

A framework for the management of spatial data quality information

Naveen Kumar Sidda
January, 2009

A framework for the management of spatial data quality information

by

Naveen Kumar Sidda

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: (Geo-informatics)

Thesis Assessment Board:

Chairman : Prof. Dr. Ir. Alfred Stein

External Examiner : Dr. P.S. Acharaya

IIRS Member : Mr. P.L.N. Raju

Supervisors I IRS : Dr. Sameer Saran
ITC : Dr. Ivana Ivanova



**INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION
ENSCHEDÉ, THE NETHERLANDS**

&

INDIAN INSTITUTE OF REMOTE SENSING , NATIONAL REMOTE SENSING CENTRE (NRSC)

DEPARTMENT OF SPACE, DEHRADUN - INDIA

Disclaimer

This document describes work undertaken as part of a programme of study at the Indian Institute of Remote Sensing and International Institute for Geo-information Science and Earth Observation. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Abstract

Metadata conveys significant information regarding the assets of the spatial datasets. Metadata is of textual information that follows some industry specific terminologies. There are colossal metadata standards that used to describe a product to the consumers. However, consumer still feels that the quality information provided could not able to cater their needs. It could not able to elucidate on the datasets to choose for their intended applications. The effective communicating of the data quality still remains as an issue which needs quantum leap.

However, there are some techniques proposed to communicate the quality information at different levels in detail. By integrating, the concepts of data warehousing and Geo-spatial Information Systems (GIS), users are able to comprehend on the data quality and dataset in a rapid and intuitive way. By these technologies, proper communicating of the data quality information will avoid the risk of misusing of the data.

This research presents a data quality tool which integrates two different domain concepts: Geo-spatial information systems and Online Analytical Processing (OLAP) in a web enabled environment. The tool helps the user in communicating the data quality of a dataset both in spatial and non-spatial graphical format. The tool communicates the quality information through visual explanations which will aid the user in getting insight about the dataset. The tool structures and records the data quality at various levels of details. The research also led to compare the quality report from the tool to web based metadata from Geonetwork.

Key words: Data quality, metadata, ISO standards, databases, dashboard reports, SOLAP.

Acknowledgements

This thesis owes its existence to the help, support and inspiration from many people. In first place I would like to express appreciation and gratitude to my supervisors Dr.Ivana Ivanova, Geo information processing, ITC for her extra wisdom and guidance from the very first day of the meeting till the last day of the final defence and Dr.Sameer Saran, Geo informatics, IIRS for his incredible suggestions, keeping me sane amidst the stress-infested thesis days and for keeping my mind at peace.

I'm also indebted to Mr. PLN Raju, In-charge Geo Informatics and Mr. Ram Mohan Rao, system administrator for their sharp and constructive criticisms.

I'm profoundly grateful to Mr. Ravi Shankar, Mr. Shiva Charan and Mr.Siddarath, software professionals for their expert suggestions and cooperation that contributed substantial to the work.

Very special thanks to all my classmates for their support and assistance which brought the whole project together.

I would like to extend my thanks to Dr. V.K Dadhwal, Dean IIRS for giving me permissions to go ahead with the thesis.

Finally, I take the opportunity to owe sincere gratitude to my family for continuous and unconditional support of all my undertakings and scholastic.

I dedicated this thesis to my parents.

Table of contents

I. Introduction.....	1
■ Motivation and problem statement.....	1
■ Research identification.....	2
■ Research objective.....	2
■ Research questions.....	2
■ Innovation aimed at.....	2
■ Research framework.....	2
■ Research report structure.....	3
II. Background and literature review.....	4
Spatial data quality.....	4
■ Introduction.....	4
■ Data quality parameters.....	5
■ Characteristics of metadata.....	8
■ Metadata standards.....	8
■ Limitation of the current data quality information.....	10
■ Spatial data warehousing.....	11
■ Introduction.....	11
■ GIS and data warehouse.....	11
■ SOLAP.....	13
■ Conclusion.....	15
III. Research methodology.....	16
■ Data preparation.....	16
■ Data quality parameters according to ISO/TC 211.....	17
■ Storing in the databases.....	17
■ Creation of quality cube.....	18
■ Generating quality report.....	18
■ Web interface to database.....	18
■ Preparing Metadata report in Geonetwork.....	19
■ Evaluating the two quality reports.....	19
■ Recommendation on the method of data quality information.....	20
■ Technologies used.....	20

IV. Results and Discussions.....	22
V. Conclusion and Recommendations.....	33
VI. References.....	35

List of figures

- Figure 2.1 (left) The problem associated with communicating the spatial metadata.(right) ideas for better communicating the spatial metadata.
- Figure 2.2 Star schema
- Figure 2.3 “Data” spatial dimension and “quality” non-spatial dimension with different level of granularities
- Figure 2.4 GIS and BI
- Figure 3.1 The overall methodological flow-chart
- Figure 3.2 Conceptual design of database
- Figure 3.3 Web interface of the data quality tool
- Figure 4.1 Web interface of the tool
- Figure 4.2 Window displaying a message regarding the database
- Figure 4.3 & 4.4 Log files for both the base and the reference data
- Figure 4.5 Quality cube along data and quality dimensions
- Figure 4.6 Quality cube presenting more details on the dataset and the quality
- Figure 4.7 Quality dashboard interface displaying the quality of overall dataset and option to get the layer wise reports.
- Figure 4.8 Detailed quality dashboard reports for the line layer.
- Figure 4.9 Reports exported to excel and PDF format.
- Figure 4.10 Interface of the Geonetwork
- Figure 4.11 Geonetwork interface displaying the ISO 19138 template details.
- Figure 4.12 Interface which has provides an option to accesses data quality tool
- Figure 4.13 Interface displaying the data quality and the rest of the information
- Figure 4.14 Interface displaying thumbnail of the data and details on the selected view
- Figure 4.15 The XML schema view

List of tables

Table 1.1 Thesis structure

Table 2.1 Sources of error

Table 3.1 Data quality elements and sub-elements

1 Introduction

Over the years, there has been a drastic increase in usage of spatial data. “Data about positions attributes and relationships of features in space are often termed as spatial data” (Morrison, 1995). Spatial data quality is defined as the concept of ‘fitness for use’ (Chrisman, 2006). Some elements of the spatial data quality are the positional accuracy, temporal accuracy, attributes accuracy, logical consistency, completeness, lineage and semantic accuracy (Morrison, 1995). These are produced and used by various organizations for numerous applications (Devillers et al., 2007b). Public have easy access to spatial data through various means, e.g. Google maps and Google Earth. For instance Google Map Maker is a new service that allows users to edit and contribute map information like draw, add map features. This increase in mass consumptions and production of spatial data has its related issues of which maintaining the quality is one of the significant issues.

The dataset is usually described by the metadata. The objective of metadata is to allow the users to judge the fitness of a dataset for their use. Metadata hides most of the information and it is insufficient to judge the quality of a dataset (Devillers and Beard, 2006). Indeed the user gets only textual information from the metadata. So, this calls for a new concept, Spatial On-line Analytical Processing (SOLAP) which aids the users in selecting the appropriate dataset (Devillers et al., 2005). SOLAP is the spatial extension of OLAP, Business Intelligence (BI) tool (Rivest, S. et al., 2005). It helps in making spatial decisions. It is introduced to resolve the limitations of the Relational Database Management Systems (RDBMS) and Geo-spatial Information System (GIS). It helps to analyse the spatial data at different levels of data granularity. Multidimensional databases are one of the data warehousing concepts used for efficient way of storing and analyzing the data (Kimball and Ross, 2002). This research proposes a framework for the management of the quality information using the concepts of multidimensional databases and the OLAP tools. This approach may be advantageous in comparing to the traditional metadata.

1.1 Motivation and problem statement

Spatial data is being extensively used for diverse applications by various experts and non-expert users (Oort, 2005). The massive increase of spatial data has raised many serious issues of which providing meaningful data quality information has become a more predominant factor. Users need a simple, quick way of communication rather than the textual metadata (Tóth and Nunes de Lima, 2005). Research is done on spatial data quality but still significant work has to be done for better quality evaluation and visualization, using latest technologies (Sonnen, 2007).

The spatial data integrated from various sources should be checked and its quality be reported to the end users who in-turn take decision whether the dataset suits to their application, i.e. fitness for the purpose. Does it really serve the purpose? The word ‘Metadata’ still is a chaotic term for many spatial data experts (Boin and Hunter, 2006). Conventional metadata accompanying the dataset is not fully informative to a non-expert user (Devillers and Beard, 2006). Hence, a method and model is required to effectively process and communicate the quality of the dataset that is stored in a spatial database. The spatial database used to be just a collection of information stored in a structured format. At present it has transformed in to geo- information system where one can make decisions (Devillers et al., 2002; Van Oosterom and Lemmen, 2001).

The Data warehousing technology is appropriate to many analytical problems and for datasets lacking in proper hierarchical structure (Kimball and Ross, 2002). This technology helps to store, handle,

query and analyze large amount of complex, structured spatial information in the multidimensional databases. This information could be communicated to the end-user via dashboard, maps and tables for visualizing the quality of a dataset(Bédard et al., 2005; Parmanto et al., 2008).

1.2 Research identification

Integrating data warehousing concepts with GIS in assessing and communicating the quality of a spatial dataset and the need for spatial extensions to OLAP functionality(Bédard et al., 2001).There is no guidance on how to structure and effectively present the data quality information to the consumers(Boin and Hunter, 2006). This can be achieved by conflating two different domains GIS and data warehouse in proposing a data quality model (Devillers et al., 2007b) . To quantify and communicate the quality of the dataset at different data granularities, rather than the ordinary textual metadata.

1.3 Research Objective

To organize and communicate the quality information of a spatial dataset, at different levels of details to the user using the concepts of dashboard and SOLAP tools in a web environment.

1.4 Research questions

The following three questions have to be answered to meet the research objective.

- a) How the data quality information has to be stored and organized in accordance with the database design principles?
- b) How the SOLAP and dashboard concepts support the communication of data quality to the user?
- c) What makes the use of SOLAP and dashboard tools different from communicating spatial data quality using metadata?

1.5 Innovation aimed at

To build data quality tool that stores and communicates the quality at different level of details to three different kinds of users i.e. naïve users, expert users and computer machine, using the concepts of Dashboard and SOLAP tools in a web environment. To evaluate the data quality tool report built to that of textual Metadata generated from Geo-network.

1.6 Research framework

Three investigations are performed to fulfil the above stated objective and research questions.

1. **The first investigation** is of preparing vector data and the required data quality elements and sub-elements according to ISO 19100 series of standards.
2. **The second investigation** is how to populate the database with the data quality information. To analyze how the SOLAP and dashboard tools effectively communicate to the end users in the form of map, table and dashboard in a web environment. In this investigation three users will be addressed i.e. naive users, expert users and computer machine.
3. **The third investigation** is comparing the report generated by the data quality tool to the report generated by Geonetwork tool, eventually, to recommend on the method for the data quality communication.

1.7 Research report structure

Table 1.1 Thesis structure

Chapter number	Chapter name	Description
1	Introduction	An introduction to the study, the research objectives and research questions.
2	Background and literature review	Background information ,literature survey on the topics spatial data quality and spatial data warehousing.
3	Research methodology	Methodology and technologies used to achieve the research objective and questions.
4	Results & discussions	The overall results obtained and discussions on the results.
5	Conclusions	Final conclusions on the thesis.

2 Background and literature review

2.1 Spatial data quality

2.1.1 Introduction

Spatial data has witnessed two major revolutions 1) digital mode 2) mass consumption of low-cost data, over the last two decades (Bedard et al., 2004). Spatial data collection is very expensive and producers can only afford to collect once (Maguire, 2004). As data collected once will be used for many times; hence, proper care should be taken while collecting or producing. When data is produced it is necessary to show the detailed specification of the product like the organization, quality details etc. To utilize the spatial data effectively, one should be aware of the quality of the dataset. As knowledge of the spatial data quality is often very important, as different users will have different quality requirements (ISO/TC211, 2006). Quality information should answer questions like “What is the level of the data quality?”, “whether the data can be used for our application?” etc. It is defined as “Quality is ability of a set of inherent characteristics of a product, system or process to fulfil requirements of customers and other interested parties” (Jakobsson, 2002).

Data quality is an important component in any GIS applications (Srivastava, 2008). In GIS applications, data from different sources, collected by using various techniques, with different level of quality can be integrated (Servigne et al., 2007). Indeed these errors get inherent in spatial database which determines the quality. These errors should be documented and dealt properly. The results that we obtained from GIS are apparently used for taking various decisions (Makela, 2007). These decisions are greatly influenced by the quality of the data-sets. Flawless dataset will lead to take a right choice regarding the data to avoid further complications. Identifying and assessing the errors are not only the factors to be considered. Conceptualization, developing, utilizing and maintaining the spatial database as well should pay attention. In order to maintain a flawless dataset, there should be continuous check on the process of data collection, data input, data storage, data manipulation, data conversion and quality control procedures (Tóth and Nunes de Lima, 2005). These are the various stages of error.

