

Discussions on the Multi-source Data Processing Technology in the Construction of GIS Databases

Xiaoyan Ji
National Geomatics Center of China
China
jxy@nsdi.gov.cn

Abstract

One of the prominent tasks in the construction of GIS database is to form a unified, standardized and co-sharing space database with complete content through integrating data from different sources. In the reality world, the geographical space data of different sources often show disparity in terms of content completeness, geometric location, attribute information and semantic definition due to differences in application target and technologies incorporated. Therefore, when we are trying to integrate these data from difference sources, we should first conduct consistency-matching processing over the data. In the construction of a global fundamental datasets that is full of translation of place names and data processing work, this paper analyzes the differences existing in the graphics and attributes of residential data coming from disparate sources, putting forward a consistency matching processing method of residential points combining the space locations with attribute similitude degree. By leveraging the existing database of place names as well as the rules and correlations among the data of residential points from different sources, and carries through program development processing automatically the translation work of most of residential site names that are not quite consistent between their name attributes and location in many cases, decreasing the work amount of manual processing and enhancing the work efficiency.

Keywords: multi-source data integrating

I. Foreword

With the rapid development of 21st Century global economic integration, China has entered into more and more frequent exchanges with other countries, and increasing demand has been felt from users on the geographical information of all countries across the world. While working on global information studies and analyses, relevant the government or corporation found that they have more and more demand on the data contents, data applicability & reliability. also on the emergency-response command, crime prevention and anti-terrorism, border administration, require that the surveying & mapping agencies able to quickly provide the geographical base map data of any region in the world, as well as various thematic geographical information systems and map products based on these base map data.

On the other hand, with its accession to WTO, China is more and more involved in global economic, trade, scientific and technological competitions, up from government departments, enterprises, public institutions and all social circles down to every common users needs to know the physical geography of other countries, so that a global fundamental datasets is needed in international economic, cultural, scientific & technological as well as travels. For these purpose, we collect relevant data and other information on a global scale according to China's technical specification, integrating them into the "Global Geographical Base Map Datasets, (hereinafter referred to as "global

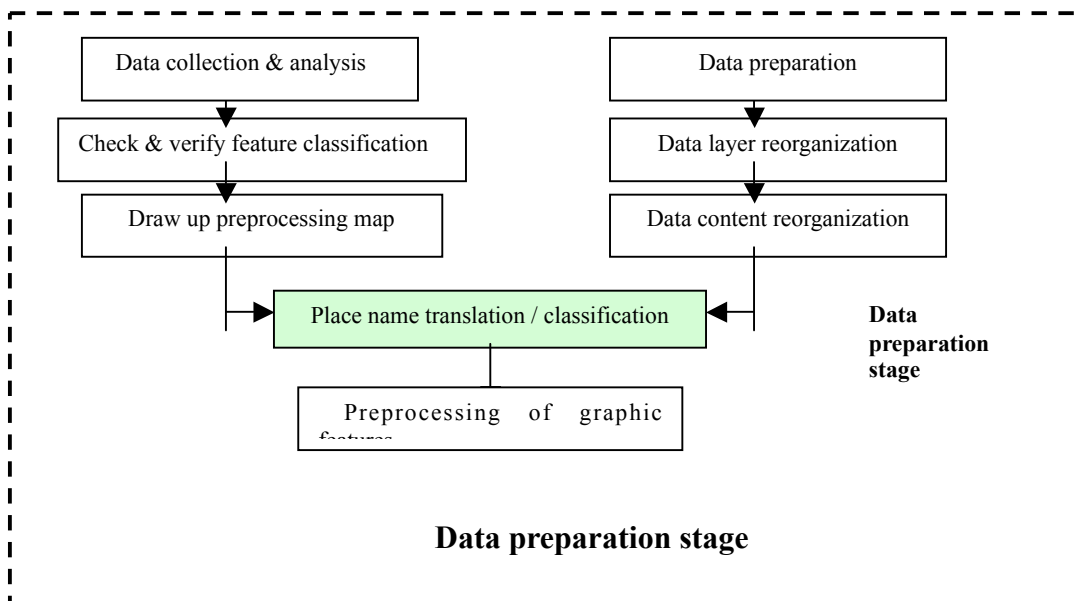
GBMD”) which should be able to provide relatively complete data contents for government administration and decision as well as public information services.

