THE MAKING OF STATISTICAL MAPS:
A PERSPECTIVE FROM THE ONTOLOGY OF POLITICAL DISTRICTS

Kai-Ling Huang[1] and Jung-Hong Hong[2]
[1]Master student, Department of Geomatics, National Cheng Kung University,
No.1, University Rd., Tainan 70101, Taiwan; Tel: +886-6-2370876#851; Fax: +886-6-2375764
E-mail: kailing112@hotmail.com

[2]Associate professor, Department of Geomatics, National Cheng Kung University,
No.1, University Rd., Tainan 70101, Taiwan; Tel: +886-6-2370876#851; Fax: +886-6-2375764
E-mail:junghong@mail.ncku.edu.tw

KEY WORDS: Ontology, Political Districts, Statistical Map

ABSTRACT: Statistical information collected by professional domains serves as a valuable reference to understand our world. Often stored in table formats, such data lacks a direct geospatial reference to make it illustratable on a map. In the past, such a geo-linking process is often completed by geospatial professionals via common geographic identifiers between the two datasets. Without such an in-depth knowledge and skills, inexperienced users may find it difficult to correctly pinpoint their data on the maps. From an ontology perspective, we argued that these required geo-links can be automatically established as long as the knowledge about political district and its links with users' statistical data can be "correctly" formalized. Following the concept of ontology, we first analyze how political district data is generated, managed and updated, and then formalize our findings for the design of political district ontology. The ontology enables us to unambiguously describe each political districts and their hierarchies or neighboring relations in the OWL format. The proposed ontology should be established and maintained by the authorized organizations to ensure its unique identification and correctness. The statistical map making mechanism has built-in knowledge to intelligently determine if the generated data can be used for mapping or even being used for producing other types of statistical maps. This collaborative framework is developed upon a number of services respectively managed by different organizations, which not only demonstrate its potential in the future SOA environment, but also perfectly reflect a better division of works for the current governments.

1. INTRODUCTION

To better understanding how our world changes, a wide spectrum of information has been collected by various professional domains over the years. Although the majority of collected data has spatial meaning, many of them is recorded by text only. Table has been the major format for domain experts to record statistical information. With its row-column structure, experts can easily extract a subset of table elements according to their tasks and transform the data into statistical charts to emphasize the pattern for the phenomena of interests. However, users may never notice the patterns of spatial correlation hidden in the data by only browsing the table-based data. For example, unless users are familiar with the location of every political district, it is difficult to imagine the spatial distribution of population with only tables-based census data. By making the tables-based data mappable, the secrets hidden in the textual data can be emphasized from a map perspective. Statistical maps are hence a common tool widely used by a variety of domains to aid the visual inspection and analysis about the spatial patterns of collected data. Despite that the use of statistical maps has been a common practice and most people were familiar with its layout since their earlier childhood, the correct making of statistical maps is not as easy as it appears to be. In the past, statistical maps were mainly produced by well-trained geospatial professionals who had both the knowledge about the data and the skill of using desktop GIS software. The progress of internet-based technology made it possible for users to request data located at remote servers. Many agencies nowadays offer their statistical maps via internet by WebGIS technology. For example, Eurostat is an on-line statistical information service from the European Union (http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home). Users are prompted with an interface for selecting cartographic data, type of map and symbols for making the statistical map they want. The Geolinked Data Access Service (GDAS) from OGC allows users to retrieve a set of geolinked data with geographic identifiers (OGC, 2004a), and Geolinking Service (GLS) can establish the link between attribute data from GDAS and its corresponding geospatial data by common geographic identifiers (OGC, 2004b). Although users can directly browse the statistical maps even if they do not own the data and the GIS software, the maps available are nonetheless often restricted to only the data already stored in suppliers' archives. Users are offered with limited flexibility to select and adjust the statistical maps they need, let alone to generate their own maps (Shih-Yu Lin, 2005). This is not very surprise because the making of statistical maps requires not only the acquisition of correct GIS-ready data, but also the knowledge for making correct cartographic decisions. If users' data cannot be rigorously and precisely controlled, the development of mapping systems may become rather complex. The majority of current statistical mapping systems therefore only "publish" statistical maps they have, and do not help users to design maps with their own data. We argued it is necessary to develop a mapping mechanism that can enable domain users to “participate” in the making of statistical maps as well as ensure the correctness of the final outcomes.

