Facilitating the Interoperable Use of Cross-Domain Statistical Data Based on Standardized Identifiers

Jung-Hong Hong
Department of Geomatics, National Cheng Kung University, Taiwan
junghong@mail.ncku.edu.tw

Jing-Cen Yang
Department of Geomatics, National Cheng Kung University, Taiwan
jingcen@mail.ncku.edu.tw

Abstract

In the big data era, the successful sharing and integration of data from various resources becomes an essential requirement. As statistical data serves as the foundation for professional domains to report the phenomena in the reality according to the selected administration units, its importance has been well recognized. However, statistical data is typically collected and published by different responsible agencies, hence the heterogeneity of how the data is designed, prepared and disseminated becomes an obstacle impeding the automatic and interoperable use in multidisciplinary applications. From a standardization perspective, this research proposes an identifier-based framework for modeling the spatial, temporal and thematic aspects of cross-domain statistical data, such that any piece of distributed statistical information can be correctly and automatically interpreted without any ambiguity for further analysis and exploration. The results indicate the proposed mechanism successfully enables a comprehensive management of indicators from different resources and enhances the easier data retrieval and correct use across different domains. Meanwhile, the interface design exemplifies an innovated improvement on the presentation and interpretation of statistical information. The proposed solution can be readily implemented for building a transparent sharing environment for the National Spatial Data Infrastructure (NSDI).

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Cross-Domain, Statistical Data, Standardized Codes, Visualization

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.31

Category Short Paper

Funding This paper is partial result from the research project (MOST 106-2627-M-006-004) granted by the Ministry of Science and Technology in Taiwan.

1 Introduction

The recent trends of open data and big data analytics have brought a new wave of information revolution, where a tremendous number of cross-domain data is available for uses in the Internet. Since the data may be acquired from various domains and stakeholders, it comes no surprise that users have to deal with unfamiliar or even unknown data structure produced by other domains [5]. In other words, big data are highly heterogeneous [1]. Despite the technology breakthrough in terms of Internet speed and storage has been remarkable, the lack of a comprehensive design, identification and encoding strategy of distributed data is impeding the successful sharing and interpretation of cross-domain applications. Failure to

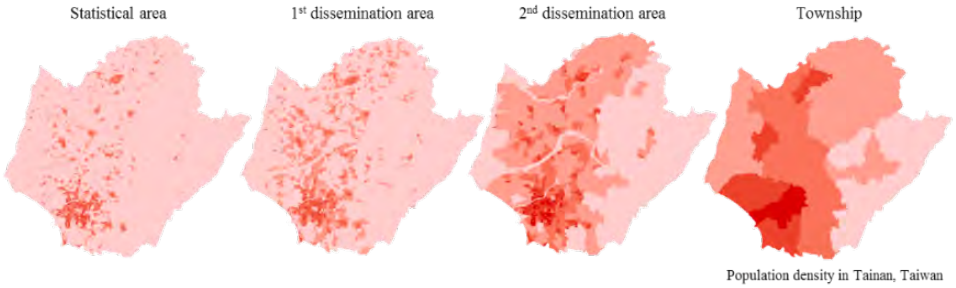

Figure 1 Top four of the most meticulous level of TGSC framework.

overcome such barriers absolutely limits the feasibility of correct decision making and any further exploration. It is therefore necessary to examine how to improve the interoperability of distributed data and enhance the application intelligence of cross-domain data.
Statistics plays an indispensable role in the sustainable development for a nation. Often managed with respect to a particular level of administrative units, statistical data is typically recorded by tables or illustrated by choropleth maps. Various domains follow this space-partitioned framework to establish and update domain statistical data according to a selected frequency. The effective integration of cross-domain statistical data enables a better understanding about continuously changing reality and correct assessment of future action plans. Every country has their own space-partitioned framework for statistical units. For example, a 7-level system named Taiwan Geographical Statistical Classification (TGSC) was established in 2012 as the common references for domain agencies to publish different granularities of statistical data to suffice different application needs (Figure 1). With the development of GIS, the distribution of statistical data evolves from tables with fixed schema [3], Web-based GIS platform (http://datashine.org.uk) to open data [2]. The correct use of statistical data, regardless of the technology being used, requires an in-depth knowledge about the data being used and professional skill for correctly manipulating the GIS software. This requirement becomes a major obstacle after the statistical data is widely and easily available to novice users. Ignorance about the meaning behind the acquired data may easily lead to wrong decisions. Worst of all, users may not even notice they are making mistakes. An interoperable solution for correctly handling and integrating cross-domain statistical data is thus necessary. This paper proposes an identifier-based mechanism for the standardized representation of distributed cross-domain statistical data. It aims to not only simplify the interpretation and processing of statistical data, but also smartly enriches the service content with related indicators and visual aids.