II. Processing of Place Names

Place name (including the name of administration areas, residents, water systems, etc.) are one of the most important geographical elements. The location & attribute information contained in these names will, to a large extent, reflect the database accuracy and data reliability. And one of the major works in the Project implementation is to translate, check, classify and process these place names. This Project leverages on multiple data resources, mainly the existing 1:1,000,000 DCW (Digital Chart of the World) vector data, the global mapping data of some countries, and China’s 1:1,000,000 vector data, with references of authoritative atlas, maps and standards from home and abroad. Based on ample analysis and tests and reflects new system environment, the original data have been reintegrated and place names have been translated in Chinese to create a new global GBMD with rational data structure, relevantly complete contents, standard place names, and to make suitable to users demands. The main process flow including:

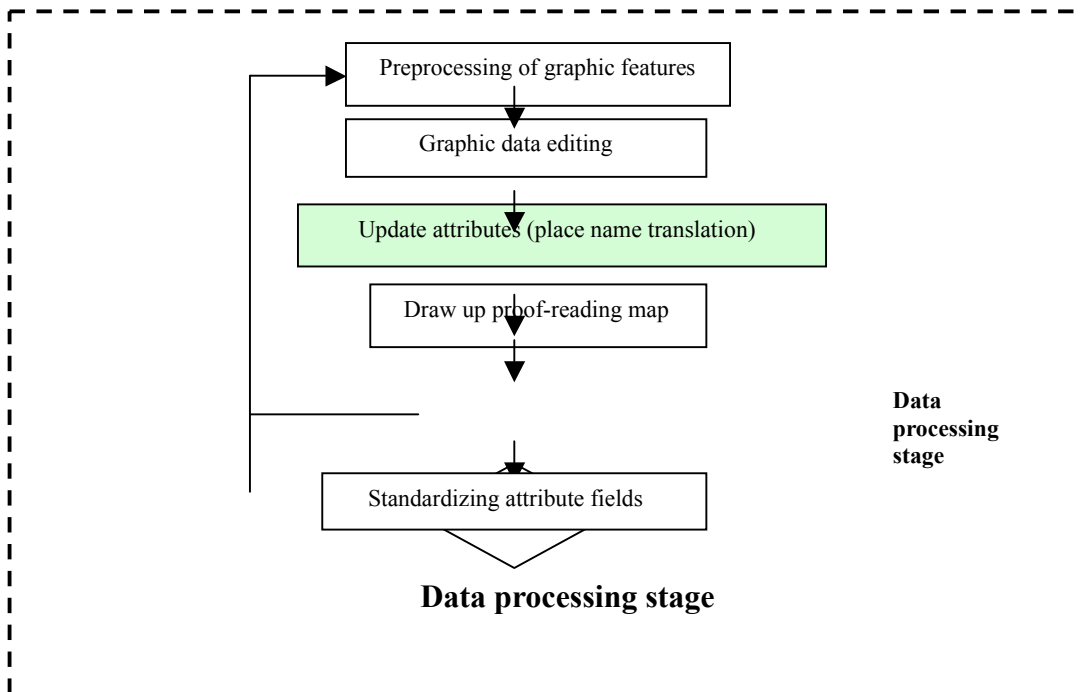
1_ Data preparation stage:

Place names translation and factor grading, this was done by labeled the place names (Chinese and other languages) and their classification on the preprocessing maps;