To bridge the technological gaps for making correct statistical maps upon users’ data, we proposed an intelligent statistical map making mechanism based on the chaining of web services offered by different organizations. It is necessary to establish an explicit geo-linking between the statistical data and the spatial description of political districts. One of the major challenge is these two datasets are collected from different and independent resources. The second challenge is the developed mechanism must be intelligently enough to determine if these two datasets can be correctly geo-linked. The third challenge is if the mechanism can automatically suggest other types of statistical maps it can produce based on the given data. The key to these three challenges is clearly the development of formalized knowledge for the political districts. To be easily and widely available to all application systems, we argued the knowledge for the political districts must be further standardized, such that the mapping mechanism can consistently acquire and parse the necessary information. Furthermore, it should be developed by its authorized organizations to serve as a reliable resource for all applications. To meet these demands, an “ontology” approach is used in this paper for building the knowledge of political districts.

The merits of the proposed approach are to successfully subdivide and reallocate the loading for making statistical map to its best responsible organizations and establish their logical chaining relationships with open web services. While the whole process may be performed by services located at different places, the developed mechanism logically and precisely links these services together to ensure the quality of the generated statistical maps. Furthermore, the proposed approach allows users to produce statistical maps of their own data with little prior mapping knowledge. The remaining of this paper is organized as follows. Section 2 discusses the collaborative framework for making statistical maps. Section 3 discusses the ontology concept and proposed ontology for political districts. Section 4 demonstrates the use of ontology for making statistical maps. Finally, section 5 concludes our major findings and suggests future works.

2. THE FRAMEWORK OF STATISTICAL MAPPING MECHANISM

To meet the great diversity of demands from domain users during the making of statistical maps, the developed mechanism must be intelligent enough to cope with data collected from different resources and make necessary decisions to ensure the quality of the products. A three-tier framework is proposed in this paper (Figure 1). The user tier interacts with the application tier by submitting their data and issues their request. With built-in cartographic knowledge, the application tier is responsible for processing the data, establishing the geo-linking relationship and requesting the making of statistical maps by interacting with respective web services provided by different authorized organizations. Compared with traditional two-tier framework, the required domain-specific data or functions of the proposed framework are managed, maintained and updated by its respective authorized organizations via web service, so as to ensure the requested data is always reliable and correct. For example, the different levels of political district in Taiwan are administered by the Ministry of Interior and the City/County governments. This increases the difficulty for GIS users to continuously track changes made to the name or spatial extent of political districts. The development of web services for political districts helps users or applications systems to acquire correct data when necessary. Figure 1 shows the major workflow of the proposed systems. Users request the making of statistical maps by submitting their EXCEL-based data and the application sends the data to the processing service to extract all necessary information upon users’ request. The extracted data is sent back to the application via XML-format data according to the developed applications schema of statistical data. The application analyzes the mapping area by interpreting the returned XML data and sends requests to the web service of geographic identifier reference system. The web service of geographic identifier reference system returns the necessary information (spatial data) about the political districts in XML (OWL) format. The application tier then analyze if the geo-linking between the two datasets can be established and if the data can be used for producing requested maps. After a map request is issued to the statistical map service, it responds with the requested map and the map is further forwarded to the users.

To take advantages of the service technology, the transferred data is encoded in open XML format following pre-designed schema to enable transparent parsing and interpretation at the application tier. Following the above workflow, all domain users need to worry about is if their own data contain the necessary statistical information and if they can correctly specify their mapping needs. The past technical demands for making statistical maps are now met by the built-in cartographic knowledge and the web services residing behind the interface of the application systems. Among all the necessary actions in the workflow, common geographic identifiers serve as the basis for establishing the required geo-linking between users’ data and political district data. The returned information must be complete and unambiguous, such that the application tier can determine if the geo-linking relationship can be successfully established. Because the changes of the reference system of political districts are administered by the governments, the authorized organizations must voluntarily publish all the versions of their administered political districts via web services. To ensure the returned information can meet the demands for cartographic design, the knowledge about political districts must be formalized beforehand. Section 3 will further explore the details for developing the ontology of the political districts.