2 Method

A necessary presumption when using statistical data shared by other domains is to correctly interpret its meaning. Four major approaches are adopted in this research to facilitate an interoperable sharing mechanism for overcome current exchange barriers and enrich the capability of decision making:

2.1 Standardized identifier framework

As statistical data typically uses quantitative measures to describe the phenomena for a selected geographic location (Where) from a particular theme consideration (What) at a

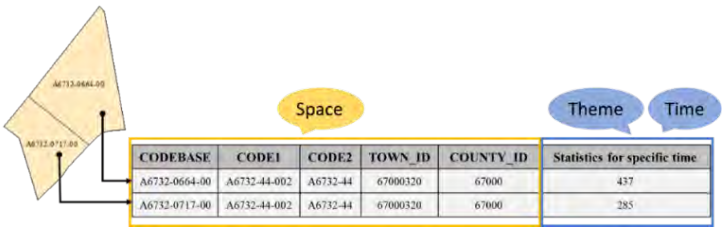

Figure 2 Standardized identifier framework.

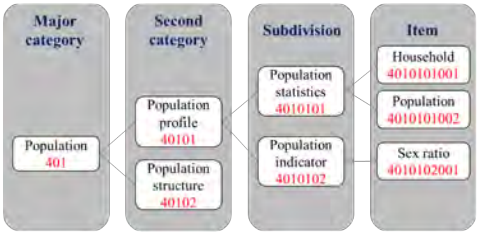

Figure 3 Theme code structure.

Table 1 Statistical method code list.

Code	Statistical method
TC	Total count
SUM	Summation
PC	Percentage
TH	Per mille
RAT	Rate
DEN	Density

given time (When), therefore these three aspects should be unambiguously modeled by unique identifiers to avoid confusion. We proposed to subdivide the attributes into two major parts, one for spatial identification and another for the temporal and thematic description of the statistical indicators (Figure 2). Every row consists of only one unique spatial attribute and a number of temporal/thematic attributes. The TGSC identifiers are directly used for representing the spatial identifiers and can be linked to its geometric representation. The theme codes from different domains are organized following a tree structure, so that every theme is given a unique identifier (Figure 3). The theme code is further extended to include the concept of the indicator (Table 1), such that 4010101001TC represents the indicator for the total count of household. The design of temporal coding system takes the time mode, time resolution, time instance and time range into consideration to ensure all temporal information can be unambiguously represented, interpreted and compared. Table 2 shows two examples. By definition, the population data of every month refer to the status at the end of the month, so we use "TI" to denote this is a time instant, "4010101002" and "TC" to indicate the data is about population and total count, and "02_201701_E" to imply the time is the last day of January, 2017. The number of deaths, on the other hand, is referred to the statistics of a period of time, so it is represented as "TP".

Table 2 Examples of standardized code.

	Population in Jan. 2017	Number of deaths in 2008
Time interpretation	Statistics at the end day of the month	Accumulated in a period
Time mode	Time instant	Time period
Standardized Code	TI_4010101002_TC_02_201701_E	TP_4010402001_TC_01_2010_0

Table 3 Examples of related auxiliary indicators.

	Original indicator	Related auxiliary indicators
Statistical concept	Total population (4010101002_TC)	Average population of 2nd dissemination area (4010101002_L4L2AVG) Standard deviation of total population of 2nd dissemination area (4010101002_L4L2STD)
Domain knowledge	Crude mortality rate (4010406001_TH)	Mid-year population(4010101002_TC) Number of deaths(4010402001_TC)

2.2 Auxiliary indicators

For a chosen indicator, auxiliary indicators are developed for aiding the interpretation of statistical results, e.g., quality measures and spatial variation. Auxiliary indicators are automatically calculated according to the concept of the selected indicator. For example, standard deviation is automatically calculated for every indicator based on average concept; the Spatial Dispersion Index (SDI) proposed by Weng and Tsai in 2006[4] is calculated for every indicator based on the concept of total count. Every auxiliary indicator is also modeled by unique and standardized codes. The package of the chosen indicator and related auxiliary indicators enriches users' understanding about the different aspects of the acquire data without revealing the raw data. Domain providers can therefore flexibly package a set of related indicators either based on the statistical theories (e.g., average and standard deviation) or domain knowledge. Table 3 shows examples about how these two types of related indicators are designed and recorded.