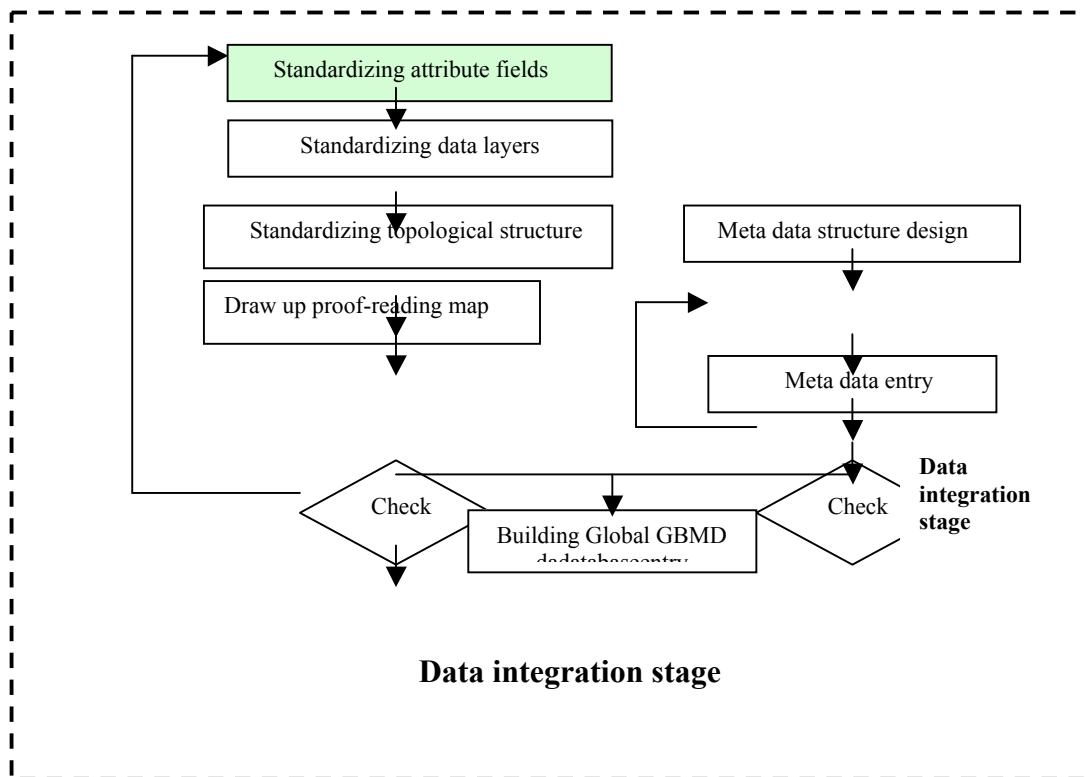
2_ Data processing stage:

updating attribute contents, this was done by entering the above labeled place names into the datasets from Preprocessing of graphic features;



3) Data integration stage:

standardizing attribute fields, this was done by standardizing and normalizing the place names data and classifications entered, to ensure the accuracy and reliability of the translated place names.



It can be seen that during the database building of global GBMD, just the processing of place names would involve a lot of work such as manual marking, manual entry, manual drawing and checking, etc. A preliminary estimation is that, more than 280,000 place names of residential points need to be translated and processed. Faced with so much names translation, data processing, standardizing and adjustment work, we have, through repeated experiment and analysis on the graphical and attribute discrepancies of place names data from different sources, worked out a matching processing method of database consistency based on a combination of their spatial locations and attribute similitude degree. This method leverages the existing database of place names as well as the rules and correlations among the data of residential points from different sources, and carries through program development processing automatically the translation work of most of residential point names that are not quite consistent between their name attributes and location in many cases, decreasing the work amount of manual processing and enhancing the work efficiency.

III. Realization of Technical Solution

Here we would take the example of residential points in place names data to show the realization of technical solution.

During the Project, we took use the 1:1,000,000 DCW data (hereinafter referred to as "DCW data") produced by US National Imagery and Mapping Agency as our basic data source, its residential points data has two types: point and polygon, the attribute fields has the place names in English, all capitalized. Another source is the world place names data

file provided by SinoMaps Press (hereinafter referred to as “SINOMAPS data”), it has been the result of many years accumulation, their file type is spreadsheet database files (.DBF), and the attribute fields contain Chinese and English place names and their coordinates. Our purpose was to leverage on these Chinese place names information to reduce manual translation & processing work. For this, we first, based on the coordinates (latitude & longitude) of SINOMAPS data, transformed them into the residential points, thus created a new layer of data source, which have Chinese and English place names. Table 1 is a list of the name attributes of these sources, and it can be seen that differences exist in the English names of the residential points in different sources.

	English names in DCW data	English names in SINOMAPS data	Chinese names in SINOMAPS data
1	DOWGHA'I	Dowgha I	_____
2	FIRUZ KUN	Firuz Kuh	_____
3	ISESAKI	Isezaki	_____
4	SHIMMACHI	Shinmachi	_____