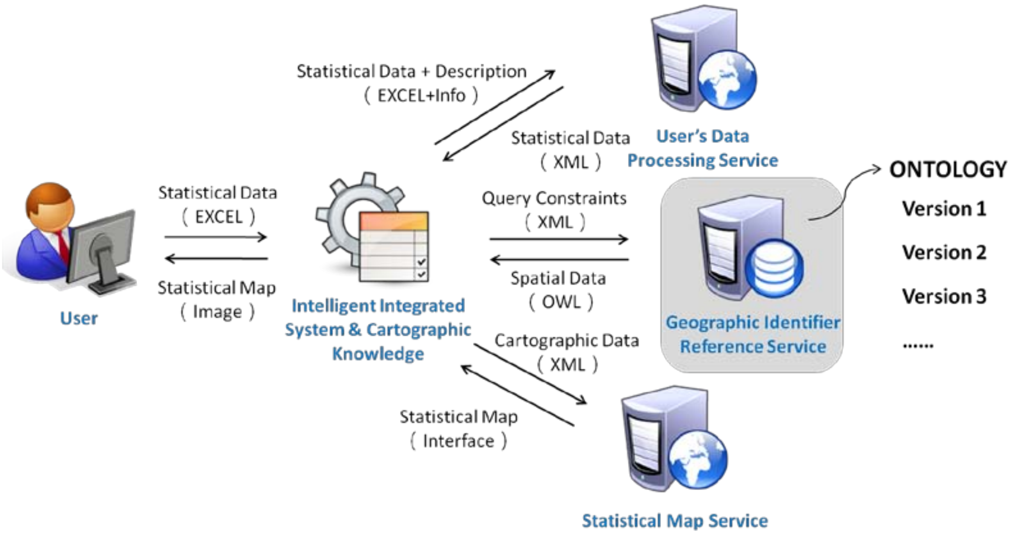


Figure 1 The workflow of service-based framework for making statistical maps

3. THE ONTOLOGY OF POLITICAL DISTRICTS

Spatial description is an essential element in the map making process to ensure users' statistical data can be illlustrated at the correct positions in the map. Based on a given geographic identifier, the major purpose of the geographic identifier reference service is to provide an open interface for applications to acquire the necessary information about the geographic identifier and its coordinate-based geometric representation, so as to serve as the foundation for establishing the required geo-linking relationships. In addition to the geometric representations, the system of political district is often designed following a hierarchical organization, where an upper level of political district is subdivided into a number of lower-level political districts. Not only a spatial partition relationship exist among the spatial extent of these two levels of political districts, their names and identifiers are also designed accordingly, where the identifier of the lower-level political district will include the identifier of the upper-level to denote their administered relationship. Whenever a change is made to the political districts, the geo-linking relationship must be changed accordingly, the correct management of historical information therefore must also be considered. To allow developed applications to establish the necessary understanding towards the acquired data, we argued all the knowledge about political districts must be formalized to meet the demands for making cartographic decisions. Ontology can transform normalized knowledge of specific domain, task, and application into the message that computers can understand to improve data sharing and interoperability (J. Lacasta et al., 2007). By doing so, the ontology of political districts can supply not only the identifier and the spatial extent of each political district, but also the relationships between them with defined vocabularies. This allows the application to make such cartographic decisions as if a requested map can be made, if the mapped area is across a number of political districts administered by different governments or even if the data can be used to make other types of statistical maps. Conceptually speaking, it is important for the design of ontology to take the essential properties of specified application into consideration. To aid the making of cartographic decision, the following four aspects must be considered during the design of ontology for political districts: political districts: identification, spatial description, valid time and hierarchical relationship.

(1) Identification: To distinguish an individual political district, a unique geographic identifier must be assigned (ISO, 2003). Common geographic identifiers are the foundation for the geo-linking process. Such a geographic identifier can be a text-string, serial number or fixed-format of codes with a specific meaning. To avoid possible conflicts, the geographic identifiers of the political districts are given by their authorized government organizations following a top-down approach. This not only ensures the unique characteristic of the political districts at the same level, but also maintains the integrity of hierarchical relationship between different levels of political districts. The collection of geographic identifiers administered by a particular organization can thus be regarded as a reference system for geographic identifiers. The success of geo-linking operation thus requires the comparison of both the geographic identifier and its reference system. As mentioned earlier, the hierarchical relationships between different levels of political districts are also included in the design of geographic identifiers. The best practice is to extend the code whenever a lower-level of political districts are created.

