

Study on resources and environmental data integration towards data warehouse construction covering trans-boundary area of China, Russia and Mongolia

J Wang¹, J Song, M Gao and L Zhu

State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China

E-mail: wangjl@igsnr.ac.cn

Abstract. The trans-boundary area between Northern China, Mongolia and eastern Siberia of Russia is a continuous geographical area located in north eastern Asia. Many common issues in this region need to be addressed based on a uniform resources and environmental data warehouse. Based on the practice of joint scientific expedition, the paper presented a data integration solution including 3 steps, i.e., data collection standards and specifications making, data reorganization and process, data warehouse design and development. A series of data collection standards and specifications were drawn up firstly covering more than 10 domains. According to the uniform standard, 20 resources and environmental survey databases in regional scale, and 11 in-situ observation databases were reorganized and integrated. North East Asia Resources and Environmental Data Warehouse was designed, which included 4 layers, i.e., resources layer, core business logic layer, internet interoperation layer, and web portal layer. The data warehouse prototype was developed and deployed initially. All the integrated data in this area can be accessed online.

1. Introduction

The trans-boundary area between Northern China, Mongolia and eastern Siberia of Russia is a contiguous geographical area located in northeastern Asia. The region has a complex ecological environment, a variety of climate zones and typical human-Earth relationships [1]. Many common issues in this region need to be addressed based on a uniform resources and environmental data warehouse, such as natural resources management, environmental monitoring, natural disaster forecasting, social sustainable development policy making, and so on [2].

Many countries, including the USA, Germany, Japan and Korea have launched long-term scientific research and integrated scientific surveys in this area [3], such as the Northern Eurasian Earth Science Partnership Initiative (NEESPI, <http://neespi.org/>). But most of these research activities focused on specific themes, without providing the general integrated databases for the study in this region.

How to build this uniform resources and environmental integration data warehouse is not only an urgent requirement, but also a big challenge [4]. Facing to this issue, Chinese, Russian and Mongolian scientists launched a joint science expedition project in this area in 2008 [1]. Based on this background, the paper discussed how to integrate resources and environmental data in the trans-boundary area towards data warehouse construction.

¹ To whom any correspondence should be addressed.



Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

2. Study area and technique flow

2.1 Study area

The joint scientific expedition area covers northern China north of the Yellow River, Mongolia, and eastern Siberia and far eastern Russia, as shown in figure 1.

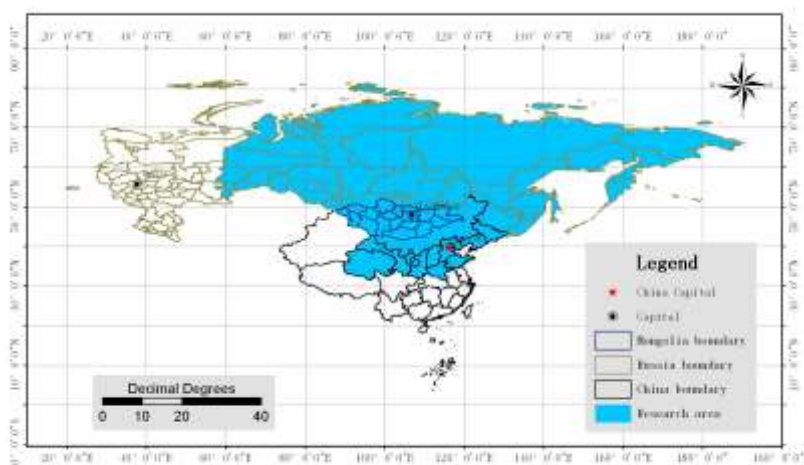


Figure 1. The sketch map of research area.

2.2 Technique flow

The technique flow of resources and environmental data integration in this trans-boundary area included 3 steps. Step 1 was data collection standards and specifications drawing up, step 2 was for multi-disciplinary and multi-sources data reorganization and integration, step 3 was the design and construction of data warehouse. The technique flow was shown in figure 2. There were cross and overlay procedures in each connected steps, because many of the data sets have pre-processing or post-processing in the whole data integration technique flow.

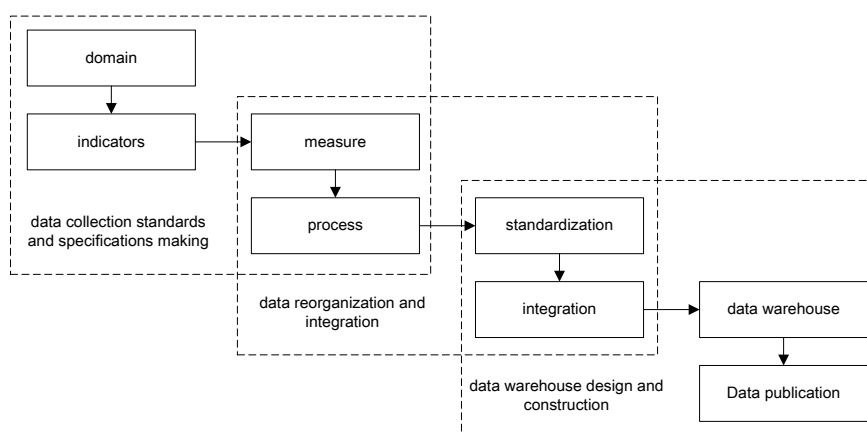


Figure 2. Technique flow of data processing and integration.

