The enhancement of geographic names database with volunteered geographic information (VGI) approach

Wei-Chia Huang[1], Jinn-Guey Lay[1], Ching-Chi Huang[1], Chang-An Chen[1], Ching-Jen Kao[2]

[1]National Taiwan University (NTU),

No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan, r99228023@ntu.edu.tw

[2]Chinese Culture University (CCU),

55, Hwakang Rd., Taipei, Taiwan, cjkao@staff.pccu.edu.tw

KEYWORDS: volunteered geographic information (VGI), gazetteer, Taiwan Gazetteer

ABSTRACT: Geographic name is an important element of spatial data retrieval and is an essential component of the national geographic information system. Due to its diverse geography, history, and culture, Taiwan has accumulated a large number of geographic names throughout history. The Taiwan Gazetteer is a geographic name database maintained by the Ministry of Interior (MOI) of Taiwan that contains over 170,000 records. This database has been widely used for government administration, cultural research, and K-12 education purposes to this day.

However, this database was developed by different organizations, each with varying data sources. Hence the contents and formats of the Taiwan Gazetteer require more rigorous review. This study will analyze the attributes of the database and propose a comprehensive framework that can deal with the diversity of Taiwan's geographic names.

Moreover, in order to collect more information and data regarding geographic names from the general public in the future, this research explores the feasibility of implementing a Web GIS that gathers geographic names as volunteered geographic information (VGI). Through this research we hope to enhance the quality and applications of geographic name databases.

1. RESEARCH MOTIVE AND PURPOSE

Geographic name is an important element of spatial data retrieval and is an essential component of the national geographic information system. Due to its diverse geography, history, and culture, Taiwan has accumulated a large number of geographic names throughout history. Previous research published in the past century on this topic is compiled and outlined in the 1990 publication “Taiwan Place Names Dictionary” (臺灣地名辭典).

With the widespread popularity of the Internet, this dictionary has been combined with other sources to form the “ Taiwan Gazetteer” (地名資料庫)—a geographic name database maintained by the Ministry of Interior (MOI) of Taiwan that contains over 170,000 records. This database has been widely used for government administration, cultural research, and K-12 education purposes to this day.

However, the database is largely unstructured today primarily due to two reasons: the switch from analog to digital technology in recent decades as well as the database’s development by different organizations, each with varying data sources. This unstructured format from the original literature does not satisfy the requirements of a structured database. The contents also require more rigorous review from researchers, educators and the public.

This study will analyze the attributes of the Taiwan Gazetteer and propose a comprehensive framework that can deal with the diversity of Taiwan’s geographic names. Moreover, in order to collect more information and data regarding geographic names from the general public in the future, this research explores the feasibility of implementing a Web GIS that gathers geographic names as volunteered geographic information (VGI). Through this research we hope to enhance the quality and applications of geographic name databases.

2. CURRENT SITUATION AND PROBLEMS FACING TAIWAN GAZETTEER

2-1. Current situation

Taiwan Gazetteer is now designated as a project under the Geographic Name Information Service (GNIS, 地名資訊服務網) (http://gn.moi.gov.tw). It contains 156,846 geographic name records as of May 2016. For ordinary users, the database provides geographic names and their details, including: categories, location, pinyin in Mandarin Chinese, abbreviations and nicknames, as well as history. As an example, Figure 1 shows all of the fields in the records for Tamsui.