(2) Spatial description: Spatial description is a necessary component for any mapping applications. Every political district must be associated with a description of its spatial extent. This geometric representation may be abstracted as a point or a surface-type object. The choice of spatial dimensionality will determine what types of maps can be made from the given spatial description. Theoretically every political district must have a unique geographic identifier and a spatial description. It is however possible that the spatial extent of a political district change while its geographic identifier remains unchanged. Under such circumstance, a new version of political district data following the proposed ontology must be established to avoid conflicts.

(3) Valid time: Because the systems of political districts may change with governments' policy, it is important to make sure the geo-linking relationship is always established upon a correct version of political districts. Whenever there is a change to the reference system, a new version of data should be created. All versions of the reference systems must be carefully managed. Because such changes can only be made by the authorized organizations, the valid time of a reference system may be denoted by its last modified time and the newly modified time. During this specified period of time, the geographic identifiers and spatial description of the political districts remain unchanged. The valid time of each version of political districts therefore must be carefully recorded and managed by their authorized organizations.

(4) Hierarchical relationship: Two important types of relationships for the political districts are the hierarchical and the neighboring relations. The hierarchical relationships lead to two different spatial relationships: "composed of" and "part of". An upper-level of political district is composed of a number of lower-level of political districts and a lower-level political district is part of its immediate upper-level political district. When the hierarchies are not limited to two, reasoning rules can be derived to infer the composition and part_of relationships between political districts. The availability of composition relationship is very important when evaluating if all the data of the same levels of political districts within a political district is available. Such information helps us to determine if we can further produce statistical maps at the upper-level of political district. The hierarchical relationship may need to change accordingly when the geographic identifiers or spatial extent change.

Instead of only supplying geometric representations, we argued that this service must effectively encode the provided data from an ontology perspective, such that the statistical map service can establish an unambiguous understanding about acquired data and correctly make necessary cartographic decisions with its built-in knowledge. The proposed ontology of political districts in this paper subdivides the required information into three major sections, namely, metadata, political district and instance (Figure 2).

(1) Metadata: As discussed above, a new version of political district data must be established according to the changes of political districts. Metadata describes the basic information for each individual version of political district data, mainly used for identifying the correct version of geo-referenced data. The Dublin Core Metadata Element Set (Dublin Core, DC) is a metadata standard widely used for describing network resources, government publications and reservation of museum and library (DCMI, 2010). We selected 7 items from Dublin Core and added 2 more items, "CoorSystem" and "VersionDifference", to form the set of metadata elements in this research. The two elements of "Date-Available" and "Date-Valid" respectively denotes the date of the beginning and end for an individual version of political district data. As the political districts in Taiwan are administered by two different levels of governments, the elements of "type" is designed to denote this difference. Finally, the element of "VersionDifference" is specially designed to record the changes when compared with the previous version. This information is helpful for users to quickly obtain the difference between two versions of political districts. Because changes to the spatial extent of political districts between two versions of data are negative to the making of statistical maps, the explicitly recording of this information is necessary.

(2) PoliticalDistrict: The class of "PoliticalDistrict" is used to define a political district in the ontology. Both the level of political districts and its hierarchical relationships with other levels of political districts must be defined. This hierarchical structure is adaptable to political districts in different countries. The parent-children relationship helps users to quickly understand the composing and part_of relationships in the hierarchy. The class of "PoliticalDistrict" contains four elements: "LevelID", "LevelName", "Identifier" and "Coordinate". The attribute of "LevelID" and "LevelName" records each level number and name of the structure of political districts. Depending on the hierarchical structure, every political district in Taiwan is given a unique code. The attribute of "Identifier" records the corresponding geographic identifier and will be used to serve as the basis for geo-linking, and "Coordinate" is used to record the corresponding geometric representation.

(3) Instance: Instances are the encoding results of political district data according to the proposed ontology. For example, both "Tainan City" and "East district, Tainan City" are the instances of the class "PoliticalDistrict". Every instance contains four elements that are pre-defined in "PoliticalDistrict". The political districts in Taiwan are subdivided into four different levels, so the domain value of "LevelID" is restricted to a number between 1 and 4.

