Geo-Web Service Tool for Spatial Data Integrability

Hossein Mohammadi, Abbas Rajabifard¹, Andrew Binns, Ian P. Williamson
Centre for Spatial Data Infrastructure and Land Administration
Department of Geomatics
University of Melbourne

Abstract

The integration of multi-source heterogeneous spatial data is one of the major challenges for many spatial data users. Users put much effort to identify and overcome inconsistency among data sets through a time-consuming and costly process. Spatial applications that rely on multi-source heterogeneous data also suffer from the lack of automatic mechanism to identify the inconsistency items and assign an appropriate solution for any particular item. An effective integration necessitates the identification of the inconsistency among data sets and the provision of necessary standards and guidelines in order to overcome the inconsistency, and then data sets can be manipulated based on the guidelines and proposed solutions.

The paper follows two main streams. Firstly, the results of a number of case studies which have been conducted in order to identify the issues and challenges of spatial data integration are discussed. Then based on identified issues the design and development of a validation tool is discussed. The tool has been designed based on an approach which is presented in the paper. The tool aims to investigate multi-sourced spatial data and identify the items of inconsistency. The tool also proposes available guidelines to overcome the inconsistency. This tool can help practitioners and organizations to avoid the time-consuming and costly process of validating data sets for effective data integration.

Keywords: multi-source spatial data, inconsistent spatial data, spatial data integration

¹ Corresponding author, Tel: ++61 3 8344 0234, Fax: ++61 3 9347 2916, email: abbas.r@unimelb.edu.au
1. Introduction

The amount of heterogeneous multi-sourced spatial data sets is growing dramatically. At the same time demands and requirements of spatial services are expanding beyond framework data sets (National Research Council, 2003). Framework data (including cadastre and topography), single-purpose value-added data sets and collaborative products meet needs of particular spatial services (Buehler, 2003), while many services and applications rely on integrated spatial data (Figure 1). In this regard, multi-sourced data integration has become a significant issue as it adds value to the data sets and ensures effective use of spatial data by broader range of users and applications. This creates many opportunities and possibilities for spatial datasets to be utilized in a variety of spatial services.

![Figure 1 spatial data value chain](image)

Table 1 shows a number of spatial data and their potential sources in the case of NSW state, Australia. Most of these datasets are managed by different custodians. For example, cadastre and topography is managed by local councils and Department of Lands, transportation network by local councils, Roads and Traffic Authority (RTA) and National Parks (Baker and Young, 2005, ASDD, 2006).

<table>
<thead>
<tr>
<th>Spatial data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>cadastre and topography</td>
<td>Local councils and Department of Lands</td>
</tr>
<tr>
<td>natural hazards</td>
<td>Department of Land and Water Conservation, Department of Infrastructure, Planning and Natural Resources</td>
</tr>
<tr>
<td>transport network</td>
<td>Local councils, Roads and Traffic Authority (RTA) and National Parks etc.</td>
</tr>
<tr>
<td>imagery</td>
<td>Department of Lands, RTA, and Department of Agriculture</td>
</tr>
<tr>
<td>vegetation</td>
<td>Department of Land and Water Conservation, National Parks and Wildlife Services, Forests NSW, Department of Defense</td>
</tr>
<tr>
<td>climate and weather data</td>
<td>Bureau of Metrology, Environmental Protection Agency</td>
</tr>
</tbody>
</table>

As shown in the above table, different organizations are responsible for different datasets. Organizations target a particular group of users and develop their own strategies and policies for capturing, managing and sharing data. The diversity of approaches utilized by organizations leads to many technical and non-technical inconsistencies and heterogeneity among datasets which hinders effective spatial data integration. Diversity in data providers/custodians not only results in some technical inconsistency including format, accuracy, currency, datum and access network among data sets, but also results in some non-technical issues such as collaboration, custodianship and licensing arrangements (Mohammadi et al, 2007 and Syafi’I, 2006).
Many spatial data applications and services integrate multi-source data. As spatial data integration is a time-consuming and costly process (Mills and Paull, 2006), minimizing the time and cost of data integration is a priority for these services and organizations (Ting and Williamson, 2000). From acquiring multi-source heterogeneous data to actual integration of data there are few steps which are costly and time-consuming processes.