Attributions	Fields
地名類別 Categories	Settlement
地名名稱 Geographic Identifier	淡水

地名別稱 Alternative Geographic Identifier	滬尾
地名譯寫 English Name	Tamsui
漢語拼音 Hanyu Pinyin	Danshui
通用拼音 Tongyong Pinyin	Danshuei
標準地名 Standardized	No
所屬村里 Location Village	Chang-geng
所屬鄉鎮市區 Location District	Tamsui
所屬縣市 Location City	New Taipei City
地名意義 Meaning	
相關位置與面積描述 Description of Location and Area	今指淡水鎮全鎮，面積 70.6565 平方公里。北到屯山里大片頭，南到竹圍關渡埔頂，西北臨臺灣海峽。
地名沿革與文獻歷史簡述 History	「淡水」地名一詞最初出現於明季嘉靖年間，係漢人見河口狀態所取的地名。1.明嘉靖年鄭成功著書《日本一鑑》記載：「夫小東（指臺灣北部）之域，有雞籠之山（指大屯山系），山乃石峰特高於眾，中有淡水出焉（指淡水河流出）。2.嘉靖末年（大約 1565 年）顧炎武撰《天下郡國利病書》載有：「今琉球告急，屬國為俘，而沿海姦民揚帆無忌，萬一倭奴竊據，窺及雞籠淡水，此輩或從而勾引門庭之寇，可不為大憂乎」。「滬尾」早期亦寫作「戶尾」（見媽祖廟內「望高樓碑」），或「虎尾」（見黃叔璥撰《台海使搓錄》）。《台灣府志》記載：「以碎石築海坪
地名相關事項訪談內容 Content of Interview	滬尾與淡水原指不同區域：淡水最初指整個台灣北部，後指今淡水河流域和台北一帶，日治時代指的是今淡水、三芝、石門、八里四鄉鎮，淡水街(滬尾街) 指的才是今日的淡水鎮。名詞互用於清同治年，滬尾名詞消失於日治大正年。
普查使用之地圖與文獻 Source Reference	滬尾街第一期、金色淡水第 17 期-周明德文
受訪者 Respondents	
相關照片 Photo file URL	
相關錄音檔 Audio file URL	
文獻調查人員 Investigators	

文獻調查時間 Time of Investigation	
田調訪問人員 Field interviewers	
田調訪問時間 Tim e of Field Interview	
WGS84 經緯度坐標 Coordinates	
語言別 Language	
命名族群 Ethnic Group	
地名年代時間 Time of Geographic Name	明嘉靖年~迄今

Figure 1. A sample of the records of a geographic name from Taiwan Gazetteer

2-2 Current Problems

To create a database, the data structure should be arranged in order of priority. However, this database has been developed by different organizations over time, each with varying data sources and conceptions. Those organizations created some fields (as shown in Figure 1), although their expectations were far from reality. Today, most of the fields remain blank due to limited research.

2-2-1 Coordinates

Since historians also conduct research on geographic names in Taiwan, the Taiwan Gazetteer contains detailed information about the history of every geographic name record, but does not have very basic fields, such as coordinates. Presently, less than 30% of records have accurate coordinates in the WGS84 datum coordinate system, although all of the records have village-level (the most basic administrative division after county) data.

2-2-2 Multilingual Support

As a multiethnic country, Taiwan has four notable ethnic groups: Hoklo (or Hokkien), Hakka, Mainlander and Indigenous. The majority are Hoklo (70%) and Hakka (15%), who are descendants of early Han Chinese immigrants from the 17th to 19th century.

Prior to the arrival of Han Chinese settlers, indigenous people resided on the island of Taiwan. Together, the indigenous people of Taiwan speak 28 different Austronesian languages while Hoklo and Hakka people speak their own Chinese dialects. Over the past three centuries, most of the geographic names in Taiwan were mixed, reformed and coined by the Han Chinese and indigenous groups. According to an act passed in 2000, public transportation broadcasts must be made in three languages: Mandarin (Standard Chinese), Hoklo and Hakka. In addition, the Dutch (1624-1662), Spanish (1626-1642) and Japanese (1985-1945) occupations of Taiwan each introduced new forms of geographic names. After World War II, a large number of immigrants from Mainland China moved to Taiwan, accounting for 13% of Taiwan’s population today.