Table 2.1 Sources of error (Cockcroft, 1997)

Stage	Sources of error
Data collection	<ul style="list-style-type: none"> * Inaccuracies in field measurements * Inaccurate equipment * Incorrect recording procedures * Errors in analysis of remote sense data
Data input	<ul style="list-style-type: none"> * Error in digitizing * Nature of fuzzy natural boundaries * Other forms of data entry
Data storage	<ul style="list-style-type: none"> * Numerical precision

	* Spatial precision (in raster systems)
Data manipulation	* Wrong class intervals
	* Boundary errors
	* Spurious polygons and error propagation with overlay operations
Data output	* Scaling
	* Inaccurate output device
Use of results	* misunderstanding of information
	* Inefficient usage of data

The spatial data quality is a major issue from both the data producer and users prospective. From producers point of view it is defined as ‘confirmation to the specifications’ from user point of view it is fitness for use. Data producers about the credential of their dataset where as the users about the availability, cost and quality levels(Yang, 2007). Data producers have to be meticulous about the different stages of the error as shown in table 2.1. Now-a-days data producers do not only imply to statutory bodies. Even common public started creating spatial data through Keyhole Mark-up Language (KML). The following are the reasons for the spatial data quality to be a major concern. Firstly GIS is a decision making tool for many applications. Secondly, spatial data is now extensively used and produced by the private sectors and common public as well. Finally, these spatial are easily accessible. The user should be well comprehended about the dataset in order to take the right decision.

Data quality assessments (Srivastava, 2008):

- The data producers are responsible for the first level of assessment. This assessment is of the data quality check based on the data specifications.
- Users do the second level of assessment. They give relevant feedback on the product, based on those feedback data has to be rectified.

Essentially there are few issues in assessing the spatial data quality (Oort, 2005).

- Evaluation of the spatial data quality is very expensive and tedious process.
- Legal issues.
- No proper algorithms or tools that can cater the needs.
- Users need has to be focused and awareness of the spatial data quality.

2.1.2 Data quality parameters

Data quality parameter provides an insight to the user on the dataset fitness-for-use. Essentially the data quality of a particular dataset is described by the data quality elements and sub-elements(Giversen, 2007). The quality parameters are completeness, logical consistency, positional, thematic and temporal accuracy(ISO/TC211, 2006). Depending on these elements the user takes decisions whether data is of relevant to their application.

Completeness

Completeness is number of committed and omitted objects, their attributes and relations in the dataset to the reference. Completeness is of feature completeness, attribute completeness and value completeness(Yang, 2007).The quality measure could of Boolean, integer, percentage or ratio depending on the method of evaluation.

Data quality sub-elements(ISO/TC211, 2006; Surveyofkenya, 2007):

Commission is of extra committed objects in a dataset.

Example: If there are 20 lakes present in reference dataset, and there 25 lakes existing in our dataset, then the quality is “commission 5%”.

Omission is the omitted objects in a dataset.

Example: If there are 100 buildings in reference dataset, and there 95 buildings exist in our dataset, then the quality is “omission 5%”.

Consistency:

Consistency is of four types: conceptual consistency, domain consistency, format consistency, topological consistency. The data quality measure of consistency could be of boolean, integer, ratio, and percentage (ISO/TC211, 2006).

“Logical consistency refers to the degree of adherence to logical rules of data structure, attribution and relationships”.

Data quality sub-elements(ISO/TC211, 2006; Surveyofkenya, 2007):

“Conceptual consistency adherence to rules of the conceptual schema”

Example: If the dataset is not consistent with the application schema, then it is conceptual consistency error.

“Domain consistency adherence of values to the value domains”

Example: If the attribute value ranges from 1 to 5. But if the value does not fall in the range, and then it is a domain consistency.

“Format consistency degree to which data is stored in accordance with the physical structure of the data set”.

Example: If the data format is of ESRI shape file and if does not contain database file (.dbf file), then it is format consistency.

“Topological consistency is the ratio of items out of rules of topological characteristics like overlapping, undershoot, overshoot and contain”.

Example: number of overshoots and undershoots.

Accuracy

Accuracy is one of the primary components in the data quality. Accuracy can further subdivided into: accuracy of attribute values spatial and temporal references (Cockcroft, 1997).

Positional accuracy

Positional accuracy refers to the accuracy of the spatial component of a database (Giversen, 2007). It shows the deviation of geographical feature location in a dataset to its ground truth. It is how well the true measurements of a object on ground match with same object in the database. It also relates to the relative or absolute positional accuracy of the features. The data quality measure for this element is error statistics i.e. Root mean square error RMSE(ISO/TC211, 2006).

Data sub-quality elements(ISO/TC211, 2006; Surveyofkenya, 2007):

“Absolute accuracy is the closeness of the reported coordinate values in a dataset to the values accepted as or being true”.

Example: If the RMSE of co-ordinate value of our dataset from the value of reference dataset is 0.45, then the “Absolute accuracy is 0.45m”

“Relative accuracy is the closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true”.

Example: The difference between relative distances of two control points of reference dataset from the control points of our dataset is 0.27m, “then the relative accuracy is 0.27m”.

Example:

“Gridded data positional accuracy is defined as closeness of the gridded data position values to the values accepted as or being true”.

Example: RMS error of the TIN grid point elevation values and value of the reference dataset is 1.2m then the “gridded positional accuracy is 1.2m”.

Temporal accuracy

Temporal accuracy is of accuracy of time measurement, temporal consistency and temporal validity. Temporal accuracy is correctness of temporal reference of an item. The quality measures are generally of Boolean values or error statistics.

Data sub-quality elements(ISO/TC211, 2006; Surveyofkenya, 2007):

Accuracy of time measurement

“Difference between time attribute recorded in our dataset to the reference dataset”.

Example: If there are 6 days difference between data of construction completed to the recorded data of construction completed in our dataset. ‘Accuracy of time measurement is of 6 days’.

“Temporal consistency is correctness of ordered events”.

Example: If the data of demolition is earlier than the construction data, then it is temporal consistency error.

“Temporal validity is the validity of data with respect to time”. It is treated with the same data quality measures used for domain consistency.

Thematic accuracy

Thematic accuracy is the accuracy of either spatial or thematic attribute of feature.

Data quality sub-element(ISO/TC211, 2006; Surveyofkenya, 2007):

“Classification correctness indicates the correctness of classification items.”

Example: If a rail is classified as road, then it is a thematic error.

“Non-quantitative attribute accuracy indicates the correctness of non-quantitative attributes.”

Example: If 5% road names of our dataset are incorrect to the reference dataset road names then the “Non-quantitative attribute accuracy is 5%”.

“Quantitative attribute accuracy is the accuracy of quantitative attributes.”

RMS error, by comparing the attribute “length” in the dataset to the length in the reference dataset, if it is 10m then the “Quantitative attribute accuracy is 10m”.

2.1.3 Characteristics of metadata

It is typical to have dataset with no metadata at all. Metadata acts as a catalogue for the user to choose the datasets. Essentially metadata conveys information regarding identification, spatial extent, quality, spatial and temporal schema, spatial reference and distribution of digital geographic data(Danko, 2000).This is applicable to data series, datasets, individual feature objects. Knowing what information about the quality of spatial datasets is available will be great benefit to the users of spatial data. Apparently, data processing, duplication of data collection is tiresome and expensive. Hence, documentation about the data has to be maintained to provide such information to the users. The core metadata provides nature and content of the data resources. Metadata as well exists at different spatial hierarchy i.e. having different data granularity data series, dataset, layer, feature and feature instance level(Devillers et al., 2007a). The producer has to spend dedicated amount of time and resources to achieve these kind of metadata quality at various levels.

For different data types like vector, raster at various levels requires various forms of metadata to be documented. Essentially metadata must be well-structured and should in reliable format for search, retrieval, metadata exchange and presentation. There are three types of metadata that has been recognized in the implementation form (within the database), export or encoding form and the presentation form(Yang, 2007).Perhaps sometimes producing metadata can be more expensive than the data production(Devillers et al., 2007a). Some of the open source tools for editing/producing metadata are available such as: Geonetwork(Geonetwork, 2008) and Map Window GIS(Map Window GIS, 2008). These facilities are also available in some of the commercial software’s.

2.1.4 Metadata standards

Today, the spatial data is produced from many sources. Often capabilities and resources of the data are unknown. Making the problem more complicated, to avoid complications metadata has to be prepared(Sonnen, 2007). “Metadata is often defined as data about data”(Boin and Hunter, 2006) . It is an essential information to assess the dataset fitness-for-use(Goodchild et al., 2002). It is very important for both the data user and producers. It contains the data quality information which plays significant role in describing the quality of the dataset. Metadata systems allow the users to search the metadata for the datasets located on network.

Different organizations have their own standards for documenting data quality information and metadata. There various data quality standards developed by international, regional and national organizations such as(GSDI, 2008; Servigne et al., 2007):

International level

- International Organization for Standardization of Geographic information/Geomatics ISO/TC 211
- International Civil Aviation Organization (ICAO)
- International Geosphere - Biosphere Program (IGBP)
- Open Geospatial Consortium (OGC)

Regional level

- Infrastructure for Spatial Information in Europe (INSPIRE) - Europe

- ICIMOD's Approach towards a Regional Geo-Information Infrastructure (RGII) In the Hindu-Kush - Himalayan (HKH) Region
- ANZLIC - The Spatial Information Council - Australia/New Zealand
- Antarctic Spatial Data Infrastructure - Antarctica
- Digital National Framework (DNF) - United Kingdom
- Central American Geographic Information Project (PROCIG) - Central America

National level

- RAVI - the Netherlands
- National Spatial Data Infrastructure (NSDI) -India
- National Spatial Data Infrastructure (NSDI) - United States
- German Umbrella Organization for Geoinformation - German
- GI Gateway - United Kingdom
- Am/Fm Geographic Information System - Italy
- Australian Government Spatial Data - Australia
- National Geographic Information - Finland

ISO 19115: Geographic information, specifies the conceptual structure of metadata. Describes the different quality components of a datasets(ISO/TC211, 2006).This defines the metadata elements, sections, and optional elements, elements essentially support the four major uses: discovery of data, determining data fitness use, data access and use of data(Yang, 2007).ISO 19115 defines more than 300 metadata elements that includes, relation ,classes and attributes(Jakobsson and Giversen, 2007). They are two mandatory packages i.e. Metadata entity set information and identification information(Yang, 2007). Data quality information contains the general assessment of the quality of the dataset. It defines the core metadata of the datasets as well. ISO/TC 211 defined the data quality parameters namely completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy, overview elements like purpose, lineage and usage as well(ISO/TC211, 2006).ISO 19138 defines information necessary to describe the quality indicator and descriptions of the list of quality measure. Users are allowed to define their own indicators.

2.1.5 Limitation of the current data quality information

The current metadata standards (e.g. ISO standards) provide comprehensive, clear definition for the data quality elements, sub-elements, their evaluation techniques. Apparently there are numerous metadata standards and guidelines to describe a product. The quality section of the current metadata gives just the basic quality information regarding the datasets(Devillers et al., 2007a). In spite of all these standards we can still say that the metadata has still some deficiencies as shown in figure 2.1. The user could not elicit much information from the metadata. There are few caveats that the producer has to look upon(Yang, 2007).

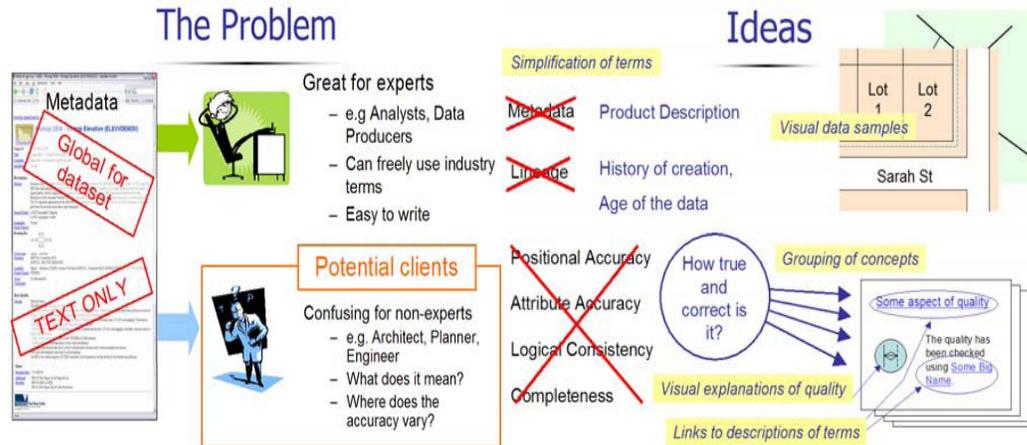


Figure 2.1 (left) The problem associated with communicating the spatial metadata. (right) ideas for better communicating the spatial metadata. (Hunter et al., 2005)

Objects in a spatial dataset are organized in hierarchical way. Whereas the, metadata lack the means to communicate the information hierarchical. Though it is possible, but it is hard to find means for users to navigate information hierarchical. Integration of spatial data and quality information is a second problem for the current data quality information. Current metadata standards stores the quality information in separate files, limiting effective use of dynamic operators. It would be beneficial to the end user if we could show the data quality visually appealing instead of textual information. As the data content changes dynamically there is difficult in updating the unstructured text. The other main factor is the lack of implementation leading to the inconsistency of the data which further puts the user in a chaotic condition. Goodchild (2002) says that the current standards give more priority to the data producers than the users. Eventually the user is incapable to choose the dataset.

