

## A PRACTICE APPROACH OF MULTI-SOURCE GEOSPATIAL DATA INTEGRATION FOR WEB-BASED GEOINFORMATION SERVICES

Huang Wei<sup>a</sup>, Jiang Jie<sup>a</sup>, Zha Zhuhua<sup>a</sup>, Zhang Hongping<sup>a</sup>, Wang Cong<sup>a</sup>, Zhang Junhui<sup>a</sup>

<sup>a</sup>National Geomatics Center of China, 28 West road, Lianhuachi, Beijing, P.R. China-  
huangwei@nsdi.gov.cn

**KEY WORDS:** Geospatial data, Data Integration, Web-based geoinformation services, Practice approach, Principles, Procedure

### ABSTRACT:

Geospatial data resources are the foundation of the construction of geo portal which is designed to provide online geoinformation services for the government, enterprise and public. It is vital to keep geospatial data fresh, accurate and comprehensive in order to satisfy the requirements of application and development of geographic location, route navigation, geo search and so on. One of the major problems we are facing is data acquisition. For us, integrating multi-sources geospatial data is the mainly means of data acquisition.

This paper introduced a practice integration approach of multi-source geospatial data with different data model, structure and format, which provided the construction of National Geospatial Information Service Platform of China (NGISP) with effective technical supports. NGISP is the China's official geo portal which provides online geoinformation services based on internet, e-government network and classified network. Within the NGISP architecture, there are three kinds of nodes: national, provincial and municipal. Therefore, the geospatial data is from these nodes and the different datasets are heterogeneous. According to the results of analysis of the heterogeneous datasets, the first thing we do is to define the basic principles of data fusion, including following aspects: 1. location precision; 2. geometric representation; 3. up-to-date state; 4. attribute values; and 5. spatial relationship. Then the technical procedure is researched and the method that used to process different categories of features such as road, railway, boundary, river, settlement and building is proposed based on the principles. A case study in Jiangsu province demonstrated the applicability of the principle, procedure and method of multi-source geospatial data integration.

### 1. INTRODUCTION

Geospatial data resources are the foundation of the construction of geo-portal which is designed to provide online geoinformation services. Today we are rather faced with an excess than with a lack of data. Digital geographic datasets have been acquired multiple times, for multiple application, and in multiple scales. With decades efforts, the national level databases of China has covered the whole territories including 1:1 million, 1:250,000 and 1:50,000 scale. While the provincial database (1:10,000) covers more than 50% territories and larger scale data cover most of downtown areas. In the meanwhile, great volume of satellite images and aerophotos have been

collected. Recently more and more images come from Chinese surveying satellites such as ZY-3, etc. According to statistics, there are already about 1158 TB images in the National image database, among them 965 TB aerophotos and 194 TB satellite images. There are several versions of low resolution ( $\leq 2.5$  meter) satellite image covering the whole land area. Most of the urban areas have been covered by high resolution images (higher than 1 meter).

Therefore the major issue for us is how to integrate these data resource and ensure a high quality data. Because it is vital to keep geospatial data fresh, accurate and comprehensive in order to satisfy the requirements of application and development of geographic location, route navigation, geo search and so on.

And for the geo-portal users, up-to-date of data is more important than accuracy in some cases such as POI searching or car navigation. Therefore, we need to solve two problems. First, how can we minimize the effort for keeping the databases up-to-date? Secondly, how can we combine the information given with different databases?

In this paper, a practice data integration approach is proposed which provided the construction of National Geospatial Information Service Platform of China (NGISP) with effective technical supports. The paper is structured as follows. Firstly, we briefly sketch the context of our work, that is, the NGISP and it's requirements for multi-source geo-spatial data integration. Then we present our integration method including principles, procedures and so on. Finally, we analyze and discuss the result of the experimental data in Jiangsu province of China and conclude the paper.

## 2. BACKGROUND OF THE WORK

Challenges arise accompanying with the further and wider application of the location-based information. One of the most urgent challenges is the one-stop access and integrated-usage of the multi-scale and distributed databases.

To solve this problem, a program was initiated by NASG (The National Administration of Surveying, Mapping and Geo-information of China) in 2009 to establish the National Geospatial Information Service Platform of China (NGISP), with Chinese name TIANDITU which means Map World.