2.3 Management mechanism

With the rules embedded in the coding system, the retrieval of data meeting specific requests can be easily completed by transforming the standardized identifiers. Two types of transformation rules respectively based on spatial and temporal perspectives are developed. The search for statistical data at finer or coarser levels is as easy as using the spatial transformation rule to replace the spatial identifier, while the search of time series data can be also easily completed by using temporal transformation rule to replace the temporal identifier. By registering the tables and the indicators in the data catalog, the search of requested data can be readily completed. Even if the requested data is not directly available, it still can be calculated if its formula is predefined and the required parameters are available (Figure 4).

2.4 Visualization technique

Users are prompted with an integrated interface that can simultaneously illustrate a number of related indicators with maps, tables or charts. The traditional choropleth maps are augmented by new visual aids like highlighted boundaries or spyglasses to make users aware

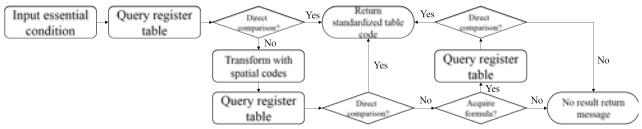

Figure 4 Searching mechanism.

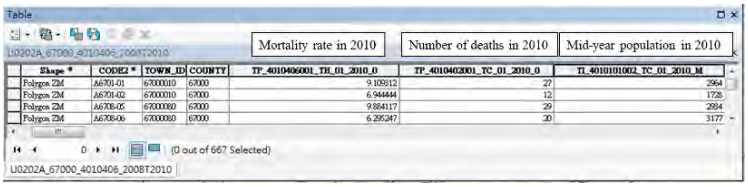

Figure 5 Subpart of mortality rate data in 2010.

Table 4 Query procedure.

Step 1	SELECT Table FROM registration table WHERE Attribute = 'TP_4010406001_TH_01_2010_0' AND Scope = '67000' AND LevelCodeVersion = 'U0202A' AND Time = '2010' Query result: Table = 'U0202A_67000_4010406_2008T2010'
Step 2	Acquire the 2nd dissemination area of mortality rate in Tainan in 2010 SELECT TP_4010406001_TH_01_2010_0 FROM U0202A_67000_4010406_2008T2010

of the possible quality or geographic distribution issue that may otherwise not directly observable. According to users' selected indicators, the developed mechanism analyzes the results of auxiliary indicators and automatically prompts users with meaningful visual illustration.

3 Result

The yearly mortality data for the city of Tainan is chosen as the test data. Figure 5 shows a subpart of the data for the year of 2010. The search for a particular indicator starts with locating the table that includes the requested indicator from the registration table. As the example of table 4 shows, the specified constraints include "TP_4010406001_TH_01_2010_0" (the standardized code for the mortality rate in the year of 2010), "67000"(the spatial code of the Tainan city)," U0202A"(the level of 2nd dissemination area) and "2010"(time constraint). After locating the table ("U0202A_67000_4010406_2008T2010"), the system proceed to retrieve the requested data in step 2. Any statistical data stored in the database can be found in a similar way. For example, the data for one year earlier can be found by using the transformation rules to change the constraint to "TP_4010406001_TH_01_2009_0" and time constraint to "2009".

Assume that the data of the year 2011 is not directly available, it can be calculated according to the predefined formula by filling in the time constraint. As figure 6 shows, the formula for mortality rate requires the number of deaths (TP_4010402001_01_Year_0) and the mid-year population (TI_4010101002_TC_01_Year_M). The requested indicator of "TP_4010406001_TH_01_Year_0" can then be calculated accordingly (Table 5). Even if the data of the number of deaths and the mid-year population is provided by different

$$Crude\ mortality\ rate = \frac{Number\ of\ deaths\ during\ a\ specified\ period}{Mid-year\ population} \times 1000 \xrightarrow{\text{Standardized code}} TP\_4010406001\_TH\_01\_2010\_0 = \frac{TP\_4010402001\_TC\_01\_Year\_0}{TI\_4010101002\_TC\_01\_Year\_M} \times 1000$$

Figure 6 Use standardized codes to represent crude mortality rate.

Table 5 Function of calculating crude mortality rate.

CalculateMortalityRate=(TP_4010402001_TC_01_2011_0/TI_4010101002_TC_01_2011_M)×1000 GenerateMortalityRate('67000','U0202A', '2011')

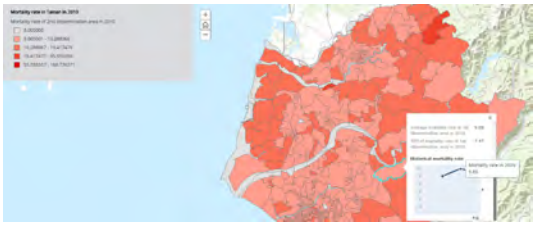

Figure 7 Historical mortality rate of 2nd dissemination area.