Because, of the above limitations with the communication of the data quality information. The end user cannot be fully comprehended on the quality of a dataset. Hence, there is an imperative guidance to all three users: naive users-who are not from GIS domain, expert users-GIS experts like the researchers who want to use the dataset for their analysis. They may need to complete understand the assets of the data set, computer machine which uses the database to query regarding the quality of the dataset.

2.2 Spatial data warehousing

2.2.1 Introduction

The notion of data warehouse (DW) which belongs to Information technology, in recent past it got evolved in GIS market as well. Integrating the functionalities of databases to the cartographic software's (Morehouse, 1985). As GIS and Business Intelligence (BI) were brought up in two separate IT environment. The potential benefits of these two industries had started coming into limelight. Data warehousing is defined as special kind of tool for data-structuring used in business intelligence field, in conjunction with Online Analytical Processing (OLAP) techniques (Codd et al, 1993; Databeacon, 2004). The OLAP mechanism has immense rich capabilities to analysis historical datasets for decision making process. This phenomenal capability is very much useful for the spatial data analysis. The interaction between GIS and data warehouse opens up a new set of analytical possibilities. This distinct combination offers a wide ranging set of research possibilities.

2.2.2 GIS and data warehouse

Recently, interfacing GIS and DW turned to web domain. This gave rise to emerging metadata and web-services standards. These combination in a dynamic web environment has a wide set of research possibilities. Indeed there are two kinds of databases one the ‘analytical databases’ and the other is ‘transaction’ or ‘production’ databases(Healey and Delve, 2007). This transaction databases are used for Online Transaction Processing (OLTP) to store the co-operate data. OLTP is for the clerical community for instance banking sectors. The database is continuously updated and is normalized for optimum performance. Its does not keep the track of the historical data, whereas the, OLAP technology maintains and analysis the historical data in a database. Data warehousing is aimed for the decision making community-managers, business analysts. ”Data warehousing is a relational database that is designed for query and analysis rather than for transaction processing, the characteristics are subject oriented, integrated, time-variant and non-volatile ”(Parida, 2005).Data are not updated on a transactional basis. Snapshots of data are added to the warehouse at regular intervals. The basic component of data warehousing are operational systems, data staging area, data presentation area, and data access tool (Kimbell and Ross, 2002).

The structure of the data warehousing comprises of fact table and dimension tables (Healey and Delve, 2007). They provide aggregation capabilities for further analysis. The fact table which consists of numerical data i.e. is the measure and the foreign keys to connect to the dimension table. A schema is defined as fact table with finite number of dimension (Malinowski and Zimányi, 2007). A fact table surrounded by dimension table is defined as a star schema which is the simplest style of data warehouse schema (Kimball and Ross, 2002).There is also snowflake schema which consists of more complex dimension tables with the normalized data in it. The concept behind this is de-normalization is to enhance performance and to speed-up queries. There is one more schema called the starflake schema it is essentially the integration of both the star and snowflake schema(Healey and Delve, 2007). A mixture of both normalized and de-normalized dimension tables. Usually bitmapped indexes are employed in construction of data warehouse. As it return queries that contains large number of rows from fact table linked to the dimension table (Healey and Delve, 2007).

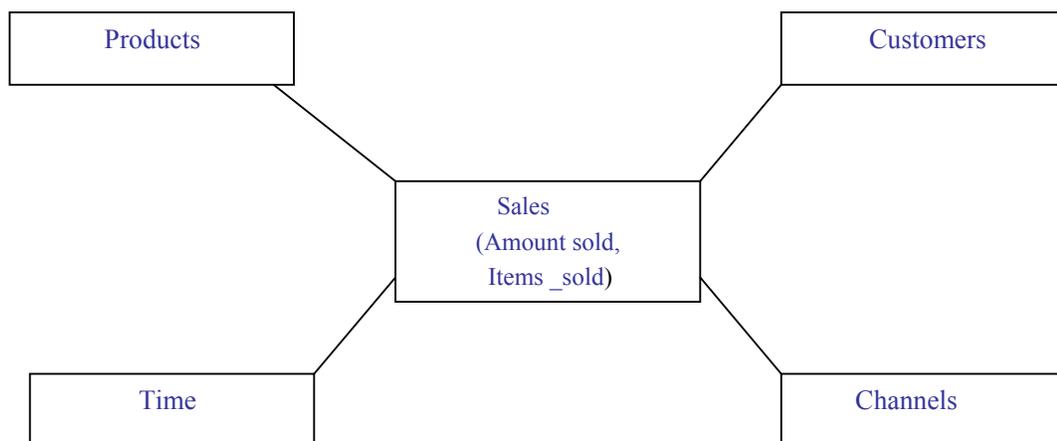


Figure 2.2 Star schema(Parida, 2005)

There is much focus on to the finished maps rather than the underlying data (Maguire, 2004). Since 80% data stored in databases have spatial or location component (Dube et al., 2007). The location dimension are widely integrated in data warehousing and OLAP concepts. Conventional database cannot handle the spatial data, this led to the concept of spatial databases and spatial data warehousing (SDW). SDW is combination of data warehouse and spatial databases technologies (Malinowski and Zimányi, 2007). They have additional support to handle the spatial data. Additional supports like store the spatial data, indexing, querying etc. Other than SDW few other concepts has also been cropped up like SOLAP, spatial data mining, spatial statistical analysis. SDW are very much an emerging field.

Multi dimensional databases (MDDDB) or the datacubes which essentially consists of dimensions, members, facts and measures, as shown in figure 2.2. A cube which is a extension of GROUP BY function. Operations like drill down, roll-up are performed on hierarchies in dimensions. Roll-up is an aggregation carried along dimension hierarchy. Drill-down to reach down to the bottom of a dimension hierarchy (Kimball and Ross, 2002). Similarly slice/dice, roll-up operations are performed on datacubes. These operators does not require any pro-efficiency in databases languages which give instantaneous answers (Devillers et al., 2007b).

OLAP tools are to carry out advance querying on relational databases, MDDDB and data warehouses (Codd et al, 1993). They are quite few number of OLAP tools in the market. Relational OLAP (ROLAP) for the relational databases. Multi-dimensional OLAP (MOLAP) is designed for the multi-dimensional databases (Fu and Hammer, 2001). Indeed OLAP operations and functions substitutes the standard complex decision support SQL queries. They are different type of functions used for the aggregation: distributive, algebraic, and holistic (Shekhar et al., 2000). Distributive functions like sum, min and count. Algebraic functions like average, variance and standard deviation. Holistic functions like median, most frequent and rank. For the spatial data aggregate functions like geometric union, geometric intersection are used. Algebraic functions like centre of gravity, while for spatial holistic functions like equi-partition or nearest-neighbour index.

In data warehousing and OLAP, multidimensional databases model has been recognized as well suited model for the decision making. Further the advantage of representing the spatial data as well been well recognized. Besides the above functionalities there are few caveats with OLAP and data ware house technologies. Problems like real-time monitoring or tracking problems, data that lack hierarchical structure (Healey and Delve, 2007). Justifying the cost involved in implementing data warehousing infrastructure is a major concern.

2.2.3 Spatial On-line Analytical Processing (SOLAP)

Multidimensional databases name itself shows that it can manage information at different dimensions and levels. Multidimensional database are the component of data warehouse used in OLAP (Kimball and Ross, 2002). These essentially support for the organizations to take decisions. OLAP is defined as “a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user” (compinfo, 2004). OLAP tools essentially support the fast retrieval of the data and the management of multidimensional data. OLAP tools rely on the multidimensional database structure also popularly known as data cube or hypercube. These provide intuitive navigation into detailed information using various operators like roll-up, drill down, drill across, slice and dice according to the dimension need to manipulate (Kimball and Ross, 2002). We can manipulate the thematic, temporal and spatial dimension just by click of mouse instead of using SQL-like queries (Spatial OLAP, 2008).

The above all functionalities of OLAP lead to the spatial extension SOLAP. It is the confluence of two different subjects GIS and Business Intelligence (BI). It is an efficient decision support system in spatial environment known as SOLAP (Devillers et al., 2007b).”SOLAP is a visual platform built especially to support rapid and easy spatial-temporal analysis and exploration of data following a multidimensional approach comprised of aggregations levels available in cartographic display as well as in tabular and diagram display”(Matias and Moura- Pires, 2005). SOLAP tools are user-driven, allow interactive exploration of the datacubes(Healey and Delve, 2007). SOLAP provide facility for the complex analysis and visualization operations for the spatial decision support system. By interacting at various levels of granularity including spatial and temporal granularities (Zhang et al., 2005).In recent past SOLAP tools have been introduced into public/private organizations.

GIS systems are not well adopted for decision making as they are often used at operational level(Rivest et al., 2005). With the intuitive and user-friendly interface SOLAP tools have high level of data interaction which is not present in GIS(Healey and Delve, 2007). For instance, spatial data quality information can be organized in multidimensional database at different level of granularity(Devillers et al., 2007b).These results can be visualised and analysed through some of the dashboard and cartographic techniques. Which otherwise, it would be difficult to show the data quality at different levels. Essentially the visualization techniques used in SOLAP are in the form of maps, tables and charts(Parmanto et al., 2008). Though, SOLAP is promising decisions support systems, there should be a check on the datacubes which is backend for the spatial decision systems .Misinterpretations of data or incorrect observations could lead to stronger adverse consequences for the organization.

Multidimensional databases indeed consist of dimensions, hierarchy, level, fact and measure(Kimball and Ross, 2002).When it comes to spatial multidimensional databases, its consists of spatial data types, spatial dimension where it members have a cartographic representation (Berderd et al,2006).Spatial data types of point, line, polygon , raster and elevation of unique characteristics and capabilities. Spatial dimensions essentially are of three kinds(Rivest et al., 2003) .The non-geometric spatial dimension which consists of non-geometric data. For example name attributes, type of feature. Second is of geometric dimension which is of the geometric data as shown in figure 2.3. Third one is the hybrid spatial dimension some levels are geometry while the others are non-geometry.

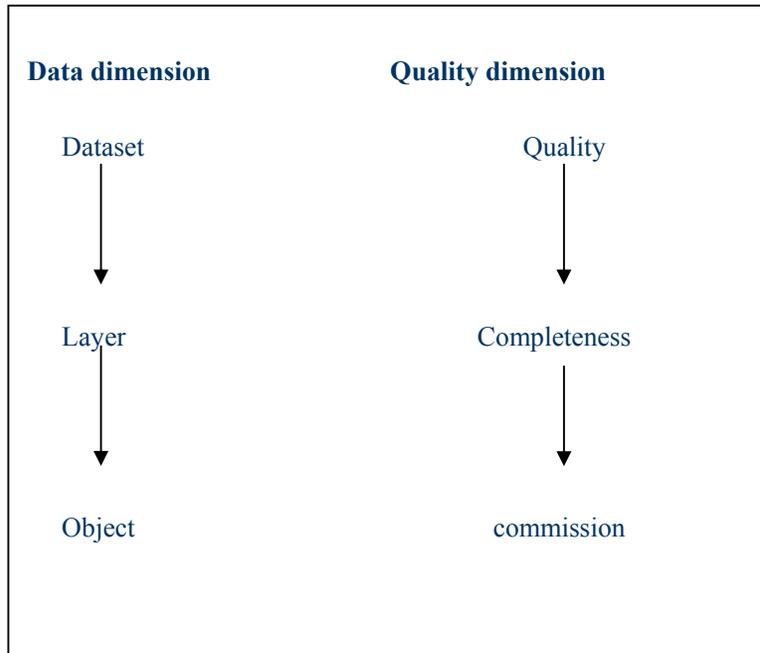


Figure 2.3 “Data” spatial dimension and “quality” non-spatial dimension with different level of granularities(Salehi et al., 2007)

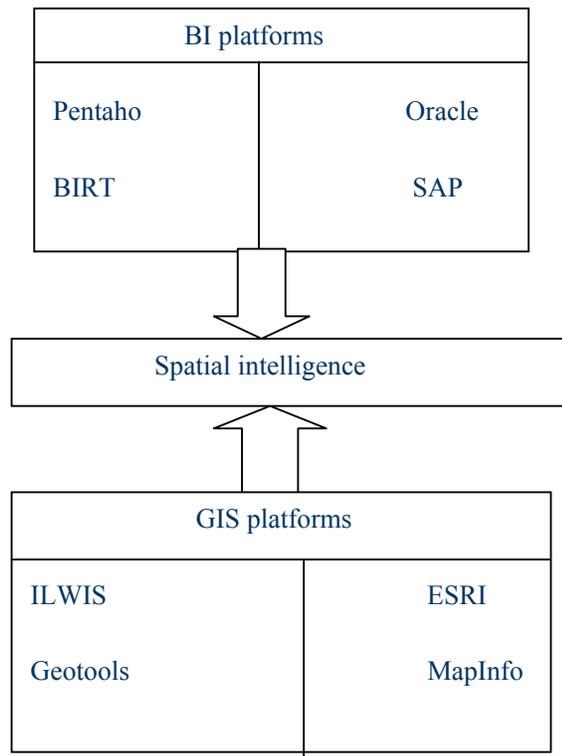


Figure 2.4 GIS and BI (Intege, 2008)

Conclusion

To summarize! User needs there must be predictable, standard and effective spatial data quality model. These can be achieved by conflating two infrastructure, data warehouse concepts like the OLAP, dashboard and data mining with GIS. The figure 2.4 above illustrates on the conflation of the two domains. This conflation gave rise to many new concepts and improved functionalities. For instance interactive exploration, multi-view reporting, spatial temporal queries (Healey and Delve, 2007). Many new concepts like spatial datacubes, spatial dimensions, spatial drill-down/roll-up, spatial aggregations etc (Iris et al., 2006) has been cropped. By this developed technologies we can analysis the spatial data and its data quality at different levels of granularity, of various different themes. As internet has grown in popularity from the past few decades it provides dynamic environment for conveying information. When compared to the conventional GIS, connecting two domains GIS and data warehouse via the web facilitates the end-user has the possibilities to access more views and combination of data for better knowledge discovery and decision making.