Practitioners put much effort and time to investigate the data and accompanying documents including metadata to find out the characteristics of the data and inconsistency with other data sets. They sometimes also go through the actual data to identify the characteristics of the data including spatial and aspatial accuracies. Beside this, the assignment of the best solution in form of guidelines or standards is another challenge which should be addressed properly (Figure 2).

Most applications that rely on integrated data suffer from a lack of automatic mechanism to identify the inconsistency among data sets and a mechanism to assign available solutions to overcome the inconsistency. Therefore a tool can sit between data provider and user and takes the responsibility of validating data against the integrability measures (Figure 3).

Based on this approach, the tool provides access to multi-source data and validates data against measures of integrability and facilitates the effective data integration. Based on the requirement of any particular application of jurisdiction, any measurable technical and non-technical
integrability measure can be adopted by the tool. Many technical characteristics of data sets can be investigated and extracted from data and accompanying documents including metadata and privacy policy. This includes format, scale, geographical extent, accuracy (spatial and aspatial), currency, datum and projection system. Also, other information like access method, restrictions on data, pricing policy and availability of metadata can be found to some extent. This provides bases for data comparing and validating.

This study consists of two major streams. One is to identify the issues and barriers associated with effective spatial data integration. This has been done by undertaking a number of case studies incorporating seven countries in the Asia and Pacific region. The second part is devoted to the design and development of a geo-web service tool to assess and validate multi-sourced spatial data based on the issues and challenges which have been found from case studies. The tool also proposes guidelines for particular inconsistency identified among data sets.

2. Spatial Data Integration Issues

In order to investigate the issues which hinder effective data integration, a number of case studies have been conducted in the region of Asia and the Pacific with the support of UN-sponsored UNRCC-PCGIAP committee. The case studies have aimed to investigate the issues and barriers of spatial data integration within the countries with different social, governmental and geographical characteristics. The methodology to investigate the case studies has been illustrated in Figure 4.

![Figure 4 Case study investigation methodology](image)

Seven countries including Japan, Singapore, Australia, Brunei Darussalam, Indonesia, Malaysia and Philippines were invited through the PCGIAP channel to provide reports on integration of multi-sourced spatial data (UNRCC, 2006). They were also asked to give presentations in a workshop on data integration (PCGIAP, 2006). A number of case studies also have been conducted through data sets investigation and visits within two states of NSW and Victoria, Australia to identify the issues of spatial data integration (Mohammadi et al, 2006). Through the
data assessments, visits, reports and the presentations, common potential issues which hinder effective spatial data integration within the context of case study countries have been identified and provide inputs to the design and development of a validation tool for spatial data integration.

The integration report template has been designed as a standardized generic proforma to enable the discovery of information, including matters concerned with member countries’ spatial information policies, laws and regulations, infrastructure implementation, institutional arrangements, technology, and integration issues as well as human resource and capacity building for spatial data integration.

The report consists of two main sections. Firstly, case study countries were asked to provide the explanatory information on following topics:

- Country Context including Geographical Context, Historical Context and Current Political and Administrative Structures
- National SDI Context including History and Status of National SDI Initiative, Historical Outline of Cadastral and Topographic Data Development and Current Administration of Cadastral and Topographic Data
- Institutional Framework for Integration – Data Provider Perspective
- Institutional Framework for Integration – Data User Perspective
- Issues in the Integration of Built and Natural Environmental Datasets including Need for Integration, Major Issues and barriers

Secondly, a questionnaire aimed to identify the main issues of integration in five major categories of technical, institutional, policy, legal and social collected information on individual items. The reports and questionnaire show the importance of solving non-technical issues along with technical issues for effective data integration.

Country reports identified that the diversity in spatial reference systems (datum and projection systems), scales and formats hinders easy data integration and requires time and cost to cope with data preparation. Integration of data models facilitates a greater degree of cross-data set analysis (Mills and Paull, 2007) from both spatial and aspatial perspectives.

International standards have been adopted in most cases, especially standards defined by ISO. All case study countries have adopted ISO (including TC211) as primary standard framework for spatial data coordination and maintenance. It shows the significance of common standard framework for spatial data management. Many of these standards including metadata, quality and spatial reference standards facilitate the integration of spatial data.

Some spatial analysis depend on data sets which monitor features at a certain point of time, hence currency of data becomes a critical issue for mentioned analysis. Completeness of data sets also assists more accurate results; however the latter two issues (currency and completeness) together with issues including geometrical and logical consistency are crucial for integration and also other purposes. But, different datasets with different currency and completeness are more
problematic. Metadata and its content can play a key role in data integration as it can provide information on consistency of data sets. Information on some of the above mentioned issues including accuracy, geographical extent, spatial reference systems etc. are included in most common metadata standards.