NGISP is the China's official geo portal which provides online geoinformation services based on internet, e-government network and classified network. It is designed as an important part of the geospatial framework for Digital China, aiming to promote geographic information resources sharing and improve the capability and efficiency for better services. Ordinary users can use the website to browse maps, locate places, measure distances or areas, plan car driving routes. Professional users can access TIANDITU's resources via service URLs to development value-added services and applications. TIANDITU also provides many Application Programming Interfaces (APIs) to facilitate integration of its service resources and various systems or websites.

Within the NGISP architecture (Fig.1), there are three kinds of nodes: national, provincial and municipal (or data centers). The nodes are connected by Internet or Intranet. The geospatial data

is from these nodes arc collected by various surveying and mapping agencies, enterprises, professional agencies, social organizations and volunteers, and the different datasets are heterogeneous.

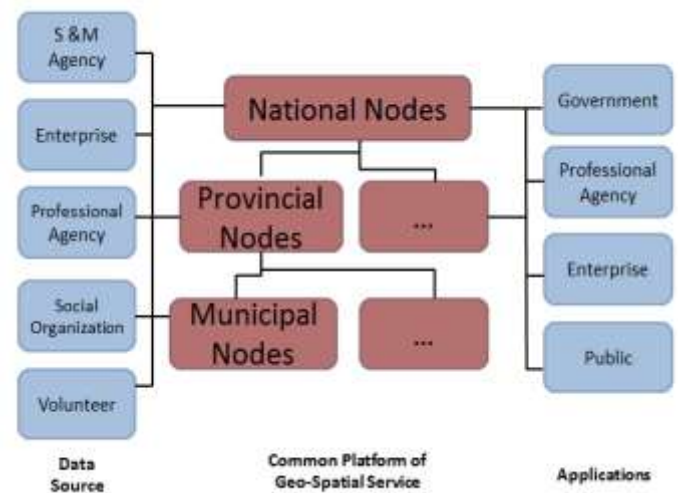


Figure 1 The Architecture of NGISP

In order to provide high quality online geospatial information service, it is necessary to integrate these multi-source datasets. However, this task is very hard in actually due to heterogeneity in terms of data modeling concepts, data encoding techniques, storage structures, access functionalities, etc.

## 3. INTEGRATION APPROACH

### 3.1 Basic Principles

The integration approach is designed based on analysis of multi-source heterogeneous datasets. These datasets collected or produced by different surveying and mapping agencies, enterprises, professional agencies or social organizations. Some of that is stored in traditional files, while some is stored in database with spatial data engine. Their reuse for web geoinformation services is a difficulty, due to poor documentation, obscure semantics of data, diversity of data sets including what information is stored, how it is represented and structured, what quality it has, which date it refers to, etc. The main problem when integrating these datasets is the different geometrical representation of spatial objects. Additional problems arise because of different views of the world and different data quality characteristics.

Therefore, to construct the one stop geo-portal, we have to identify what data is needed. It is necessary to look around to

all these datasets to find out which ones is the best for the needs. Most likely, part of that will indeed exist, spread into pieces over various heterogeneous data stores, part of it will not exist and will have to be acquired. Eventually, the newly acquired data and the many pieces of reused data will be integrated into a single, uniform, non-redundant data store, which will serve as the underlying database for the online geoinformation services. So the first important thing of data integration is to define the basic principles, that is, 1) developing a correct understanding of the semantics of multi-source data, i.e. what they really mean, 2) establishing an accurate correlation structure, and 3) choosing a well-suited integrated description based on integration goals and on the available data conversion techniques. According to the principles, operators can integrate various data sources into a single framework. The principles include the following aspects:

- (1) Location precision: The location precision of data from different sources is inconsistent in general, the high precision data is prevail.
- (2) Geometric representation: Dataset with more accurate geometric representation is the best one in this case.
- (3) Up-to-date state: We have to select the latest features from multi-source datasets.
- (4) Attribute values: It is a headache problem, the same attribute fields of different datasets have different values, we must evaluate which one is correct or latest.
- (5) Spatial relationship: The results of the fused data elements of all kinds of information must be logically consistent, the conflicts, such as house located in the river, are not permitted.

These principles can be treated as requirements about the quality of the data integration results.