3 Research Methodology

This chapter briefs about the methodology followed to achieve the research objective and questions. The overall methodology can be understood by the following flowchart as shown in figure 3.1.

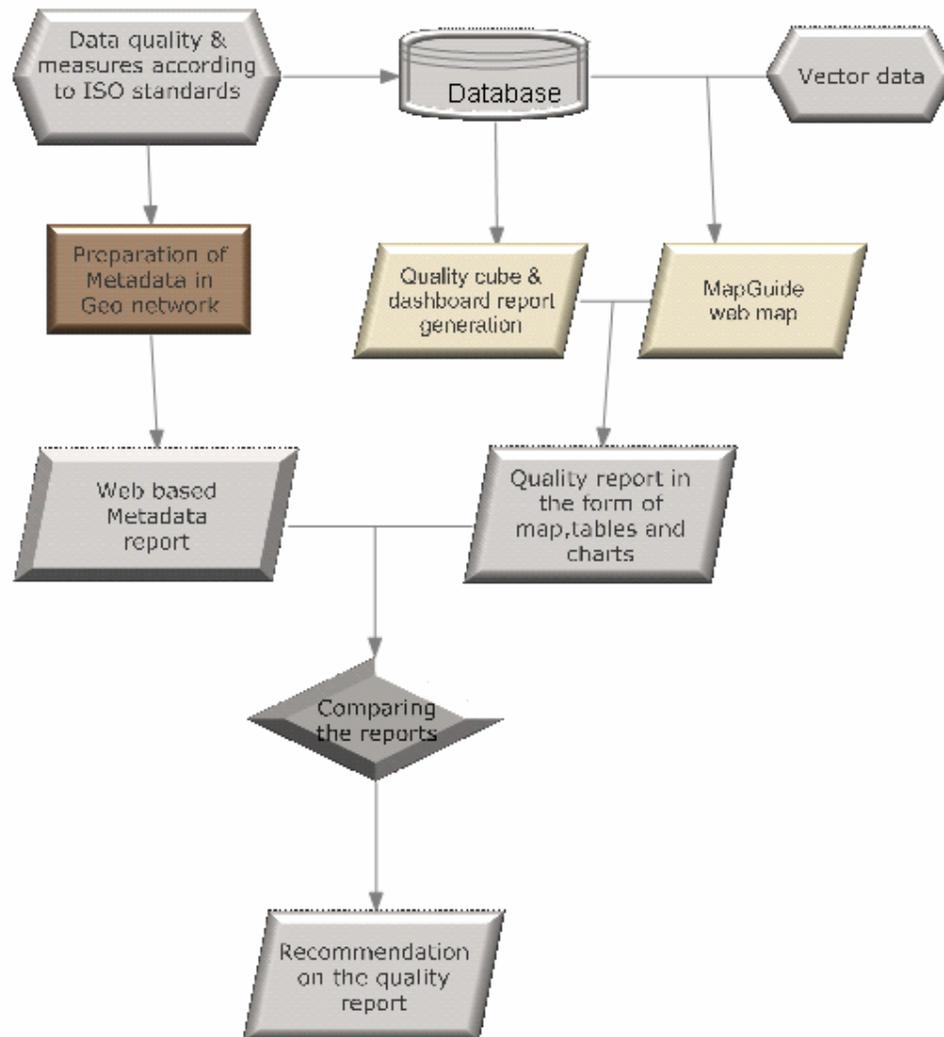


Figure 3.1: The Overall methodological flowchart

The elaborate descriptions of the entire process are illustrated step-by-step:

Step1: Data preparation

As the objective of the thesis, is to store and communicate the quality of a dataset. Prepared the dataset i.e. vector dataset of geographically referenced in the shapefiles format. Which essentially consists of 3 layers - for point layer it is buildings, for line layer it is railway lines and for polygon layer it is the water bodies. This dataset is adequate to demonstrate the full capabilities of the technology.

Step 2: Data quality parameters according to ISO/TC 211

Once the dataset was prepared, the applicable data quality elements, sub elements and data quality measures are identified. The following elements are stored in the database and communicated of the elements mentioned below. Completeness, and Positional accuracy are evaluated and stored in the database automatically once the dataset is loaded, rest of the elements are done manually. The following elements are defined according to the standards of ISO/TC 211.

Table 3.1 Data quality elements and sub-elements

<u>Data quality element</u>	<u>Data quality sub-element</u>
▪ Completeness	Commission Omission
▪ Logical consistency	Format consistency Domain consistency Conceptual consistency Topological consistency
▪ Positional accuracy	Horizontal accuracy
▪ Temporal accuracy	Temporal validity
▪ Thematic accuracy	Classification correctness Non-attribute classification correctness

Step 3: Storing in the databases

After the dataset and the data quality elements are prepared. The schema for the database is designed as shown in figure 3.2. Then the parameters that require for calculating completeness and positional accuracy are analyzed from the map. For instance, positional accuracy the co-ordinates for both the base and reference shapefiles are read from the map. Then the RMSE are calculated by using stored procedure at various dataset levels. Similarly other data quality elements are eventually populated in the database with the values to generate quality cube and dashboard reports.

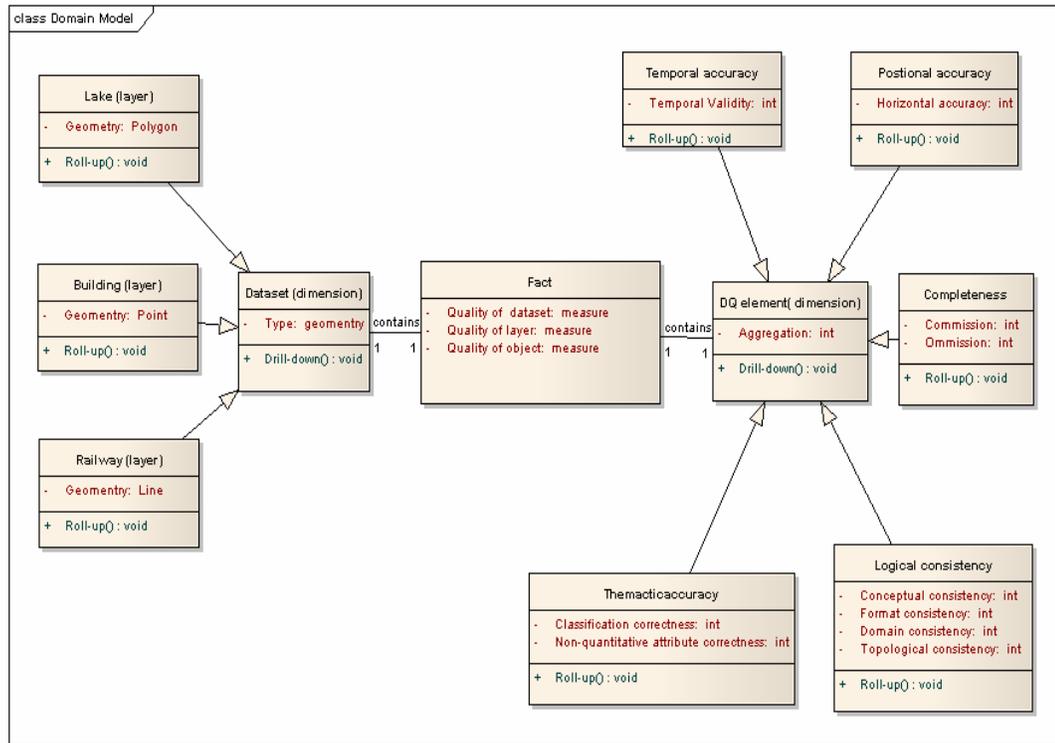


Figure 3.2 Conceptual design of database

Step 4: Creation of quality cube

After the database is ready, cube is being populated with the values from the database dynamically. The quality cube essentially is of two dimensions. One dimension is along the dataset and the other is along the quality dimension. It consists of detailed drill down table report. All the reports at various levels along dataset dimension i.e. overall data, layer and individual objects. Similarly along the quality dimension i.e. detail reports on the data sub-quality elements level are shown. All this is being accomplished by creating stored procedure and views on the database.

Step 5: Generating quality report

After generating the quality cube, the output is in the form of tables, whereas the quality dashboard reports are in the form of column and line charts. The dashboard report page show the quality of entire dataset, along with the data quality elements displayed with the legend. The dashboard reports are provided with a dropdown box, which has an option for drill down to get layer wise detailed quality report. This is done by querying the database to get the detailed quality report.

Step 6: Web interface to database

For the tool to be complete and to show the quality report in the form of map, tables and chart in a single interface, a web interface has been designed. The data quality tool that works in a web environment designed in AutoCAD MapGuide. The MapGuide interface has been customized; buttons like Quality dashboard and Quality cube are put up in the interface. In the layer canvas it contains all the layers. The Map canvas will be displaying the vector data. If the users click on the individual entities the data quality will be displayed in the property box, which dynamically fetches the data from

the database. All the table and charts that are generated from the database are called to the web page in the tool.

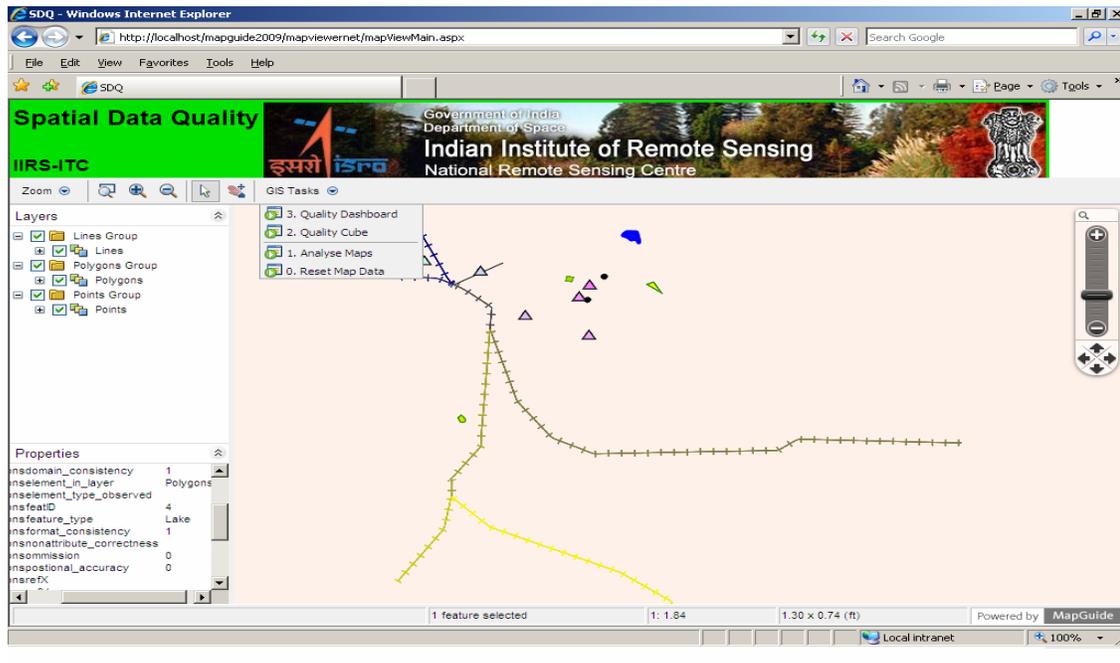


Figure 3.3 Web interface of the data quality tool

Step 7: Preparing metadata report in Geonetwork

Firstly, installation of the Geonetwork tool so that both the tools are in a web environment. Prepared the textual metadata for the same dataset used for the data quality tool. Generated thumbnail for the data, based on the input file in JPEG format to have graphical overview. Apparently, metadata contains more information regarding the dataset other than data quality information. It has identification information, spatial reference system, and point of contact. This was prepared using ISO 19139 template. All the metadata textual information is feed into the catalogue of Geo-network. The data quality tool built is as well linked to Metadata report, so that user will have an option to access both the reports.