Case study visits and studies in Australia consist of investigating data sets from two cities in the state of Victoria and two cities in the state of NSW. In order to have a comprehensive study on the integration practice and the issues, the major organizations in the case study jurisdictions at state and national levels (Geoscience Australia and PSMA) have been visited. State and national level spatial data sets including cadastre, topography, vegetation, administrative boundaries and addresses is also collected and investigated for data integration issues within these regions (Figure 5).

![Figure 5 Four case study regions in Australia](image)

Case studies show despite the importance of technical issues for effective spatial data integration, non-technical issues are of a high significant and should be addressed and investigated simultaneously with the technical issues.

Hence, it was identified that effective spatial data integration does not only require providing technical mechanisms to facilitate integration including the geometrical and topological match of data and a correspondence of attributes (Usery et al., 2005). It also includes the establishment of appropriate institutional, policy, legal, and social mechanisms which is required to facilitate the integration of multi-sourced spatial data. The 17th UNRCC-AP’s Resolution 4 (2006) also insisted and highlighted legal, policy, social, and institutional issues of integration (Figure 6).
Among the case study countries Australia, Philippines and Malaysia (to some extent) has decentralized the spatial data coordination, however they still suffer from inconsistency and barriers in data integration. Australia’s states develop and comply with their own methodologies in data management, which differs from other. This leads to inconsistency among institutional arrangements, policies, spatial data and technical tools. The study of data sets in different Australian jurisdictions showed that data categorization does not comply with a standard pattern. For example, the state of Victoria maintains 1 and 23 layer(s) respectively in Address and Administrative boundary themes, while the state of NSW maintains 11 and 3 layers for Address and Administrative boundary themes. At the same time, Geoscience Australia maintains 4 layers as Administrative boundaries.

Data models are also inconsistent across different jurisdictions of Australia. The state of NSW maintains a data model with different road categories as separate entities, while PSMA maintains road categories as a property of a single road object (Figure 7).
The different levels of maturity from data management perspective, utilizing different funding and cost recovery models, lack of a solid custodianship arrangements are some of the issues which have been observed from Australia’s case study. Other issues are as follows (Mohammadi et al, 2006):

- Utilizing different software
- Different logical data model (road casement as polygon or line)
- Duplication of the same data set (Alexander-Tomlinson, 2006)
- Attribute inconsistency
- Lack of single channel for data access
- Different IP and privacy policies
- Aversion to sharing of data
- Lack of well-developed software to support data integration
- Lack of supporting legislations

The issues and barriers of data integration including institutional, policy, legal and social issues are major concerns, as highlighted in most country reports. Therefore, Effective multi-sourced data integration requires the provision of a number of technical and non-technical mechanisms to overcome the barriers. Effective data integration is not achievable easily unless these prerequisites are identified and coordinated under the holistic framework of SDI. Without the establishment of standards, policies and technical tools in the context of SDI, multi-sourced spatial data integration remains problematic.

Having the capability to adopt diverse standards and propose the solution within different jurisdictions and amendment of the inconsistent data sets to comply with the guidelines and standards can highly invigorate the tool to facilitate data integration and avoid somehow the time and cost of integration.

Among many technical and non-technical issues associated with spatial data integration which have been identified here, some applicable issues have been chosen to incorporate in the design and development of an spatial data integrability tool.

3. **Spatial Data Integrability Tool**

In order to be used in different spatial applications, there are some characteristics which should be considered for an effective spatial data integrability tool. The ability to access remote and local databases and different formats together with complying with open systems and languages is necessary for a tool which aims to facilitate the integration of multi-source heterogeneous spatial data. The diversity of formats and modeling languages and the distribution of spatial data across internet, demands an open web-based tool. In order to meet these requirements web-service architecture was proved to be an appropriate solution (Newcomer, 2002). Geo-web services (GWS) are web applications with spatial interests (Aditya, 2003) which are designed to support spatial interoperability across the web (W3C, 2007).
In order to utilize an appropriate GIS engine which both supports distributed and open databases, a number of available tools including GeoServer (2006), GeoTools (2006) and uDig (2007) have been evaluated. WFS is not stable in GeoServer environment (GeoServer, 2006). GeoTools is an open and free tool which has been utilized by broad range of developers. The uDig is an open source spatial data viewing and editing environment, with special emphasis on the OpenGIS standards for internet GIS, the Web Map Server and Web Feature Server standards. uDig provides a common Java platform for building spatial applications with open source components. The uDig collaborates with GeoTools for WMS and WFS support. Based on the capabilities of the uDig (user-friendly Desktop Internet GIS) in working with different spatial data types and web services and capitalizing on GeoTools capabilities, the uDig was selected.