Table 1 Name Attributes of 4 Residential points in Two Data Sources

From Table 1 we can see that, the English name attributes of the same residential points, though from different data sources, do share much similarity; the only differences are the capitalized/lower case letters and certain characters, therefore this attribute similitude degree can provide basis for place names matching. Usually the name attributes are in the form of character strings, the number of identical characters between different character strings can be viewed as the similarity degree of these name attributes. While software programs can be used to automatically compare and judge the identical letters of name attributes, and standardize upper/lower case letters. Based on this, the similarity degrees can be estimated, where a higher similarity degree indicates a more likely correct matching of residential points. Also through software programs, the Chinese names of these high similarity residential point names can be automatically linked to DCW data, thus completing the translation of place names data. However, identical same names may exist for different places within a country or even in different countries, if this similarity is only based on the name attributes, error will occur. To limit the influence of this uncertainty and get a higher matching efficiency and accuracy, we need to set a boundary limit to the candidate residential points. Precision discrepancies of these data sources do create a difference in the geometric locations of the residential points, but according to geometric matching principles, after eliminating overall and local coordinate difference, the same residential points will always located in a limited boundary zones to same area data of difference sources,. Thus, based on certain buffer distance we can determine the boundary zones of candidate residential points, and to confirm or negate matching according to the similarity degree of name attributes within that zone. A combination of the two will enhance the matching efficiency and accuracy.

The specific steps of the solution in the global GBMD database building are as follows:

1. Data preprocessing:

Since different data sources were collected by different departments, great differences do exist in terms of point precisions, data formats and attribute definitions. Thus a

preprocessing of these data is necessary before we enter into the matching of residential points of different sources, main work include: transform and unify data formats, standardize attribute fields, etc.

2. Determine candidate residential points according to buffer zone distances.

While it's a main step in this matching process, the determination of candidate residential points will not only decide the matching efficiency but also the accuracy. Given the many residential points with similarity of name attributes in different data sets, if it were to carry out a comparison calculation of each residential point in one data set against all residential points in another data set, the workload would be unnecessarily much higher, and a lot of system resources would be consumed. More importantly, the more candidates, the harder to get the most accurate attribute similarity degree, i.e. the matching process would be encumbered, nor will it guarantee the accuracy of this matching. Thus in deciding the candidate residential points, we should reduce information redundancies as much as possible, which is a precondition to obtain accurate matching results. Specific method is described as follows: Take DCW (which is of relatively higher geometric accuracy) as base, and generate a distance buffer zone for each residential point based on a threshold value of certain distance; then overlap SINOMAPS data (which is of lower accuracy) with it, select residential points to be matched according to the buffer zone extension to determine the candidate matching point sets; after that, calculate and compare the name attribute similitude degree between candidate points and the point to be matched, take the one point of highest similarity as the final matching point. At the same time, apply the same ID number to each residential point and its candidate point, so that it can be retrieved in the next matching process.

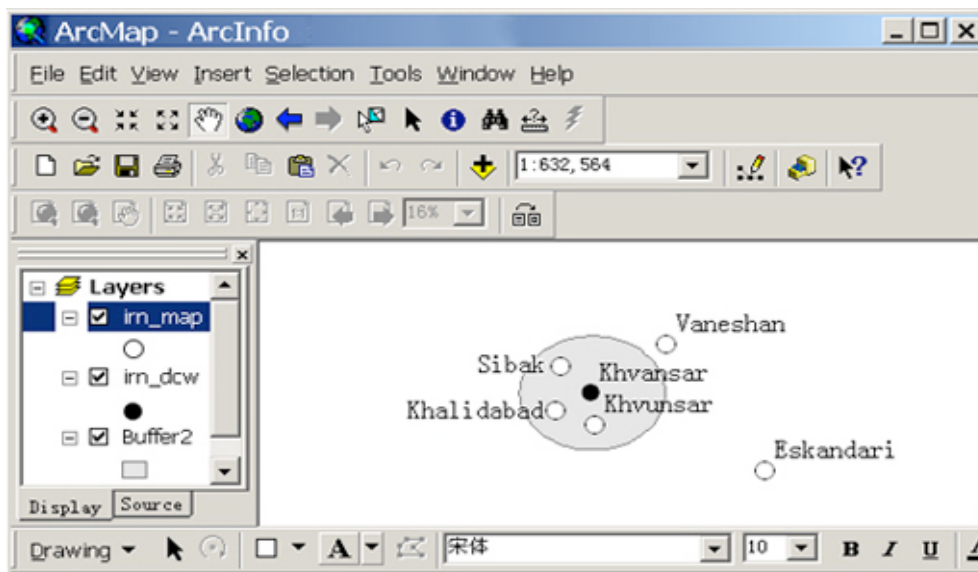


Fig. 1 Determine the extension of candidate population points according to buffer zone distances