Step 8: Evaluating the two quality reports

The reports from the two technologies vary in representation. Though, both the reports are web-based. The report from the data quality tool built is in both spatial and non-spatial graphical format where as the metadata report from geo-network is in the form of text. The Geonetwork Metadata report consists of other information like spatial reference system, identification information, extent, online resources and data quality information. But if the users want to choose a dataset for their application, then he has to be provided with more detailed information on the quality parameters and the data i.e. breaking the information into smaller sub-categories in a visual comprehensive manner. The data quality report has the advantage of interactive exploration and multi view reporting.

Step 9: Recommendation on the method of communicating the quality information

Depends upon the type of users and the user's requirement. Essentially these reports address three users: naïve users, expert users and computer machine. For naïve user the metadata report may not be so comprehensive to judge the quality of the dataset as it just gives the textual metadata. Metadata report uses terminology more of industry specific and technical. Naive users may find difficult in understanding the terms used in the standard whereas the report from the data quality tool is more visual comprehensive as it is in the form of map, table and charts. For expert user metadata report perhaps sufficient to get an idea about the dataset, but if they want to have more detailed quality report for a dataset then the quality report from the data quality tool may be more useful as it gives quality details at various levels. The report is based on visual presentation techniques. The database from the data quality tool can be used by any other computer machine for the flexible and extensive querying the data quality stored, integrate with the other databases, up-to-date information and developers can use to build the application.

Technologies used

As the tool is a web based, the following technologies are used:

- Windows Server 2003, Web Edition used for building and hosting Web applications, Web pages, and XML Web services using the ASP.Net technology. It is designed to be used primarily as an IIS Web server.
- IIS- Internet Information Server provides set of Internet-based services for Windows Sever 2003.
- Microsoft.net framework-It has Base Class Library cover a large range of programming needs that includes user interface, data access, database connectivity, web application development and network communications. .
- Microsoft SQL server is a full featured relational database management system that offers a wide variety of administrative tools like the enterprise manager, query analyzer, SQL profiler etc.
- MapGuide introduced by AutoCAD is a web-based map-making platform that to quickly develop and deploy web mapping applications and geospatial web services. It has features an interactive viewer that includes support for feature selection, property inspection, map tips, AJAX viewer and operations such as buffer, select within, and measure. MapGuide includes an XML database for storing and managing content, and supports most common geospatial file formats, databases, and standards. The MapGuide platform can be deployed on Linux or Microsoft Windows, supports Apache and IIS web servers, and offers extensive PHP, .NET, Java, and JavaScript APIs for application development.

For the data quality tool, MapGuide is deployed in Microsoft Windows 2003 server platform with the Microsoft internet information service (IIS), using .NET programming environment, SQL server database in the backend.

Geonetwork is a web based Geographic metadata catalogue system based on principles of Free and Open Source Software (FOSS), international and open standards. It is a metadata catalogue for describing, searching, editing and publishing geospatial information. The metadata template system allows creating metadata quickly. It supports ISO 191938, FGDC, Dublin core templates.

4 Results and Discussions

This chapter illustrates on the results shown by the data quality tool and Geonetwork followed by the discussions on the report generated by the tools. The study is to propose a data quality tool to explore-synthesis-analysis and report the data quality for the intend users. The study sees the combination of the visualization techniques and web technology possible to deliver the data quality to the users which will aid in judging the quality of a dataset.

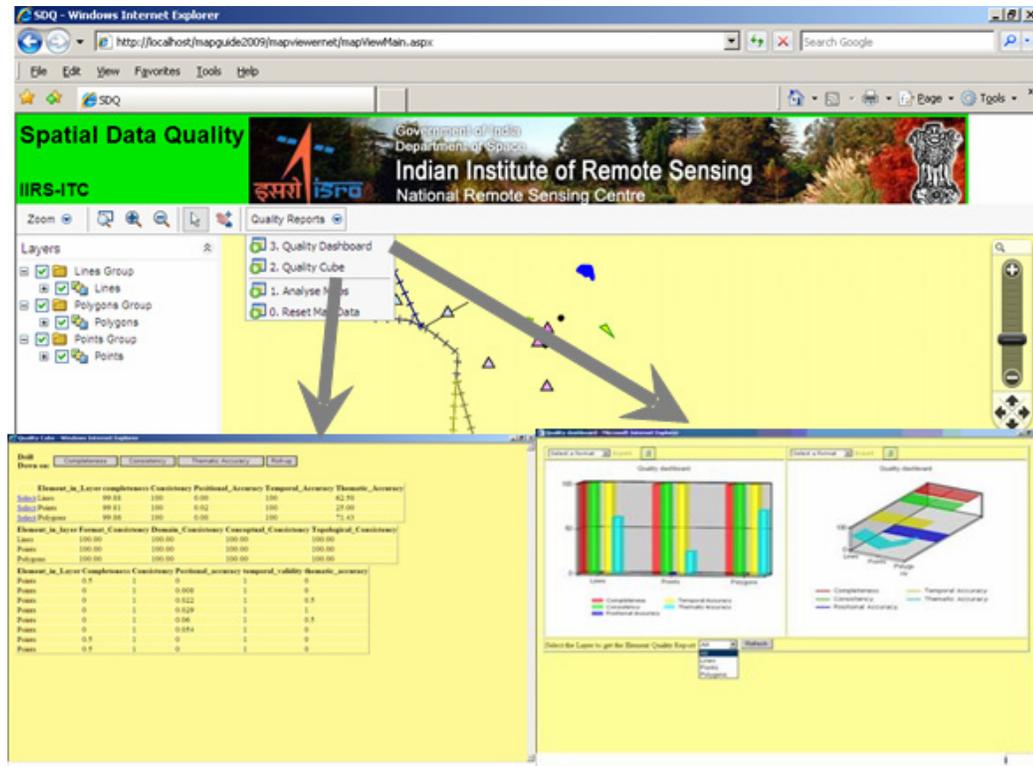


Figure 4.1 Web interface of the tool

Spatial data quality web tool has many interactive components like the dropdown menus, combo box and the property box. The mapping functionalities within the tool include zoom in, zoom out, pan, identify and quality reports. Quality report menu which includes Quality cube and Quality dashboard enable the user to explore the quality. Through the web browser the users are requested to send the information required to the server, the results retrieved are displayed on to the map, charts and table which are handled by the query programming. By using the “identify” function of the tool the user gets the quality information displayed in the property box of the map. Eventually, a graphical web interface tool to visualize the detailed data quality is prepared as shown in the figure 4.1



Figure 4.2 Window displaying a message regarding the database

Reset map data: The functionality of the reset map data is to truncate all the values stored in the database. This is shown by a webpage displaying a message that map data so far collected have been deleted as shown in the figure 4.2. Then the database can be used for the fresh analysis of the vector dataset.

```

Untitled Page - Windows Internet Explorer
asp:page End InitComplete
asp:page Begin PreLoad
asp:page End PreLoad
asp:page Begin Load
<<<<Dump of lstShapes:39
LayerName:Lines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:2|Area:0|Position:
(75.76254,26.93288)|Dimensions:1|Length:0.02563941
LayerName:Lines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:3|Area:0|Position:
(75.78609,26.91871)|Dimensions:1|Length:0.03337745
LayerName:Lines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:4|Area:0|Position:
(75.87988,26.84776)|Dimensions:1|Length:0.2536271
LayerName:Lines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:5|Area:0|Position:
(75.78502,26.85633)|Dimensions:1|Length:0.09833974
LayerName:Lines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:6|Area:0|Position:
(75.76388,26.78575)|Dimensions:1|Length:0.05453981
LayerName:Lines|AttributeType:|ShapeObserved:Line|FeatID:7|Area:0|Position:(75.82207,26.77983)
|Dimensions:1|Length:0.1167649
LayerName:Points|AttributeType:Police staion|ShapeObserved:Point|FeatID:2|Area:0|Position:
(75.83022,26.92301)|Dimensions:0|Length:0
LayerName:Points|AttributeType:Railway station|ShapeObserved:Point|FeatID:3|Area:0|Position:
(75.83456,26.90147)|Dimensions:0|Length:0
LayerName:Points|AttributeType:Police station|ShapeObserved:Point|FeatID:4|Area:0|Position:
(75.80629,26.91255)|Dimensions:0|Length:0
LayerName:Points|AttributeType:Policestation|ShapeObserved:Point|FeatID:5|Area:0|Position:
(75.78641,26.9374)|Dimensions:0|Length:0
LayerName:Points|AttributeType:|ShapeObserved:Point|FeatID:6|Area:0|Position:(75.76145,26.9432)
|Dimensions:0|Length:0
LayerName:Polygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:2|Area:8.777497E-07|Position:
(75.84601,26.98859)|Dimensions:2|Length:0.003502309
LayerName:Polygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:3|Area:1.916509E-06|Position:
(75.85443,26.98488)|Dimensions:2|Length:0.008134381
LayerName:Polygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:4|Area:2.111773E-05|Position:
(75.85362,26.95712)|Dimensions:2|Length:0.0223947
LayerName:Polygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:5|Area:7.433504E-06|Position:
(75.82588,26.93279)|Dimensions:2|Length:0.01085822
LayerName:Polygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:6|Area:8.2357E-06|Position:
(75.77814,26.85402)|Dimensions:2|Length:0.01085721

```

```

Untitled Page - Windows Internet Explorer
Written: 50:LayerName:Polygons|AttributeType:|ShapeObserved:Polygon|FeatID:7|Area:1.027979E-05|Position:(75.86341,26.92864)|Dimensions:2|Length:0.01738169
Written: 51:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:1|Area:0|Position:(75.76263,26.95219)|Dimensions:1|Length:0.0516589
Written: 52:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:2|Area:0|Position:(75.76254,26.93288)|Dimensions:1|Length:0.02563941
Written: 53:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:3|Area:0|Position:(75.78609,26.91871)|Dimensions:1|Length:0.03337745
Written: 54:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:4|Area:0|Position:(75.87988,26.84776)|Dimensions:1|Length:0.2536271
Written: 55:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:5|Area:0|Position:(75.78502,26.85633)|Dimensions:1|Length:0.09833974
Written: 56:LayerName:RefLines|AttributeType:Railway Line|ShapeObserved:Line|FeatID:6|Area:0|Position:(75.76388,26.78575)|Dimensions:1|Length:0.05453981
Written: 57:LayerName:RefLines|AttributeType:RailwayLine|ShapeObserved:Line|FeatID:7|Area:0|Position:(75.82207,26.77983)|Dimensions:1|Length:0.1167649
Written: 58:LayerName:RefLines|AttributeType:|ShapeObserved:Line|FeatID:8|Area:0|Position:(75.78549,26.93562)|Dimensions:1|Length:0.02493593
Written: 59:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:1|Area:0|Position:(75.8414,26.93409)|Dimensions:0|Length:0
Written: 60:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:2|Area:0|Position:(75.83488,26.92964)|Dimensions:0|Length:0
Written: 61:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:3|Area:0|Position:(75.83022,26.92301)|Dimensions:0|Length:0
Written: 62:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:4|Area:0|Position:(75.83395,26.92104)|Dimensions:0|Length:0
Written: 63:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:5|Area:0|Position:(75.83456,26.90147)|Dimensions:0|Length:0
Written: 64:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:6|Area:0|Position:(75.80629,26.91255)|Dimensions:0|Length:0
Written: 65:LayerName:RefPoints|AttributeType:Police Station|ShapeObserved:Point|FeatID:7|Area:0|Position:(75.78641,26.9374)|Dimensions:0|Length:0
Written: 66:LayerName:RefPoints|AttributeType:|ShapeObserved:Point|FeatID:8|Area:0|Position:(75.76145,26.9432)|Dimensions:0|Length:0
Written: 67:LayerName:RefPolygons|AttributeType:Lake|ShapeObserved:Polygon|FeatID:1|Area:1.400309E-06|Position:(75.84834,26.90015)|Dimensions:2|Length:0.004501585
    
```

Figure 4.3 & 4.4 Log files for both the base and the reference data

Analyse map: Issued a stored procedures that populates the database by making use of the above two files shown in the figure 4.3 and 4.4. Essentially the two files are for the base and the reference data. Data quality parameters like the completeness and positional accuracy are populated automatically. For instance the positional accuracy i.e. RMSE is calculated from the position coordinates using the above two files and the files is been generated when click on Analyse map button. While the other quality elements are done manually in the database. With this approach the data quality is stored in SQL server in the background. The functionality of the two buttons are not for the users, it just demonstrate how the database being populated.