3.1. System Architecture and Design

The uDig has been developed with a strong emphasis on supporting the public standards being developed by the OGC (OpenGIS Consortium), and with a special focus on the Web Map Server and Web Feature Server standards. One of the most remarkable strengths of the uDig project has been the attempt to leverage as much existing technology as possible in the development process. As a result, uDig has become a case study in a number of the latest modular software development tools in the business community including Eclipse RCP (2007). The uDig inherits its extreme modularity from Eclipse (Garnett, 2005). The functionalities of GeoTools are also utilized in Eclipse application. GeoTools functionalities allow developers to focus on the creation of GIS user interface especially with utilizing WMS and WFS (Figure 4).

![Figure 4 Integratability tool development stack](image)

The Eclipse RCP (Rich Client Platform) is structured around the concept of plug-ins which can only be extended and customized through extending existing plug-ins (Garnett, 2004). In order to facilitate this process, the Eclipse RCP has provided a number of core plug-ins based on which a uDig application can be built up (Figure 5).
The integratability application has been developed by extending the uDig user interface plug-in. (Figure 6).

In the execution of integratability tool, the uDig is responsible for the provision of interface features and running environment, while Eclipse is the development environment (Figure 7).

Running the application within the Eclipse environment provokes the uDig with specific configurations and arrangements which are dictated by plug-in (Eclipse application). The uDig is customized by the configurations and necessary tool and functionalities are set up. The interaction between Eclipse and uDig has been illustrated in Figure 8.
Figure 8 also outlines the main components of the integratability tool as a whole. The application evaluates multi-sourced spatial data sets against a number of measures. The measures can be defined based on the technical and non-technical requirements of the jurisdiction. The aim can be targeted to compare a number of data sets or to validate data sets against a predefined standard measures. In this case a number of technical and non-technical issues including format, metadata availability, spatial extent, accuracy and restrictions on data are selected. The major sources of information on the issues are the metadata and data set itself. A java class has been developed to parse XML metadata and also extract a number of information from data sets. If there is no inconsistency among data sets and there is no restriction on data use, a report is provided to the user with the characteristics of the data sets; otherwise the application identifies the items of inconsistency and proposes available solution for the inconsistency in terms of guidelines or standards. If this is the case, user can amend data and attempt to fulfill the test again.

In order to perform above-mentioned structure, five major classes have been developed within the integratability plug-in including data access (data-accessMethod), metadata parser (metadata-parserMethod), comparison (data-compareMethod), reporting (reportMethod) and display (displayMethod) classes (Figure 9).
Figure 9 Main classes of integratability tool

Data access class is responsible for connecting to the data source and obtains data and metadata if available. This class saves information of data. It also sends metadata information including its source address or URL (Uniform Resource Locator) to metadata parser class. Information on data format and source is also stored by this class, which is passed to the display class if data display is needed. Metadata parser class parses metadata and based on the list of measures extracts corresponding information from metadata. Then extracted information is passed to the data comparing class. This class acquires information from data access class and based on the criteria measures data sets. The outcomes of this class provide information for reporting class to prepare report on items of inconsistency. The reporting class also proposes guidelines to user for further amendments of data. This class provokes display method if there is no inconsistency between data sets. Display class obtains data set source information from data access class and displays data in uDig display environment.

Data access, reporting and display classes contain user interfaces in uDig environment, while all classes contain background codes running in Eclipse environment.

3.2. User Interfaces

The uDig and Eclipse are quite compliant together. Any coding in Eclipse with uDig user interface plug-in extension point is executed in uDig environment. The integratability tool develops a number of steps in order to collect information on data sets, compare data sets, provide report to the user and finally collate data sets.