The attribute of “LevelName” records the name of the four levels of political districts: Province-City/County -Township/District-Village. The hierarchical relationships between different levels of political districts are recorded with parent-children relationships.

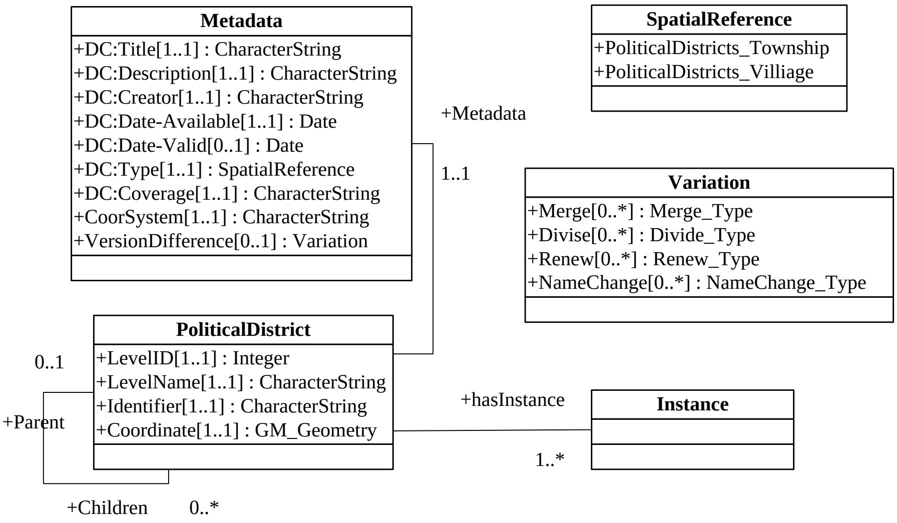
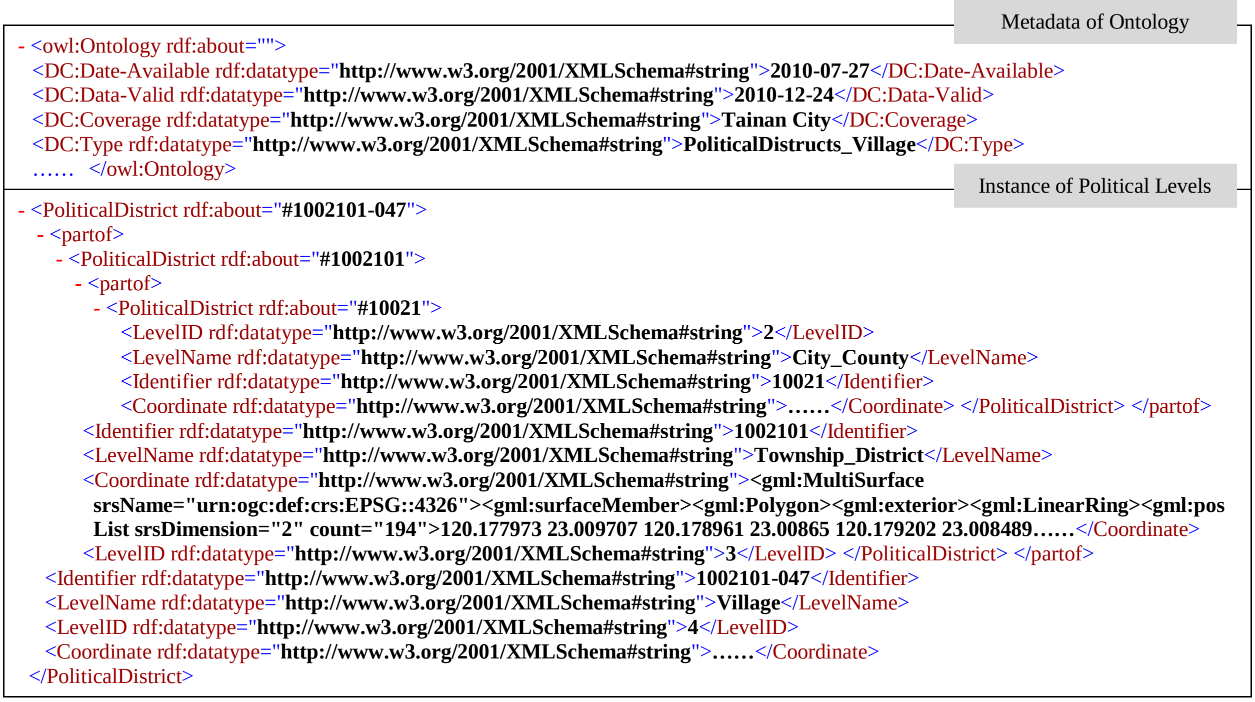