Interactive components with the user:

Quality cube

Element_in_Layer	completeness	Consistency	Positional_Accuracy	Temporal_Accuracy	Thematic_Accuracy
Select Lines	99.88	100	0.00	100	62.50
Select Points	99.81	100	0.02	100	25.00
Select Polygons	99.86	100	0.00	100	71.43

Element_in_layer	Format_Consistency	Domain_Consistency	Conceptual_Consistency	Topological_Consistency
Lines	100.00	100.00	100.00	100.00
Points	100.00	100.00	100.00	100.00
Polygons	100.00	100.00	100.00	100.00

Figure 4.5 Quality cube along data and quality dimensions

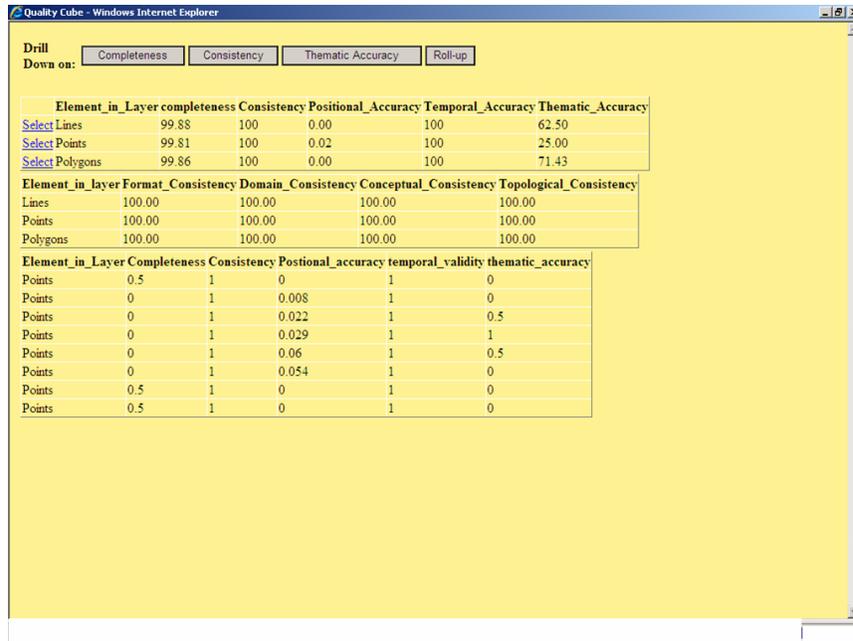


Figure 4.6 Quality cube presenting more details on the dataset and the quality

To generate the quality cube a template has been designed and applied against each term so that the data in rows and the data quality elements are in columns dimensions. The values in the cube are been automatically populated once the database is ready. These are achieved by creating the stored procedures and views on to the database. The drill down and roll-up functionalities could help the naïve users and the expert users in getting the idea about the data quality. Hyperlinks are included in the Quality cube to the web interface at the layer level so the users can begin reviewing detailed quality information. Thus the Quality cube aid the user to understand and explore the quality resources it contains.

Quality dashboard

Quality cube and dashboard approach has been chosen because this enables the users specially the naïve users to built quick view on the data quality. They are assumed to give an advantage for its familiarity. This visualization technique aims to provide better descriptive and explanatory visualization, using the graphical cues in assessing the data quality. The quality dashboard reports as well will be generated dynamically from the database. Dashboards with the legends are intended to support the users in assessing the quality of a dataset. Detailed quality data is recorded in the server database which can be queried through the web interface. The representations used in the Quality dashboards are the column and line charts. Users are facilitated with a drop down menus to inspect the data quality in more detailed way i.e. layer wise reports. Expert user may get much benefit from this facility. It shows whatever values populated in the quality cube in the form of charts.

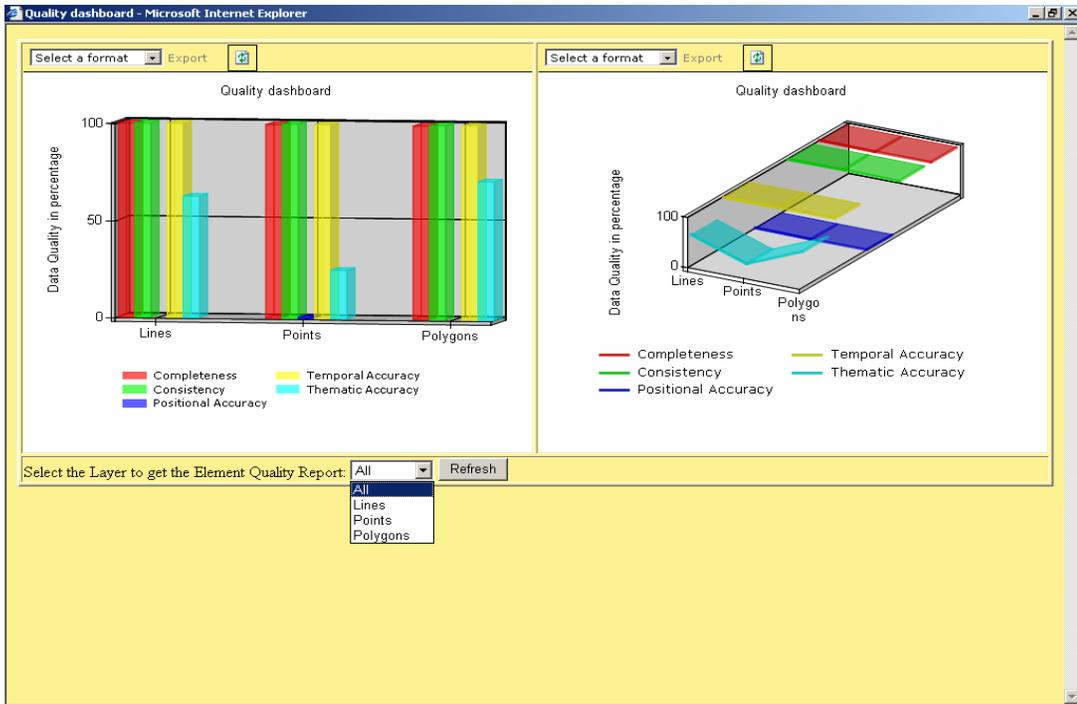


Figure 4.7 Quality dashboard interface displaying the quality of overall dataset and option to get the layer wise reports.

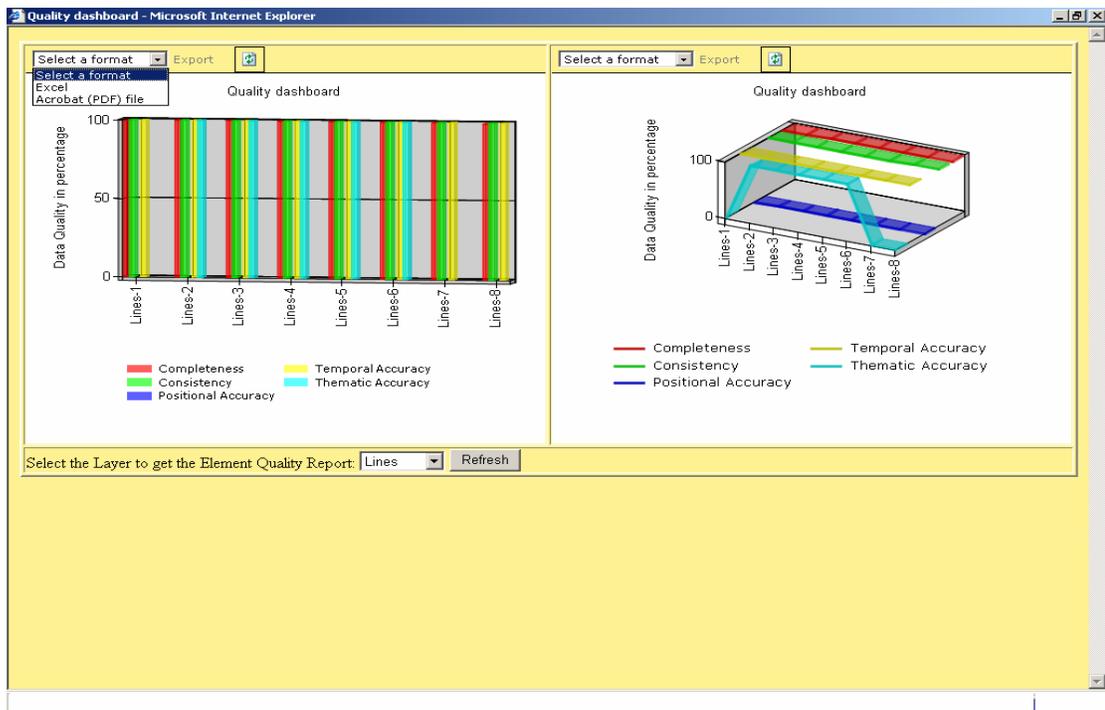


Figure 4.8 Detailed quality dashboard reports for the line layer.

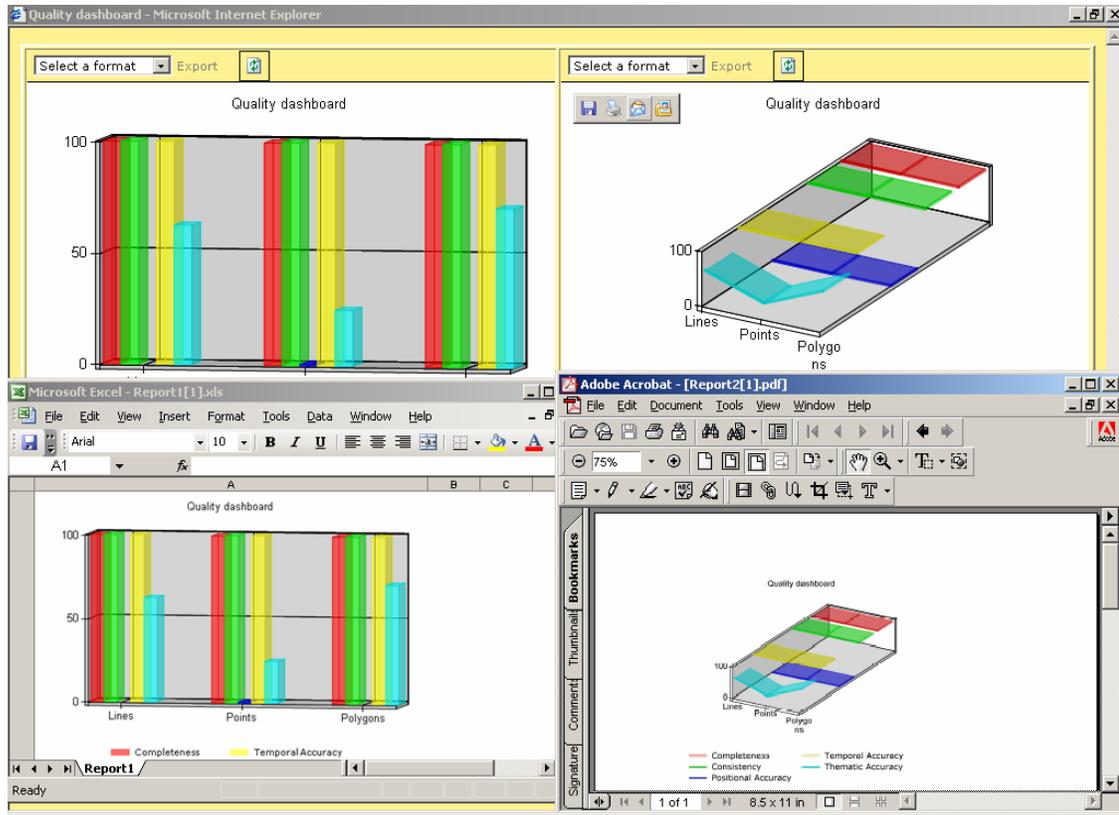


Figure 4.9 Reports exported to excel and PDF format.

The data quality plotted to Y-axis and data plotted to x-axis. The users can intercept the quality of the dataset from the graphs. For instance the users can intercept that the completeness for the line are 100%. These options could be more beneficial to the naïve users. Users are even provided with an option to export the reports to excel and PDF format. The reports can be saved to the local machine, printed, and can even be e-mailed.

Geonetwork

In this study, geonetwork has been chosen as a comparative interface. Geonetwork is an open source metadata catalogue that describes about the geospatial information. Both the data quality tool and Geonetwork are web-based applications. The metadata online template system allows creating and editing metadata quickly, with administrator privileges.



Figure 4.10 Interface of the Geonetwork

Template can be fully customized according to the editor or the administrator choice. The ISO 19139 metadata template that is generated from Geo-network other than the data quality it carries Identification information, Point of contact, Equivalent scale, Geographic bounding box, Reference system and Metadata author information. All the information is in the text format. As the data quality elements stored in the data quality tool are according to the ISO standards, hence ISO 19139 templates are used in the Geonetwork. There are three tabs on left side of the Geo-network i.e. default view, advanced view and the XML view as shown in the figure 4.9. Editing the metadata standard online in three different modes i.e. default, advanced and XML. Default view which displays the full view of the metadata. The data quality tool as well been linked to the Geonetwork shown in the figure 4.11.