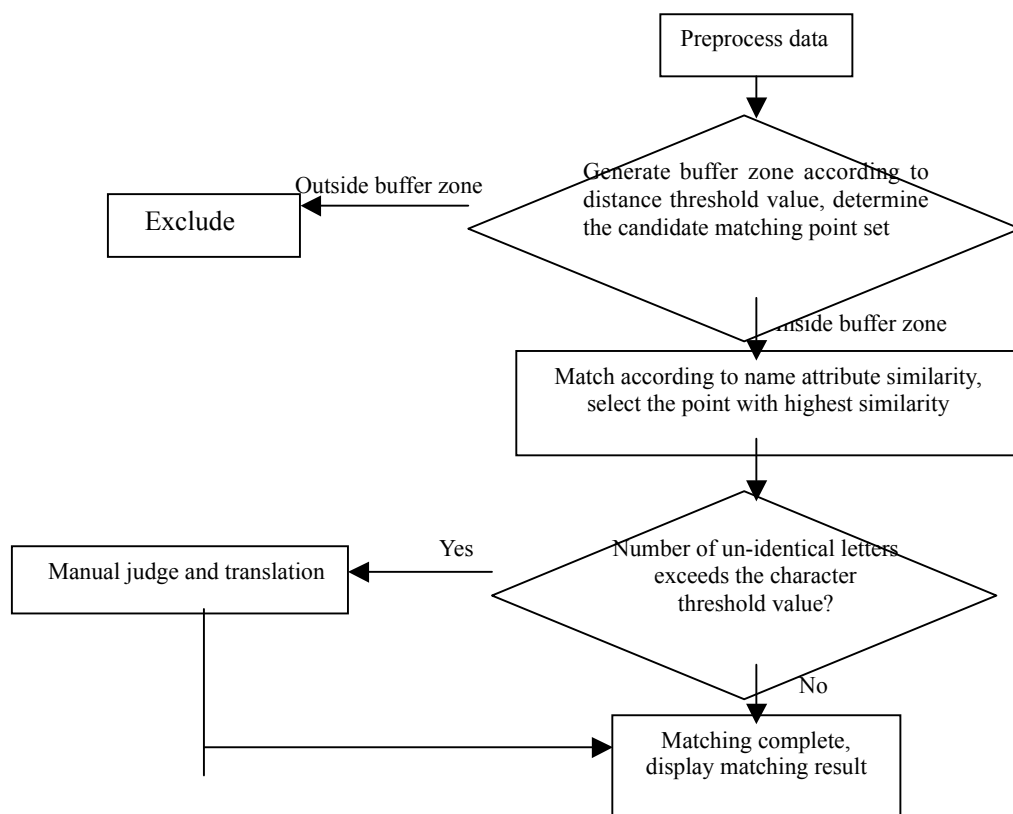
For instance, the black dot in Fig. 1 indicates a residential point to be matched, it's ID number is 1 (ID = 1). The grey circle is the buffer zone generated according to the distance threshold value, within this zone there are three residential points (white dots) from SINOMAPS data. According to the method described above, we take these three points as the candidate points for that residential point, and give them the same ID number (ID = 1). In the next matching processing, we will need to calculate and compare the similarity

between these three candidates and the point to be matched, whereas we do not have to make comparisons with residential points outside this buffer zone.

3. Match residential points according to name similitude degree:

Obtaining the similitude degrees of name attributes between these residential points is a key step in this consistency matching processing. Identical ID numbers will lead the way to the candidate points within the buffer zone, software programs are employed to calculate the number of identical character strings of these names, thus obtaining the similitude degree between each candidate and the point to be matched, the point of the highest attribute similarity will be determined as probably the same residential point. Take the three candidate points in Fig. 1 as an example, the point to be matched is named “Khvansar”, a calculation of the above method will render that, the candidate point named “Khvunsar” has seven same characters between the two, i.e. their similarity degree is 7, whilst the other two points have similarities below seven, i.e. their similitude degree is less than 7. Therefore, we can determine that the one with highest attribute similitude degree “Khvunsar” and the point to be matched “Khvansar” are same residential point. After that its Chinese name will be linked to the attribute table to complete the translation of residential names.

As a summary, a basic workflow of the consistency matching process in the building of global GBMD database can be drawn as follows:



IV . Test and Conclusion

Based on the above-described method, we took the place name data of some Asian countries as an example to carry out a technical experiment on the residential points of DCW data and SINOMAPS data. Data preprocessing were done in ArcGIS and Oracle databases; the automatic estimation and calculation of the distance values and attribute similitude degree of residential points were done with programs written in SQL language and ARC/INFO AML language, they also link the attribute data of identical points to

complete the consistency matching processing of residential points of the same area in two data sources; the matching result was then taken as the basis for name attributes integration of the residential points, thus completing the translation of name attributes of residential points in DCW data. Table 2 shows the result of translation by using the above described method for the names of residential points of some Asian countries.