Metadata
+DC:Title[1..1] : CharacterString
+DC:Description[1..1] : CharacterString
+DC:Creator[1..1] : CharacterString
+DC:Date-Available[1..1] : Date
+DC:Date-Valid[0..1] : Date
+DC:Type[1..1] : SpatialReference
+DC:Coverage[1..1] : CharacterString
+CoorSystem[1..1] : CharacterString
+VersionDifference[0..1] : Variation
SpatialReference
+PoliticalDistricts_Township
+PoliticalDistricts_Villiage
+Metadata
1..1
Variation
+Merge[0..*] : Merge_Type
+Divise[0..*] : Divide_Type
+Renew[0..*] : Renew_Type
+NameChange[0..*] : NameChange_Type
PoliticalDistrict
+LevelID[1..1] : Integer
+LevelName[1..1] : CharacterString
+Identifier[1..1] : CharacterString
+Coordinate[1..1] : GM_Geometry
0..1
+Parent
+hasInstance
Instance
1..*
+Children
0..*

Figure 2 Structure of the ontology of political districts

To facilitate a better interoperability between services, the responded data content in this paper is encoded with Web Ontology Language (OWL) (Frank van Harmelen et al., 2004). We established different versions of political ontology by protégé software (Figure 4(a)). Figure 3 shows portions of the political districts of Tainan city from 2010-07-27 to 2010-12-24. As the data describes the status of political districts at a particular period of time, the developed applications can determine if this is the correct version of political district data for further action. In addition to the geographic identifier and spatial description, note that the encoded information also offers information about the part_of relationship (ID 1002101 is part of ID10021).


Metadata of Ontology

```
- <owl:Ontology rdf:about="">
  <DC:Date-Available rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2010-07-27</DC:Date-Available>
  <DC:Data-Valid rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2010-12-24</DC:Data-Valid>
  <DC:Coverage rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Tainan City</DC:Coverage>
  <DC:Type rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PoliticalDistructs_Village</DC:Type>
  ……  </owl:Ontology>
```

Instance of Political Levels

```
- <PoliticalDistrict rdf:about="#1002101-047">
  - <partof>
    - <PoliticalDistrict rdf:about="#1002101">
      - <partof>
        - <PoliticalDistrict rdf:about="#10021">
            <LevelID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">2</LevelID>
            <LevelName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">City_County</LevelName>
            <Identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">10021</Identifier>
            <Coordinate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">……</Coordinate> </PoliticalDistrict> </partof>
        <Identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">1002101</Identifier>
        <LevelName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Township_District</LevelName>
        <Coordinate rdf:datatype="http://www.w3.org/2001/XMLSchema#string"><gml:MultiSurface
          srsName="urn:ogc:def:crs:EPSG::4326"><gml:surfaceMember><gml:Polygon><gml:exterior><gml:LinearRing><gml:pos
          List srsDimension="2" count="194">120.177973 23.009707 120.178961 23.00865 120.179202 23.008489……</Coordinate>
        <LevelID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">3</LevelID> </PoliticalDistrict> </partof>
    <Identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">1002101-047</Identifier>
    <LevelName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Village</LevelName>
    <LevelID rdf:datatype="http://www.w3.org/2001/XMLSchema#string">4</LevelID>
    <Coordinate rdf:datatype="http://www.w3.org/2001/XMLSchema#string">……</Coordinate>
  </PoliticalDistrict>
```

