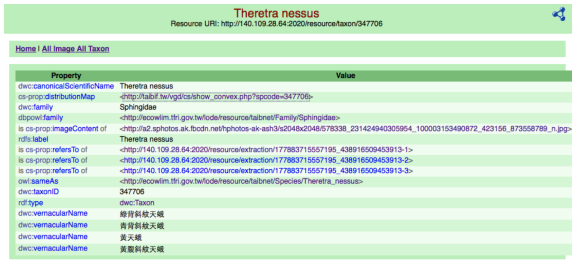

(a) The entry for the extracted species name.


(b) The entry for the extracted geographic name.

**Figure 7: Extracted from the thread in Figure 6, the specie name and the geographic name are published.**

## 4.3 From Information Reuse to Data Quality Improvement

The formalized crowdsourced information actually is a good resource for improving the generation of crowdsourced information itself. The databases of the formalized crowdsourced information can be used to help construct better content management systems (CMSs). These CMSs can provide ecologists with tools to explore ecological observations via taxonomy of species names and maps, as shown in Figure 9. These tools can also become part of user interfaces to help citizen science participants use structured data. Moreover, the use of NLP toolkits and formalized names can be used to improve the input of crowdsourced information. By using JavaScript, a semantic annotation plug-in is developed for disambiguating the use of place names and species names. As shown in Figure 10, word segmentation and name entity identification is invoked automatically when a participant starts to enter text. In this example, when the string "太平山" (pronounced *Tai-Ping-Shan*, meaning "Mt. Peace") is entered, all geographic entities with "太平山" as part of their names appear as a list in a pop-up box. As shown in Figure 10, there are at least four Mt. Peace in Taiwan, each with a unique ID. After the content creator selects one geographic entity from the list (in this case, the one with ID `tgn:92473`), this ID is inserted into the text alone with the name "太平山". This is very useful for the purpose of




(a) A taxon of *Theretra nessus* which is the extracted species name in Figure 7. This entry is connected to LODE via an owl:sameAs link.


(b) A map comparing the spatial distributions of the species drawn from, respectively, official ecological records and crowdsourced ecological records.

**Figure 8: The extracted species name is (a) linked to LODE and (b) its spatial distributions compared.**

disambiguation from now on. This is just one example for which formalized crowdsourced information can be used to assist participants with structured data.

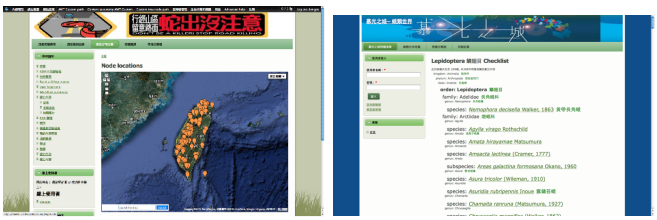
# 5. CONCLUSIONS AND FUTURE WORKS

Social media brings new opportunities to the citizen science domain. Information crowdsourced from social media is considered valuable for scientific works. This study proposed an approach to transferring unstructured crowdsourced information to structured data for scientific purposes. This approach has been successfully implemented to facilitate social-media based citizen science projects. We believe it has broader application in user-generated content management as well, and it promises to be a good start in solving important design problems in citizen science projects on the Web.

The use of NLP toolkits brought us experience in word segmentation and entity identification (at least in the context of Chinese language text threads). Crowdsourced and formalized species names and place names have been used to enrich both general and specialized lexicons, and the lexicons are used again to improve the function of word segmentation and identification tools. That is, we are forming positive feedback loops among information gathering, lexicon generation, and tool development. Currently, the efficiency of the word segmentation tool is not good enough, and there is room for improvement for the tools we have built. In the future, we will improve these tools and investigate new ways to apply them in other contexts.

# 6. REFERENCES

[1] A. Bifet and E. Frank. Sentiment knowledge discovery



(a) *Road-killed* (b) *Enjoy Moths*

**Figure 9: Two Content Management Systems for formalized crowdsourced ecological data. (a) *Road-killed* (`http://roadkilled.biota.biodiv.tw/`) (b) *Enjoy Moths* (`http://enjoymoths.biodiv.tw/`**


**Figure 10: A semantic annotation plug-in for entering geographic names in Facebook posts.**


in Twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, DS'10, pages 1–15. Springer-Verlag, 2010.

[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] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia—a crystallization point for the web of data. *Web Semantics*, 7(3):154–165, sep 2009.

[4] 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.

[5] J. P. Cohn. Citizen Science: Can Volunteers Do Real Research? *BioScience*, 58(3):192–197, 2008.

[6] A. de León, F. Wisniewki, Boris Villazón-Terrazas, and O. Corcho. Map4rdf–faceted browser for geospatial database. In *USING OPEN DATA: policy modeling, citizen empowerment, data journalism*, 2012.

[7] B. De Longueville, R. S. Smith, and G. Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 73–80, New York, NY, USA, 2009. ACM.

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

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

[10] M. M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus Law to Volunteered Geographic Information. *Cartographic Journal, The*, pages 315–322, Nov. 2010.

[11] J. Heinzelman and C. Waters. Crowdsourcing crisis information in disaster-affected Haiti. Technical Report Special Report 252, United States Institute of Peace, Washington, DC, 2010.

[12] O. Hyvärinen and E. Saltikoff. Social Media as a Source of Meteorological Observations. *Monthly Weather Review*, 138(8):3175–3184, Apr. 2010.

[13] A. MacEachren, A. Robinson, P. Jaiswal, A., A. S., Savelyev, J. Blanford, and P. Mitra. Geo-Twitter Analytics: Applications in Crisis Management. In *Proceedings, 25th International Cartographic Conference*, pages 1–8, 2011.

[14] 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.

[15] M. Okazaki and Y. Matsuo. Semantic twitter: Analyzing tweets for real-time event notification. In *Recent Trends and Developments in Social Software*, volume 6045 of *LNCS*, pages 63–74. Springer Berlin / Heidelberg, 2011.

[16] 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.

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

[18] D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 138–145, 2010.

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

[20] R. Stafford, A. G. Hart, L. Collins, C. L. Kirkhope, R. L. Williams, S. G. Rees, J. R. Lloyd, and A. E. Goodenough. Eu-social science: The role of internet social networks in the collection of bee biodiversity data. *PLoS ONE*, 5(12):e14381, 2010.

[21] A. Stefanidis, A. Crooks, and J. Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, pages 1–20, 2011.

[22] D. Sui and M. Goodchild. The convergence of GIS and social media: challenges for giscience. *International Journal of Geographical Information Science*, 25(11):1737–1748, 2011.

[23] C. Wood, B. Sullivan, M. Iliff, D. Fink, and S. Kelling. eBird: Engaging birders in science and conservation. *PLoS Biol*, 9(12):e1001220, 12 2011.