行标识	F_CODE	F_CODE_DES	NAME	FENAME	FCNAME	SOC
3992	AL105	Settlement	BAGAN DATUK	Bagan Datuk	巴明拿惹	MYS
961	AL105	Settlement	BAGANGA	Baganga	巴冈阿	PHL
1861	AL105	Settlement	BAGARCHHAP	Bagarchhap	伯格尔恰希	NPL
170	AL105	Settlement	BAGGAO	Baggaao	巴高	PHL
3076	AL105	Settlement	BAGH	Bagh	巴格	IRN
4136	AL105	Settlement	BAGHA	Pak Kad	巴卡	IND
121	AL105	Settlement	BAGHRAN	Baghran	巴格兰	AFG
101	AL105	Settlement	BAGRAME	Bagrame	巴格拉米	AFG
4019	AL135	Native Settlement	BAHARU	Kg. Baharu	甘榜巴鲁	MYS
4013	AL135	Native Settlement	BAHARU BUKIT TINGGI	Kampong Bukit Tinggi	武吉丁宜	MYS
4038	AL105	Settlement	BAHAU	Bahau	马口	MYS
5943	AL105	Settlement	BAHUMOTEPE	Bahumotewe	巴洪莫泰沃	IDN
4255	AL105	Settlement	BAHROR	Behror	贝赫罗尔	IND
477	AL105	Settlement	BAHUD	Balud	巴卢德	PHL
5601	AL105	Settlement	BAIA	Bayah	巴亚	IDN
736	AL105	Settlement	BAIS	Bais	拜斯	PHL
5706	AL105	Settlement	BAJA	Bayah	巴亚	IDN
6192	AL105	Settlement	BAJAH	Bayah	巴亚	IDN
2878	AL105	Settlement	BAJESTAN	Bejestan[Bijistan]	贝杰斯坦	IRN
1841	AL105	Settlement	BAJURA	Bajura	巴朱拉	NPL
4007	AL135	Native Settlement	BAKAH	Pak Kad	巴卡	MYS
3032	AL105	Settlement	BAKAL	Bakal	巴卡尔	IRN
2566	AL105	Settlement	BAKANAS	Bakanas	巴卡纳斯	KAZ
2575	AL105	Settlement	BAKBAKTY	Bakbakty	巴克巴克特	KAZ
2531	AL105	Settlement	BAKHTY	Bakhty	巴赫特	KAZ
4081	AL105	Settlement	BAKO	Pak Khop	巴科	MYS
840	AL105	Settlement	BAKUNG	Bacong	巴孔	PHL
3689	AL105	Settlement	BALA	Bala	巴拉	TUR
1653	AL105	Settlement	BALA DHAKA	Bala Dhaka	巴拉塔加	PAK
841	AL105	Settlement	BALABAC	Balabac	巴拉巴克	PHL
1947	AL105	Settlement	BALAKANDI	Bäläkündi	巴拉甘迪	BGD
633	AL105	Settlement	BALAMBAN	Balamban	巴兰班	PHL
281	AL105	Settlement	BALANGA	Balanga	巴朗牙	PHL

Table 2 Translation Result of Residential points Names of Some Asian Countries

Note that, due to the multi-language nature of Asian place names, quite big difference is shown in their English names of DCW data and those of SINOMAPS data, which caused the relatively low ratio of correct matches from this method. Great improvement in this ratio can be found for European and American countries and regions, practice showed that the ratio of correct matches can arrive at around 50%, and the employment of software programs for automatic calculation and judge is of such high accuracy that the work efficiency and translation accuracy are enhanced. However, given the abundant semantic information of vector databases and the complex nature of topological relations, this vector data matching itself is a complex process, it is unavoidable to have mismatches or failures of place names data. Therefore, manual intervention is still needed in the work process. Despite that, the application of this method for automatic and interaction matching & translation of place names attributes from different sources can, on a global scale, reduce by 50% the workload of manual identification, processing and data entry of translating the place names of most countries, thus enhancing the work efficiency and translation accuracy.

References: Lin Lishu, Ji Xiaoyan, Jiangjie, “Technical study of the consistency matching of residential points from multi data sources”, *Symposium of the 9th Annual Meeting of China GIS Association*, Oct. of 2005, P.989.