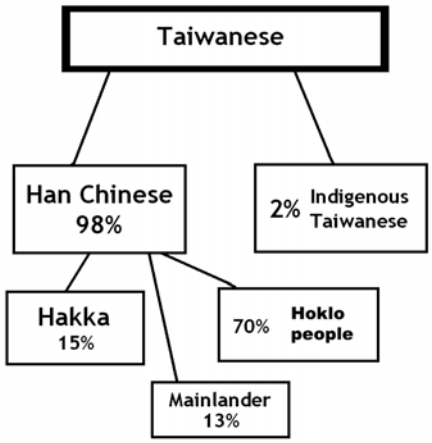

Figure 2. Breakdown of the origins of Taiwanese people today (Source: Wikipedia)

However, the Taiwan Gazetteer has limited multilingual support. For example, Danshui, a historical harbor of Taipei, is officially spelled as “Tamsui” based on the Hoklo pronunciation “Tam-tsui.” In this database, the pronunciations of native speakers have been recorded and converted to wav format. Pronunciations include Mandarin and one of the local languages: Hoklo, Hakka or an indigenous language. In the future, Romanization of geographic names should be introduced in the records as well.

2-2-3 Unclear titles of records

In the category of public facilities, many records unfortunately share the same title. For example, “gongyuan“ (公園, “park”) has been used by 1,290 records. Users may encounter difficulties when using the database as a result of the ambiguity.

Figure 3. Repeated “gongyuan” in search results

Another similar case appears in the category of roads. For example, the “Formosa Highway” (福爾摩沙高速公路) is listed as the name of 121 different records. They each share the same title, but are categorized according to the different administrative districts in which they are located.

2-2-4 Not enough Categories

Taiwan Gazetteer only has five categories; hence thousands of records share categories. For users who need more specific data (e.g. distributions of Buddhist temples), this limited selection of categories can be of little help.

Categories	Number of Records	Remarks
Settlements	43,954	
Administrative districts	8,108	
Natural entities	9,495	Mountains, rivers, islands ….
Public facilities	50,109	Markets, schools, hospitals, parking lots …
Roads	45,947	Highways, expressways, roads, streets …

Figure 4. Five categories and their contents

2-2-5 Incompletion of a specific category

Although there are 226 railway stations in Taiwan, only few have records in the Taiwan Gazetteer. We also find the similar situation with public schools, hospitals, government offices, police stations, and information centers.

3. REVIEWS- GAZETTERS IN OTHER COUNTRIES

In this paper, we will review gazetteers in four other countries – China (民政区划地名公共服务系统), Japan (国土地理院電子国土ウェブ), South Korea (National Geographic Information Institute, NGII) and United States (U.S. Board on Geographic Names, BGN) and discuss any lessons that we can learn from them.

3-1 Categories

Compared to the Taiwan Gazetteer, the counterpart databases of China and the U.S. have a diverse selection of categories. For example, China has 22 categories and 75 subcategories. Some categories seem match to service of civil administration instead of geographic names. For instance, there are marriage registration centers, community service centers, nursing homes and even lottery stations.

U.S. BGN has 65 categories with clear definitions. Figure 5 shows the table for select categories.

Populated Place	Place or area with clustered or scattered buildings and a permanent human population (city, settlement, town, village). A populated place is usually not incorporated and by definition has no legal boundaries. However, a populated place may have a corresponding "civil" record, the legal boundaries of which may or may not coincide with the perceived populated place. Distinct from Census and Civil classes.
Island	Area of dry or relatively dry land surrounded by water or low wetland (archipelago, atoll, cay, hammock, hummock, isla, isle, key, moku, rock).
Unknown	This class is assigned to legacy data only. It will not be assigned to new or edited records.

Figure 5. Table of categories for select geographic names

In addition, the NGII of South Korea has only 3 categories – Administrative, National and Cultural – and 17 subcategories. Japan’s gazetteer resembles an electric map like Google Maps and does not offer categories or other fields of geographic names.

3-2 The support of VGI

VGI-like systems, which allow users to edit existing geographic names, are available in the gazetteers of South Korea and United States. Users can offer comments on a form on the website that will be sent

to the gazetteer's administrators. Administrators can subsequently examine the comments and make any appropriate edits to the gazetteer.