Collecting Information

The very first user interface of the integratability tool which is run from uDig is a window which collects information on data sets. The window asks the user of the system information on the format, the source of data, metadata standard and place if applicable and finally any restriction that the user is aware of. At this stage user is able to enter as many data sets as required. This preliminary information is saved for each data set and retrieved as required (Figure 10).
The system supports different formats including WFS, WMS, ESRI’s shapefile, images and a few other formats. Based on the format entry, system provides the user with tools to browse data in the local or remote database. The system also proposes some sample web services in the case that user has chosen WFS and WMS.

Another part of the system asks about the metadata standard for XML metadata. At this stage only XML metadata is readable by the system and other formats such as text and html is not compliant with the integratability tool. There are five major metadata standards including ISO/TC211, CEN/TC287, Australia and NZ (ANZLIC), Europe (DCMI), US (CSDGM) which have been fed to the system and the system is able to match any metadata standard with the corresponding schema. The schemas have been developed and customized base on the information required for the integration purposes. These steps are handled by \textit{data-accessMethod} class. The information collected at this step is saved for next steps.

\textbf{Comparing data sets}

Both actual data and metadata contains information which is required for data integratability assessment. Much information including geographical extent, format, attribution and datum can be collected from actual data. Depending on its standard, metadata also contains much information on data including custodianship, jurisdiction of data, completeness, currency, any restrictions on data, pricing, scale and attribute and spatial accuracy (Figure 11). These characteristics are essential for data integration; therefore the comparison of the characteristics results in the ability of data sets for integration. The
In order to utilize above-mentioned items for integratability assessment, *metadata-parserMethod* class parses the XML metadata and extracts necessary information. This task is carried out by utilizing XML schemas which have been developed based on the required information and metadata standards. The information then is input to the *data-compareMethod* class. This class also extracts information from actual data and compares them to identify the items of inconsistency. The result of this step is passed to the reporting class to form the report.

**Reporting to the user**

The results of comparison forms a report with the list of data sets characteristics. The items of inconsistency also are highlighted in the report. The report also provides the user with some guidelines or documents to remove the inconsistency. The *reportMethod* class is responsible to form the report (Figure 12).
User can amend data sets based on the guidelines and remove inconsistencies. The uDig offers edit and view tools, users can alter and view data. If there is no inconsistency left and there is no restriction on data use, data sets can be collated and displayed.

**Data collation**

The displayMethod class receives formats and sources for homogeneous data. This class possesses methods to read and display data in different formats. The uDig provides display and basic GIS tools. Figure 13 shows an example of data collation as the final step of integratability tool.

![Figure 13 Information collection (uDig interface)](image)

**Discussion**

Overall, the development of integratability application within the context of SDI, provides a strong tool for data suppliers and users where investigation and comparing data sets is their major task. The customizability of the application creates more opportunities for utilization of the application where standardized data assessment is required. Instructions and guidelines prepared within the context of SDI also provide an appropriate basis for those who want to amend data in a standardized manner.
For organizations and jurisdictions, the integratability application can also indicate the degree of successfulness in data interoperability and integratability.

**Conclusion**

Spatial data integration is a common task carried out by many spatial services. Despite much time and costs associated with data integration, effective spatial data integration is very difficult if not impossible in many cases, especially at a National level, where there are many stakeholders involved. Spatial data integration is problematic and associated with many technical and non-technical issues which hinder effective data integration at different jurisdictional levels. Therefore a holistic framework is essential to consider all associated issues and identify and solve the problems.

This framework however, can also be utilized by spatial data stakeholders to develop institutional arrangements, legal and policy tools and also social capacities to facilitate the integration of multi-sourced spatial data so that it is used to its maximum potential.

Capitalizing on interoperable tools and XML technology, integratability tool is able to assist users to assess, validate, edit and integrate spatial data sets. This tool can incorporate both technical issues and specifications and guidelines to provide a holistic system for effective data integration. Integratability tool can be utilized not only to integrate multi-sourced spatial datasets but also to assess the integratability of these datasets. The prototype tool acts as an intermediate tool between user and data provider and facilitates the evaluation and identification of inconsistency items. The tool also evaluates the integratability of multi-sourced spatial data against set criteria and provides instructions and guidelines to amend data based on these criteria. Based on the requirements of respective jurisdiction or organization, the tool can be customized to accommodate different measures. This tool can also help users to prepare datasets before integration and also can assist practitioners to develop the required guidelines and specifications to be used by data users and providers to prepare data prior to use.
References