### 3.2 General Procedure

The general data integration procedure is illustrated as a process chart in Figure 2, is a seven steps procedure. There are two manners to perform the procedure: 1) manual approaches, i.e. those methodologies which aim at providing database administrators with adequate tools to perform integration; basically, such tools are schema definition or manipulation languages, and 2) semi-automatic approaches, where the aim is to have integration automatically performed by a tool based on the correspondence assertions given by the database administrator. We believe that the second approach is better

suits the needs of on-line service data integration.

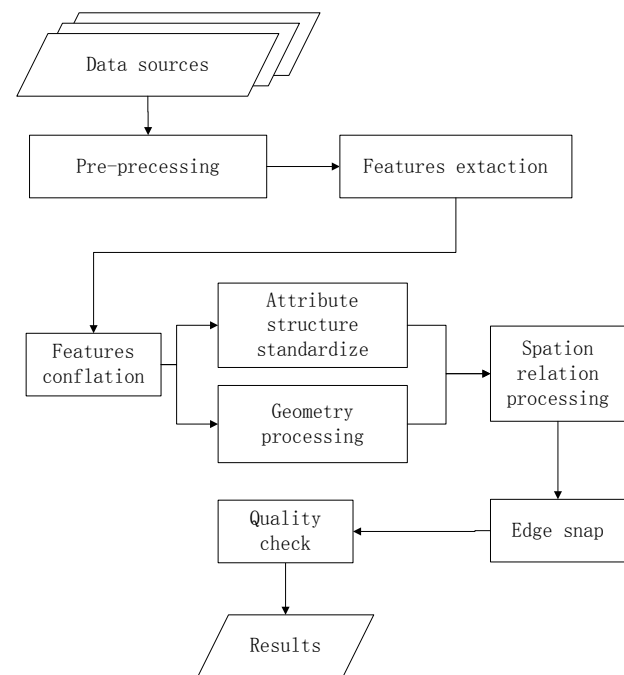


Figure 2 General Data Integration Procedure

#### (1) Pre-processing

This step normalizes the input datasets to ensure that they are compatible. For instance, they must have the same coordinate system. This may also involve format translation and any other basic preparation of the datasets.

Another important work of pre-processing is data quality assurance. Some integration tasks require that datasets have a given level of internal consistency. For instance, coverage alignment algorithms require that the input datasets are in fact a clean coverage. During this step the internal consistency of the datasets is verified and if necessary improved.

#### (2) Features Conflation

Conflation is the action of unifying two distinct datasets into a new dataset. This may be relatively easy to do or extremely difficult, depending on the complexity of representation and the size and quality of the datasets involved.

During this step common features between the datasets are matched including attribute and geometry information. This may be done either in an automated fashion using one or more conflation algorithms, or via manually determined matches. Once features have been matched the match information can be used to improve the quality of one or both of the input datasets. Features conflation involves updating one dataset with information from the other. This information can be either

attributes or geometry to be added to an existing feature, or entire features to be added to the dataset.

### (3) Edge Snap

Because the datasets are processed with map sheets, it is necessary to snap the common features between adjacent sheets.

## 4. EXPERIMENT AND CONCLUSION

The experiment performs the data integration procedure with the same geographical area. The main conflated features include road, house, river and greenbelt and place name. It is obvious that the results of integration have more detailed information and more up-to-date (Figure 3).

This paper has proposed a practice approach of multi-source geospatial data integration for web based geoinformation services. Firstly, we defined the principles about feature choosing, conflation and spatial relationship processing, etc. It is necessary and important for the work of data integration since the principles explicitly describes the requirements of NGISP online services and it is the quality standard of data integration results. Then the procedure is provide including three main phases: data pre-processing, features conflation and map sheet snap. The case study in Jiangsu province shows that the approach is practicable and the integration results are entirely satisfactory. In the future, the challenge we faced is how to enhance the automation of data integration in order to extremely improve the efficiency of data production and update.

### Reference

DAI Tegen ,QIU Dongsheng ,BAO Guangshu ,YAO Rui ,XIE Wenbing, Geo scientific data model and research of data fusion technology, J . CENT. SOUTH UNIV. TECHNOL, 32 (3), 221—223.  
 CHEN Huanxin, SUN Qun, XIAO Qiang, XIAO Jihua, Application of Spatial Data fusion in the Production and Updating of Spatial Data. Geomatics and Information Science of Wuhan University. 39 (1), 117-122



Figure 3 Experiment Results