3. Data collection standards and specifications

In order to ensure the data quality, several data collection standards and specifications were drew up before the data was collected in the field. These data collection standards and specifications included more than 10 domains. The popular 10 domains were land cover survey based on remote sensing, soil sampling, forest ecology survey, grassland ecology survey, water resources survey, aquatic organisms and ecosystems survey, typical lake environment survey, social and economic survey, living environment survey, and aerosols remote sensing monitoring. Various domains' indicators and measures in related data collection standards and specifications were listed in Table 1.

Table1. Data collection indicators and measures.

| Domain | indicators | measures |
|--------------------|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| Land cover | cropland, forest, grassland, water, built up area, unused land | (1) artificial visual interpretation (2) data update based on global land cover data sets |
| Soil | organic matter, nitrogen, phosphorus, potassium, salinity, etc. | (1) sampling collection in the field (2) data collection from literature |
| Forest | distribution, type, quality, site conditions, biomass, etc. | (1) sampling collection in the field (2) biomass retrieval based on remote sensing (3) data collection from literature |
| Grassland | distribution, type, quality, site conditions, biomass, etc. | (1) sampling collection in the field (2) biomass retrieval based on remote sensing (3) data collection from literature |
| Water resources | distribution, volume, and their physical conditions of Lake, reservoir, and river | (1) field survey and visiting (2) data collection from literature and official year book |
| Water quality | PH value, salinity, dissolved oxygen, transparency, total phosphorus, total nitrogen, etc. | (1) cross-sectional investigation (2) sampling collection in the field |
| Aquatic organisms | Plankton, benthic algae, large invertebrate, fish, water bird, etc. | (1) sampling collection in the field (2) data collection from literature |
| Social economy | Population, Gross Domestic Products, employee, industry, agriculture, etc. | (1) field survey and visiting (2) questionnaire survey (3) data collection from official year book or literature |
| Living environment | Social system, living system, physical system, support system | (1) field survey and visiting (2) questionnaire survey (3) data collection from official year book or literature |
| Aerosol | Distribution, aerosol optical depth | (1) AOD biomass retrieval based on remote sensing (2) Sampling monitoring |
| others | - | - |

4. Data reorganization and integration

4.1 Frame of data reorganization

Since the various data set types of entities needed to be harmonized, these datasets were divided into two kinds of data category types firstly, i.e., regional area dataset and the typical area dataset. Then, each category was subdivided into multiple data sets for the reorganization and integration of data entities, including 20 resources and environmental survey databases in regional scale, and 11 in-situ observation databases on typical area. When the specific dataset was stored, each dataset was constructed in four subfolders in file system. Table 2 shows the data entities storage architecture and content

Table2. The data type and content in each dataset.

| | Class_type | Name | Content |
|---------|----------------|------------------|---------------------------------------------------------------------------------------------------------------|
| dataset | Folder | Data | data entity |
| | Folder | documents | documents record, including metadata table (.xls), data document (.doc) and data reorganization report (.doc) |
| | document(.txt) | note.txt | the data processing and data source record |
| | document(.mxd) | dataset_name.mxd | data layers display* |

*Note: data stored in a relative path.

4.2 Regional and typical area data reorganization

Regional area dataset was subdivided into three subtype datasets in accordance with the correlation of the data entities, i.e., physical geography datasets, resources and environment datasets, and social and economic datasets. These datasets cover the fields of basic geographic, land use/cover, ecological geography division, soil, wetland, desert, water resources, water environment, aquatic organisms, forestry, grassland, urbanization, aerosol and greenhouse density, population and social economy, etc. All the regional data were transferred to coverage or grid format. Typical area datasets focus on specific samples/investigation/surveillance data in the field, including soil samples investigation, lake water samples, important fishes, aerosol samples monitoring, etc.

Taken the land cover data set with 500m resolution in the area for an example, the metadata and description information was listed as below:

- Content: land cover data set with 500m resolution in Northeast Asia area.
- Data source description: raw data from European Space Agency's Glob Cover through global co-production of global data sets. Downloaded from <http://ionial.esrin.esa.int/>.
- Data processing method: the classification system was transformed from raw data firstly, and then using resampling methods to generate the land cover data.
- Data quality description: the accuracy was 73%, and some obvious errors were excluded during the processing of data tables generation.

5. Data warehouse design and construction

The resources & environmental data warehouse platform aims at providing a web application for data integration, sharing, and distribution among the scientists scattered on different locations and countries.