BGN even allows users to propose a name for any unnamed place. However, if a user wishes to name a place after a person, the person must have been deceased for at least five years.

4. PROPOSALS FOR VGI INTRODUCING AND MAINTAINING THE GAZETTEER

4-1 Open data

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data.

MOI has released all data of Taiwan Gazetteer in a csv format. As a 3-star open data project, Taiwan Gazetteer does not require users to use a specific software in order to open and edit the content. A simple text editor will suffice.

In order to perform more sophisticated functions, an Internet service is deemed necessary. The key idea of 4-star open data project is to use URIs (uniform resource identifiers) for distribution data in which every piece of data has its own unique web address.

The Taiwan Gazetteer is now a part of GNIS. Users must gain access via the GNIS website. However, developing a 4-star open data project may take a lot of time, hence we suggest goals for the short- and long-term.

As a short-term goal, every record must have its unique URI. Due to the diversity of geographic names in light of geography, history, and culture, a URI can be highly useful for citation and education purposes. As a long-term goal, all requirements of a 4-star open data project must be satisfied, i.e. every field must have their own URIs.

4-2 VGI

Once the unique pages of every record have been created, the introduction of VGI will be possible. In this project, we define VGI as feedbacks from ordinary users about the Taiwan Gazetteer.

We propose two possible approaches to support VGI. First, the feedback mechanism can be "attached"

to the end of the page. Users who already registered and activated their accounts will be able to submit comment. We expect that users can contribute especially to coordinate data and multilingual information. Furthermore, users can also comment about any unclear or ambiguous titles. The MOI will review those comments and edits periodically, then make any corresponding updates following an examination by professional researchers.

This approach to supporting VGI has been used in other systems maintained by MOI -- for example, the Government Data Open Platform (data.gov.tw), which is where the Taiwan Gazetteer is released in CSV format. Users can login through Google, Facebook, GitHub accounts or a state ID to provide any comments. Figure 5 shows how this process works for data.gov.tw.

Figure 5. Taiwan Gazetteer released as CSV format (data.gov.tw/node/7063)

The introduction of Wiki is an alternative approach that affords more flexibility. We can clone all of the existing data from Taiwan Gazetteer to a wiki-based site where users can read, modify existing pages or create new pages, just like in Wikipedia. We implemented a website on Wikia, a commercial Wiki-based website founded by Wikipedia founder Jimmy Wales, as an example.

Figure6. Taiwan Gazetteer as shown in Wikia

However, this approach may face some copyright issues. As mentioned before, Taiwan Gazetteer is the digital version of the “Taiwan Place Names Dictionary” mixed with other sources that may not be public domain. According the copyright laws of Taiwan, publications shall be made public fifty years after the author’s death. For now, it seems practically impossible to do a wiki-based website.

4-3 Improvement of categories

Upon reviewing the gazetteers of four other countries, it is apparent that Taiwan Gazetteer must be improved on many levels. U.S. BGN has the clearest and most number of categories, and serves as a good model for improving our project in the future.

Our team has submitted suggestions for categories with clearer definitions. We propose no changes to the five existing main categories, but hope to supplement these categories with subcategories. For example, we can add hospitals as a subcategory of public facilities. The examination by MOI is currently in process and new categories and subcategories will be implemented next year.

5. DISCUSSION

In order to collect more geographic names from the general public, we explores the feasibility of implementing a Web GIS that gathers geographic names as volunteered geographic information (VGI). However, since many researcher papers have not been released to the public, a feedback mechanism attached to the end of the page is better way to allow users to participate in the VGI project.

We also realize that an open-data page is necessary for building a VGI. The information delivery of geographic names can be two-way once every page has its own URIs and feedback comments.

REFERENCES

1. Hausenbals M., 2010. 5-star Open Data. Retrieved Sep 20, 2016 from 5stardata.info