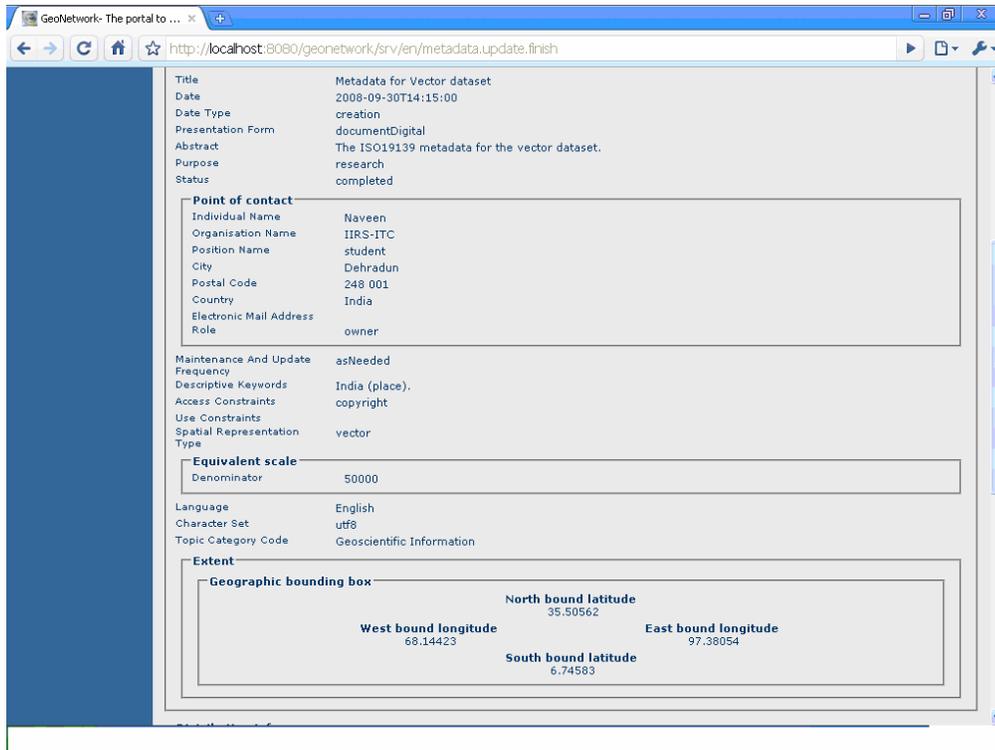


Figure 4.11 Geonetwork interface displaying the ISO 19139 template details.

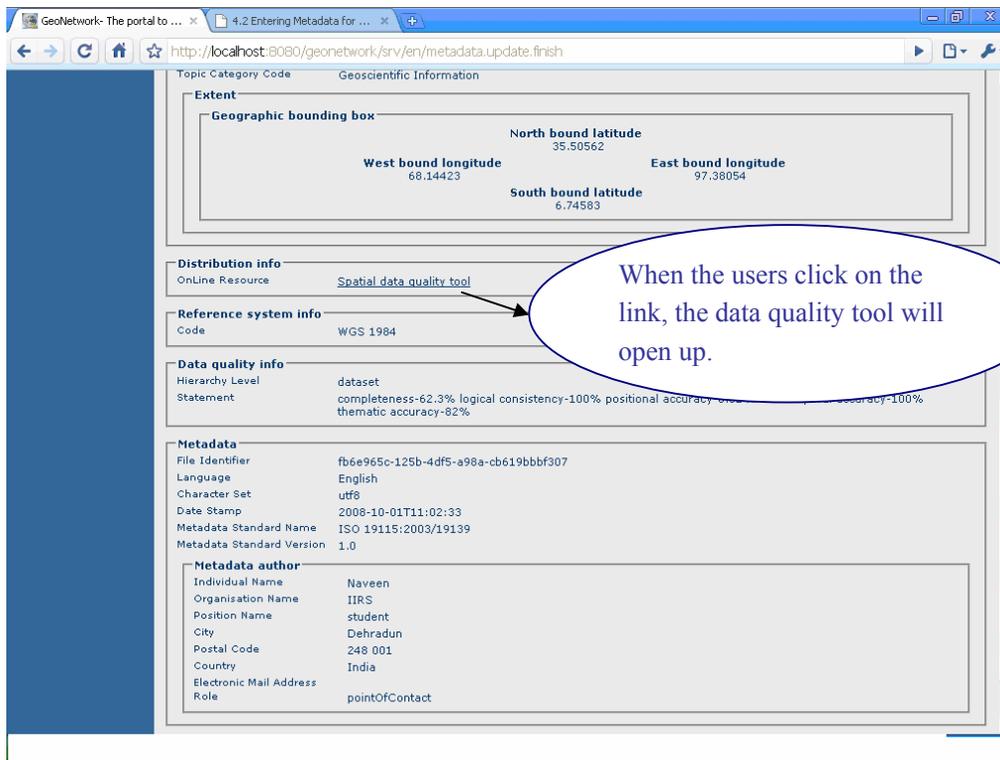


Figure 4.12 Interface which has provides an option to accesses data quality tool

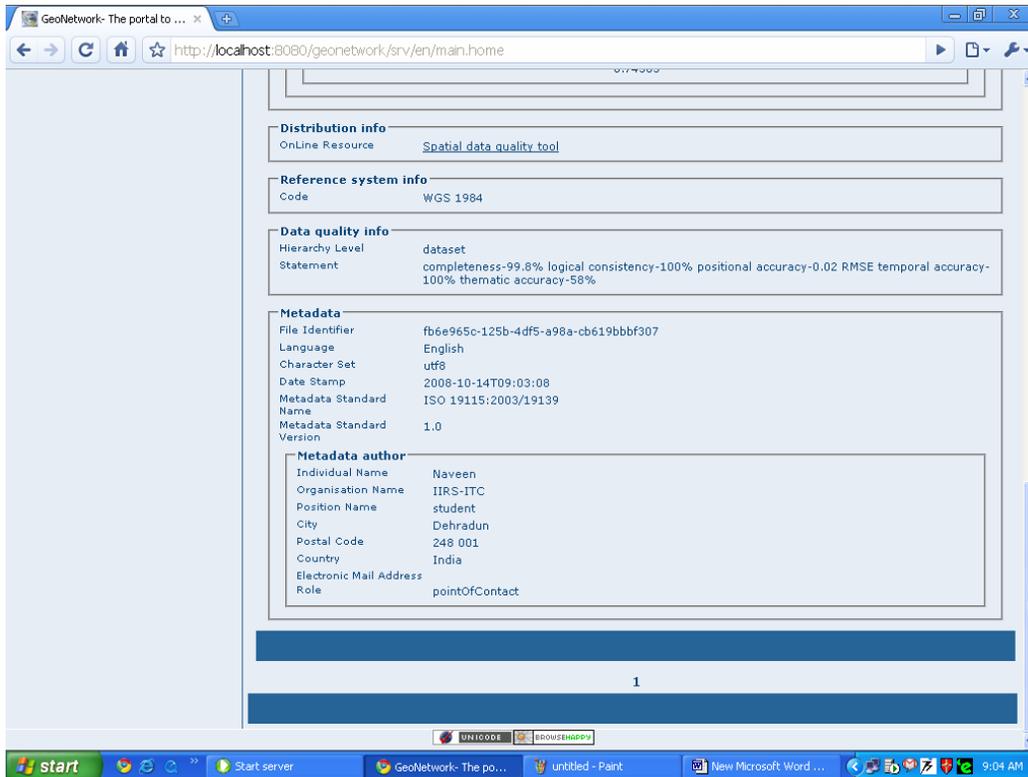


Figure 4.13 Interface displaying the data quality and the rest of the information

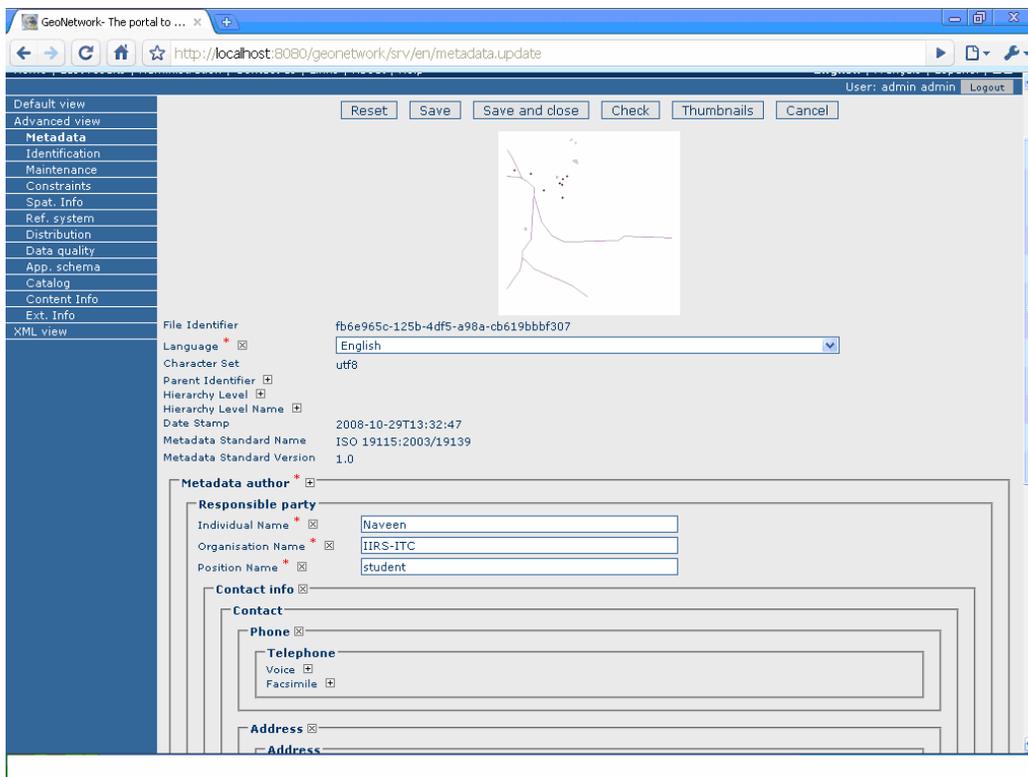


Figure 4.14 Interface displaying thumbnail of the data and details on the selected view

Advanced view displays metadata information according to the users chooses. It gives same report as the default view but it facility to display the entire detailed field through the tabs under the advanced view. The user can explore what ever information he likes's by clicking on the tab.

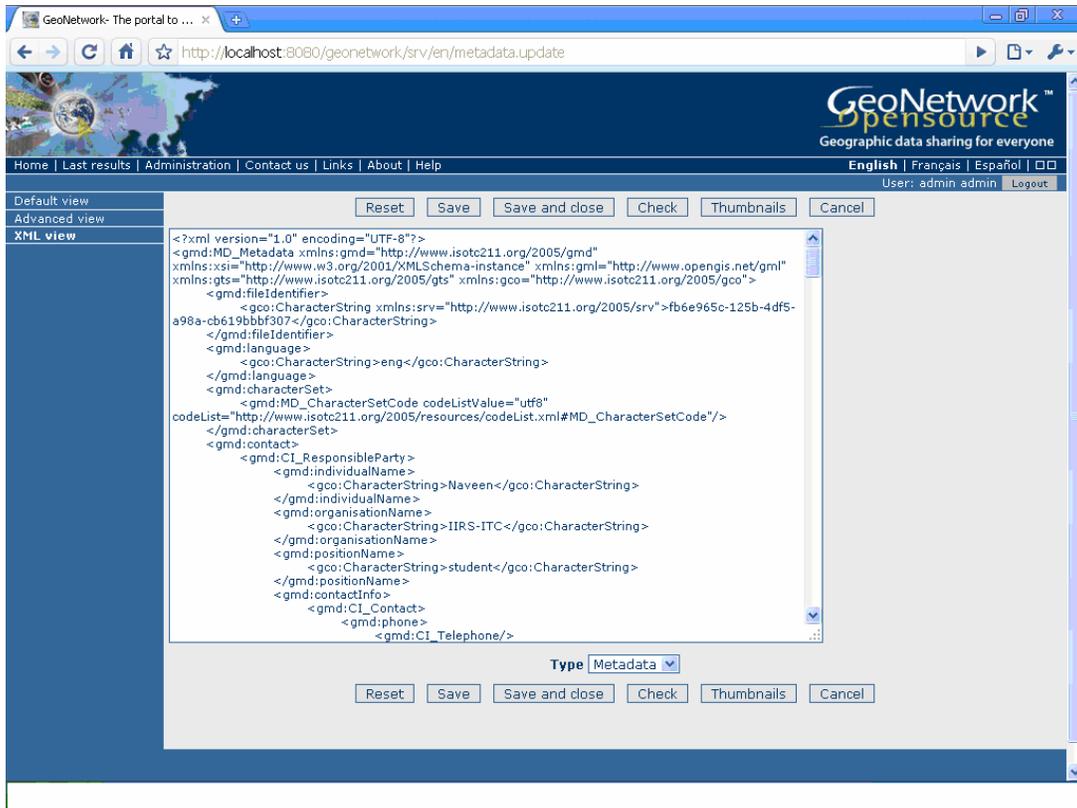


Figure 4.15 The XML schema view

Metadata editing and visualization is based on XML view as well, metadata storage is in a XML style sheet. It shows the entire metadata content in the hierachal structure composed of tags.