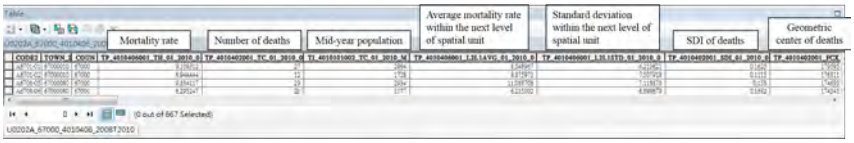

Figure 8 The package of related statistical indicators.

responsible agencies, the search mechanism can still easily find the required data as long as they are willing to comply with the rules of standardized identifiers.

After acquiring the requested time-series data, the interface is designed simultaneously illustrate multiple aspects of indicators for easier visual inspection. Figure 7 shows the interface can show the mortality rate for the 2nd dissemination area for a single year and the historical status after users select a particular dissemination area.

In addition to the mortality rate data, auxiliary indicators related to mortality rate according to statistical model and domain demands are also available. The related auxiliary indicators include the standard deviation within the next level of spatial unit (TP_4010406001 _L4L1STD_01_2010_0), SDI of deaths (TP_4010402001_SDI_01_2010_0), etc (Figure 8). Higher standard deviation usually implies a higher spatial variation within the dissemination area. The geometric center and SDI index number allow users to assess the geographic distribution of features within the dissemination area.

Based on the analysis of the auxiliary indicators, users can easily identify dissemination areas that require special attention. In figure 9, polygons with highlighted boundary imply the 2nd dissemination area with high spatial variation on mortality rate based on the analysis of its corresponding 1st dissemination area. Users can use the Spyglass tool to visually inspect the detailed geographic distribution.

4 Conclusion

In the cross-domain data sharing environment, the proposed standardized is capable of enabling the enrichment and interpretation of individual domain of statistical data, as well as the transformation, integration and visualization of cross-domain statistical data. Every

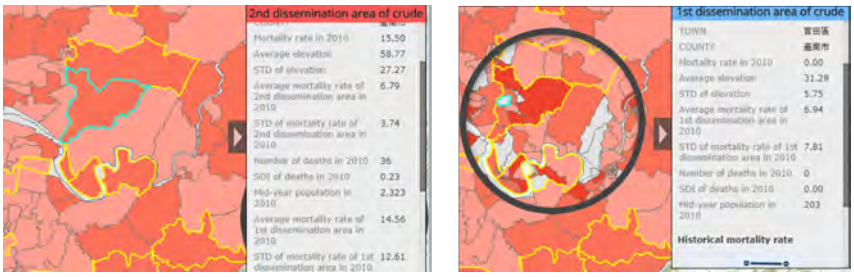

Figure 9 Different levels of statistical data with spyglass interface.

individual piece of distributed statistical data in the proposed mechanism is standardized and self-described, which enables users to develop automatic processing mechanisms and reduce the tedious efforts for conquering the heterogeneity among different domains. In addition to the requested data, users are automatically provided with multiple auxiliary indicators based on the consideration of statistical theory or domain knowledge. In addition to the traditional illustration strategies of table and choropleth maps, users are prompted an innovated interface with awareness capabilities of explaining the illustrated results based on the auxiliary indicators. Based on the consensus identifier framework, the result can be further extended for distributing statistical data in the Internet in the future, e.g., data request via API-based service or Resource Description Framework (RDF).

References

1 Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.

2 Evangelos Kalampokis, Efthimios Tambouris, Areti Karamanou, and Konstantinos Tarabanis. Open statistics: The rise of a new era for open data? In *International Conference on Electronic Government and the Information Systems Perspective*, pages 31–43. Springer, 2016.

3 Corinna Koebnick, Annette M Langer-Gould, Michael K Gould, Chun R Chao, Rajan L Iyer, Ning Smith, Wansu Chen, and Steven J Jacobsen. Sociodemographic characteristics of members of a large, integrated health care system: comparison with us census bureau data. *The Permanente Journal*, 16(3):37, 2012.

4 Pei-Wen Weng and Bor-Wen Tsai. Spatial dispersion index: old conception, new formula. *Journal of Taiwan Geographic Information Science*, 4:1–12, 2006.

5 Yu Zheng. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data*, 1(1):16–34, 2015.