Figure 3 OWL file of the ontology

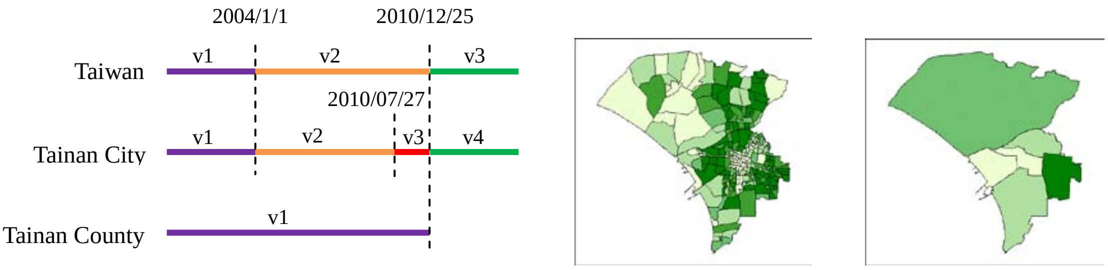
4. TEST CASE

The data we chose to test in this paper is the village-based population data of Tainan City in Sept. 2010. The following discussion uses the making of choropleth map as an example to demonstrate the use of ontology. With its quantitative nature, the measurement levels for the population data is “ratio”. In order to query the correct version of ontology, the application requires users to specify the level of reference system and the valid time of the submitted population data. This constraint is then sent to the geographic identifier reference service to request the correct version of political data. After requested data is returned, the application automatically starts to establish geo-linking between the two datasets with common geographic identifiers. If such 1:1 relationships can be established, the application proceeds to make statistical maps with built-in cartographic knowledge. If not, users are warned that the requested map cannot be made.

(1) One version: The ideal scenario is the statistical data only corresponds to one version of political district data. Because the time of test data is Sept. 2010 and the coverage is Tainan City, the correct version is version 3 of Tainan City, whose valid date is from 2010/07/27 to 2010/12/24 (refer to the metadata section in Figure 3). Figure 4(b) shows the village-based choropleth map of the tested population data.

(2) No appropriate version or multiple versions: If there is no version of data satisfying users' request, the service instead send a warning message to the application. As the two datasets are created independently, the mismatch may be caused by either zero or more than one version of data is available. For the latter case, the application needs to check if there are changes to the spatial extent of political districts in the mapping area. If such changes exist, no statistical maps can be made. .

(3) Level upgrade: With the composition information available, the application can further determine if the information for political districts at a particular level have been completely created. Under such circumstance, it is possible to produce the same type of statistical maps at the upper-level of political districts. Because the spatial description of political districts at different levels has been retrieved, the data processing only requires basic GIS aggregation operation. Figure 4(c) is the political level upgrade case from Figure 4(b).



(a) (b) (c)

Figure 4 Test case of application of political ontology

5. CONCLUSIONS AND FUTURE WORK

Although statistical maps are widely considered as a common tool for displaying the spatial description of analyzed phenomena, its creation and distribution are still impeded by the lack of GIS software training and knowledge about data. Available web services provides a convenient way for domain users to access geospatial data, but domain knowledge is still not enough for using data correctly. This paper defines the knowledge of political districts by the concept of ontology with an open format that computers can understand automatically. We suggest recording the required information about the political districts and the relationships between different political levels. Because of the formalized and standardized knowledge, the mapping mechanism can read the information automatically and determine if the data can be transformed into statistical map and data computing. The proposed approach successfully reallocates the responsibility of map making to individual services and chains them together from a cartographic perspective. The proposed framework which contains built-in and complete cartographic knowledge provides a new way for statistical map making and also improves the production and application of statistical maps.

6. REFERENCE

DCMI, 2010, Dublin Core Metadata Element Set, Version 1.1

Frank van Harmelen and Deborah L. McGuinness, 2004, OWL Web Ontology Language Overview, W3C Recommendation, http://www.w3.org/TR/owl-features/

ISO19112, 2003(E), Geographic information — Spatial referencing by geographic identifiers

Lacasta, J., J. Nogueras-Iso, R. Be´jar, P.R. Muro-Medrano, F.J. Zarazaga-Soria, 2007, A Web Ontology Service to facilitate interoperability within a Spatial Data Infrastructure: Applicability to discovery. Data & Knowledge Engineering, 63(3). 947-971.

Lin, Shih-Yu, 2005, Web-Base Thematic Map Service in OpenGIS Environment, Master's Dissertation of Department of Geomatics, National Cheng Kung University

Open Geospatial Consortium, 2004a, Geolinked Data Access Service, Version 0.9.1

Open Geospatial Consortium, 2004b, Geolinking Service, Version 0.9.1