Discussions

This study is interested in proposing a solution for communicating the data quality. To see how the developed tool supports naïve users, expert users and the computer machine. In this sense to address the issues of the naïve and the expert users a comparative interface Geonetwork is been chosen.

A comparison between the use of data quality tool developed and the Geonetwork may provide a better insight whether the data quality tool is really usefully to the users (naïve user, expert users and the computer machine). However, the objective of the study is to communicate the data quality information using the concepts of dashboards and SOLAP. Hence the comparison with the Geonetwork and evaluating the data quality elements was not considered as a priority in this study. Nevertheless the data quality tool is presented to provide an illustration on how the data quality is communicated in the form map, tables and charts in a web environment.

In this respect the developed interface can be used to assist the users in exploring the data quality of the dataset. The tool communicates the data quality in the form of maps, tables and charts which will aid the naïve users in understanding the data quality visually. The tool gives the details of the data and the data quality parameter at various hierarchy levels i.e. along dataset and quality elements as well. This can be much useful to the expert users to comprehend on the data quality of a dataset in detail. The database that is used by the tool can be integrated with any other database, build any web applications and can get the up to date quality information, will also facility for better spatial decision making. As the tool being a web based, it is a distributed service to interact with various other computer machines i.e. client- sever architecture which can be established through remote procedure calls (RPC).

The metadata report from geo network carries information like the point of contact, identification information, geographic bounding system and metadata author other than data quality in the textual format. The naïve users may be confused with the technical terms used in the document. For expert users it doesn't provide with the detailed quality information of the dataset. Rest of the information's other than data quality elements may be trivial to the users in choosing the dataset.

Data quality tool supports the narrative, interactive and hierarchal navigation style of visualizing the data quality. This enables the users to explore synthesis and analyze the data quality at various levels of the dataset and the data quality elements. In this regard the data quality may help the users in understanding the data quality than the Geonetwork, a typical metadata catalogue.

5 Conclusion and Recommendations

This chapter briefs on the conclusion of the data quality tool prepared that answers the research objective and questions posed in the study, followed by the limitation of the study and future recommendations.

Conclusion

This thesis started with a purpose to fulfil the objective and answer the questions. As the Objective of the study is to organize and communicate the quality information of a spatial dataset, at different levels of details to the user using the concepts of dashboard and SOLAP tools in a web environment. The data quality tool was built to achieve the objective and to answer the research questions. This thesis essentially addressed three user's naïve users, expert users and the computer machine. The data quality tool built would be able to cater the needs of the three users' in communicating the quality of a dataset. Eventually the users could able get insight about the quality of dataset required for the application.

The following research questions were deduced to fulfil the objective of the thesis:

- How the data quality information has to be stored and organized in accordance with the spatial databases design principles?

In accordance to the objective, the data quality element and sub elements has to shown at different levels of dataset. The database has been designed such that each entity in the database has values that are being populated in the database automatically. While the other quality elements are done manually. Eventually the values in the databases are populated to the quality cube and dashboard reports dynamically. For all this process the stored procedures and views are written on the database, compiling group of SQL queries using GROUPBY operators and aggregation functions to get the desired result at various hierarchy levels.

- How the SOLAP and dashboard concepts, support the communication of data quality to the user?

Essentially, the data quality built based on concepts of SOLAP and dashboards is of human computer interaction in communicating the data quality. The user can explore the data quality of the dataset at various levels. One at the dataset level, two at the layer level and three, individual objects. This analysis can be done even along the data quality level as well. The users can get a quick insight into data quality elements and sub-quality. For example if a user clicks on a particular entity on the map, all the data quality elements get displayed in the property box. Apparently the communication of the data quality is done by visual explanations.

- What makes the use of SOLAP and dashboard tools different from communicating spatial data quality using metadata?

The data quality tool report provides an interactive exploration and multi-view reporting interface. Whereas the metadata report generated from the Geonetwork is essentially textual report. For naïve user quality report from the data quality will be more interactive as it visual comprehensive in the form of maps, tables and charts. Where as the metadata report needs some expertise to understand the document as it contains many technical terms which may put naïve user in state of confusion. Moreover the metadata report is of text. From data quality tool expert user can get insight about the data and the quality elements in various levels dynamically. The metadata report is a static one which only gives the quality of a dataset. The

database that is used by the data quality tool can be used by any other computer machine for the intend applications.

Limitations of the study

This research faced a few limitations. It could not able to accommodate the data overview elements lineage, purpose and usage due to humongous list of quality elements and time reason.

Future recommendations

Further research can be performed to embellish the work done in the thesis. It would be interesting if the tool could be able to calculate the other data quality parameters along with its sub-elements automatically. Overview elements should also be considered. That would complete the data quality tool. Dashboard report can be made more interactive, the drill down and roll-up operations can be preformed on charts itself This can be achieved by a technology FLEX from ADOBE. If we could integrate with FLEX technology, the reports will be more interactive.

6 References

- 1) Boin,A.T., and Hunter,G.J., 2006. Do spatial data consumers really understand data quality information? In: M. Caetano and M. Painho (Editors), 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
- 2) Bedard, Y., Devillers, R., Gervais, M. and Jeansoulin, R., 2004. Towards Multidimensional User Manuals for Geospatial Datasets: Legal Issues and their Considerations into the Design of a Technological Solution, Third International Symposium on Spatial Data Quality (ISSDQ), Austria.
- 3) Bédard, Y., Merrett, T. and Han, J., 2001. Fundamentals of spatial data warehousing for geographic knowledge discovery, Geographic Data Mining and Knowledge Discovery. Taylor & Francis, pp. 53-73.
- 4) Bédard, Y., Rivest, S. and Proulx, M.J., 2005. Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective.
- 5) Chrisman, N., 2006. Development in the Treatment of Spatial Data Quality. Fundamentals of Spatial Data Quality. ISTE, London, 21-29 pp.
- 6) Cockcroft, S., 1997. A Taxonomy of Spatial Data Integrity Constraints. Geoinformatica: 327-343.
- 7) Codd, E.F.,Codd S.B.,Salley,C.T. 1993. Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate Report. E.F. Codd & Associates, Sunnyvale, CA.
- 8) Danko, D., 2000. ISO 19115 Geographic Information - Metadata, ISO/TC211 Geographic information/Geomatics.
- 9) Databeacon, 2004. Why OLAP – An overview of On-line Analytical Processing.
- 10) Devillers, R. and Beard, K., 2006. Communciation and Use of Spatial data Quality Information in GIS. Fundamentals of Spatial Data Quality. ISTE, London, 237-250 pp.
- 11) Devillers, R. et al., 2007a. How to Improve Geospatial Data Usability: From Metadata to Quality-Aware GIS Community, Spatial Data Usability Workshop. Aalborg, Denmark.
- 12) Devillers, R., Bedard, Y. and Jeansoulin, R., 2005. Multidimensional management of geospatial data quality information for its dynamic use within GIS. Photogrammetric Engineering and Remote Sensing, 71(2): 205-215.
- 13) Devillers, R., Bedard, Y., Jeansoulin, R. and Moulin, B., 2007b. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. International Journal of Geographical Information Science, 21(3): 261 - 282.
- 14) Devillers, R., Gervais, M., Bédard, Y. and Jeansoulin, R., 2002. Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual, OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management.
- 15) Dube, E., Badard, T. and Bedard, Y., 2007. Building Geospatial Business Intelligence Solutions with Free and Open Source Components.
- 16) Fu, L. and Hammer, J., 2001. CUBIST: A New Approach to Speeding Up OLAP Queries in Data Cubes, University of Florida, Gainesville.
- 17) Geonetwork, 2008. FAO Geonetwork, The portal spatial data and inforamtion, <http://www.fao.org/geonetwork/srv/en/main.home>.
- 18) Giversen, J., 2007. Recommended use of ISO 19100 quality standards at the NMCAs.
- 19) Goodchild, M.F., Wenzhong, S. and Fisher, P.F. (Editors), 2002. Spatial data quality. Taylor & Francis, London etc., 216-229 pp.

- 20) GSDI, 2008. Spatial data Infrastructure Links <http://www.gsdi.org/SDILinks.asp>.
- 21) Healey, R.G. and Delve, J., 2007. Integrating GIS and Data warehousing in a Web Environment: A Case Study of the US 1880 Census. *International Journal of Geographical Information Science*, 21: 603-621.
- 22) Hunter, G.J. et al., 2005. Next-Generation Research Issues in Spatial Data Quality, *Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis*, Melbourne.
- 23) Integeogeo, 2008. Integeogeo official homepage, <http://www.integeogeo.com/>.
- 24) Iris, J., Napoli, A. and Guarnieri, F., 2006. Contribution of spatial multidimensional technology for natural hazard data analysis.
- 25) ISO/TC211, 2006. ISO 19138:2006 Geographic information - Data quality measures. International organization for standardization(ISO).
- 26) Jakobsson (Editor), 2002. *Data Quality and Quality Management –Examples of Quality Evaluation Procedures and Quality Management in European National Mapping Agencies*. Spatial Data Quality. Taylor & Francis, London, 216-229 pp.
- 27) Jakobsson, A. and Giversen, J., 2007. Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies.
- 28) Kimball, R. and Ross, M., 2002. *The data warehouse toolkit : the complete guide to dimensional modeling*. John Wiley and Sons, Inc., 446 pp.
- 29) Maguire, D., 2004. *Spatial Databases: Unique and Complex*, Geospatial Today.
- 30) Makela, J.M., 2007. The Impact of Spatial Data Quality on Company's Decision Making. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34.
- 31) Malinowski, E. and Zimányi, E., 2007. Logical Representation of a Conceptual Model for Spatial Data warehouses. *Geoinformatica*: 432-455.
- 32) MapWindowGIS, 2008. Map Window GIS official homepage, <http://www.mapwindow.org/>.
- 33) Matias, R. and Moura- Pires, M., 2005. Spatial On-Line Analytical Processing(SOLAP): A Tool the to Analyze the Emission of Pollutants in Industrial Installations. *IEEE*: 214-217.
- 34) Morrison, J., 1995. Spatial data quality. In *Elements of spatial data quality*. The international cartographic association, 1-12 pp.
- 35) Oort, P.V., 2005. *Spatial data quality: from description to application*. Netherlands Geodetic Commission, Delft, The Netherlands.
- 36) Parida, R., 2005. *Principles and Implemenation of Data warehousing*. Laxmi Publication.
- 37) Parmanto, B., Paramita, M., Sugiantara, W. and Burke, D., 2008. Spatial and multidimensional visualization of Indonesia's village health statistics. *International journal of health geographics*.
- 38) Rivest, S., Bedard, Y., Proulx , M.J. and Nadeau , M., 2003. SOLAP: A New Type of User Interface to Support Spatio-temporal Multidimensional Data Exploration and Analysis., ISPRS, Quebec, Canada.
- 39) Rivest, S. et al., 2005. SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *Isprs Journal of Photogrammetry and Remote Sensing*, 60(1): 17-33.
- 40) Salehi, M., Bedard, Y., Mostafavi, M.A. and Brodeur, J., 2007. Towards Specialized Integrity Constraints for Spatial Datacubes. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 34.

- 41) Servigne, S., Lesage, N. and Libourel, T., 2007. Quality Components, Standards and Metadata. *Fundamentals of Spatial Data Quality*, 179-208 pp.
- 42) Shekhar, S., Lu, C.T., Tan, X., Chawla, S. and Vatsavai, R.R., 2000. *Map Cube: A Visualization Tool for Spatial Data warehouses*. University of Minnesota.
- 43) Sonnen, D., 2007. Emerging Issue: Spatial Data Quality, *Directions Magazine*.
- 44) Spatial OLAP, 2008. Spatial OLAP Concepts, <http://spatialolap.scg.ulaval.ca/concepts.asp>.
- 45) Srivastava, R.N., 2008. Spatial Data Quality: An Introduction by Ravi Nishesh Srivastava, <http://gislounge.com/spatial-data-quality-an-introduction-by-ravi-nishesh-srivastava/>.
- 46) Surveyofkenya, 2007. Manual for the Spatial Data Product Specification Creation. In: SurveyofKenya (Editor).
- 47) Tóth, K. and Nunes de Lima, V., 2005. Data Quality and Scale in Context of European Spatial Data Harmonisation, 11th EC-GIS workshop, Alghero.
- 48) Van Oosterom, P.J.M. and Lemmen, C.H.J., 2001. Spatial data management on a very large cadastral database. *Computers, Environment and Urban Systems*, 25(4-5): 509-528.
- 49) Yang, T., 2007. *Visualisation of Spatial Data Quality for Distributed GIS*, The University of New South Wales, Sydney, 199 pp.
- 50) Zhang, Y. et al., 2005. Spatial Data Cube: Provides Better Support for Spatial Data Mining. *IEEE*: 795-798.