5.1 The architecture of the data warehouse

The architecture of the data warehouse software platform insisted of four layers. They were resources layer, core business logic layer, internet interoperation layer, and web portal layer from bottom to up, as shown in figure 3.

The resource layer provided data storage function for resources and environmental spatial-temporal data and related descriptions information, such as metadata, descriptive documents, thumbnails, and so on. The core business logic layer covered data integration, access, query, and visualization components. Three different engines for specialized database, distributed files and metadata were designed for calling the various data repositories in the resource layer. The user authentication and permission module was designed for the data security. The interoperation layer provided two kinds of interfaces for calling the functions of the data warehouse platform over the Internet based on HTTP and TCP protocol. The web service interface based on HTTP and the RPC (Remote Process Call) based on TCP had high effectiveness for data access. The web portal layer directly interacted with the user through the web pages. Web portal for user authorization was built separately so as to unify user authorization for multiple web portals on data warehouses located in different places.

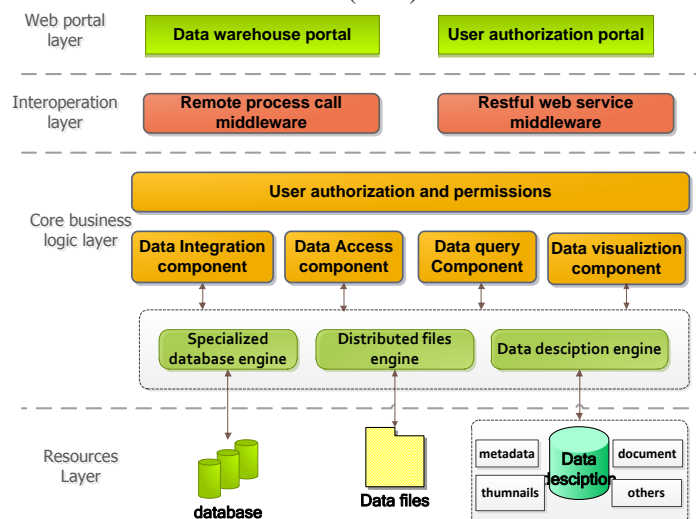


Figure 3. The architecture of data warehouse platform.

5.2 Development of data warehouse

North East Asia resources and environmental data warehouse prototype was developed in J2EE environment. MySQL database was selected as the metadata management storage and PostgreSQL was selected for spatial data management. It was deployed and running on CentOS Linux, supported by cluster, cache, proxy and schemas technologies. There were 6 clusters serving for the data access, i.e., database cluster, file server cluster, application server cluster, http server cluster, database front-end cache and web pages cache. The running portal interface was shown in figure 4.



Figure 4. The portal of data warehouse platform.

6. Discussion and conclusion

Based on the joint scientific expedition among China, Russia and Mongolia, the paper discussed how to integrate these resources and environmental data in this trans-boundary area towards data warehouse construction. Through 3 technique steps, a final data warehouse prototype was designed and development. According to the uniform data collection and reorganization standards and specifications, 20 resources and environmental survey databases in regional scale, and 11 in-situ observation databases were reorganized and integrated in the data warehouse. Some discussion was listed as below.

- (1) Standard and specification not only ensure that the different disciplinary data has uniform data structure or storage format, but also ensure the data content quality. Only those resources and environmental data with high quality would be integrated and shared in the data warehouse.

While, data quality control was a key issue for data integration in this region for long term, especially for long temporal serials and multi-scale data sets integration.

- (2) Resources and environmental data contain different data source, type, format, size and other characters. The paper gave an integration solution for 2 kinds of data, i.e., regional scale survey data and in-situ observation data in typical area. Different classification method may produce different dataset classes, and the data entity reorganization and integration would be changed following. The classification method would be researched in the further study. Meanwhile, related metadata, data document, thumbnail and other description information were very useful when those data were integrated.
- (3) Although the data warehouse prototype was developed initially, many functions had to be taken into account in the future, such as multi-standard metadata support function, data online mapping and analysis function, attribute data contrast and visualization function, etc. In the coming “Big Data” era, how to advance the data mining and analysis capacity would be a long term mission for North East Asia Resources and Environmental Data Warehouse construction and application.

Acknowledgments

This paper was supported by Chinese Academy of Sciences (Grant No: KZZD-EW-08) and National Scientific & Technology Basic Work Program of China (2007FY110300).

References

- [1] Dong S, Li Y, Li F, etc. 2011 *J. Resour. Ecol.* **2** 250
- [2] Wang J, Zhu L, Sun C 2011 *J. Natur. Resour* **26** 1129
- [3] Mun Y, Ko I, Janchivdorj L 2008 <http://www.kei.re.kr>
- [4] Wang J, Zhu L, Yang Y and Zhang L 2011 *J. Resour. Ecol.* **2** 266