

Chapter 10

Improving Spatial Data Quality through Spatial ETL Processes

Elzbieta Malinowski

University of Costa Rica, Costa Rica

Sehyris Campos

University of Costa Rica, Costa Rica

ABSTRACT

The growing availability of spatial data related to different aspects makes managers at different levels of administration aware of the possibilities to enhance decision making processes through map visualization. Currently, the so-called location intelligence is identified as an important trend for the next-generation business intelligence solutions. However, before considering spatial data as “first-class citizens,” users that are not experts in geo-fields (e.g., cartography, surveying) should learn about possible problems that may arise while using and producing spatial data. These problems must be solved to improve spatial data quality and to increase the benefits that this data can deliver. Unfortunately, there is still a poor connection in applying scientific solutions, international standards, or technological advances for improving spatial data quality in everyday usage of this data. In this chapter, the authors refer to different problems that may exist in handling spatial data and show several examples of how these problems can be detected and solved using spatial ETL tools. Problem detection is based on a set of control parameters derived from the international standard.

INTRODUCTION

Current advances in technology and information systems and new initiatives, e.g., creating Spatial Data Infrastructures (SDIs), open the possibility to publish and share spatial (or geographic) and

conventional data among interested parties. A SDI is defined as “the technology, policies, standards, human resources, and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data” (OMB, 2010). These spatial data repositories distributed over the Internet at national or global levels mainly represent public sector agencies that could play both roles: data

DOI: 10.4018/978-1-4666-4892-0.ch010

providers and consumers (Granel, Gould, Manso, & Bernabé, 2009; Elwood, 2008). SDIs' purpose is to serve as a foundation for an Information Society willing to use and share spatial data over the Internet that "allow nations to better address social, economic, and environmental issues" (GSDDI, 2012).

SDIs and other kinds of spatial data repositories over the Internet, e.g., Google Maps, Microsoft Bing Maps, make data available to different kinds of users. This data may form part of different Business Intelligence (BI) solutions (Bédard, 2005; Intelli3, 2010; Badard & Dubé, 2009; Bimonte, Wehrle, Tchounikine, & Miquel, 2006; Di Martino, Bimonte, Bertolotto, & Ferrucci, 2009) and the so-called location intelligence is mentioned as an important trend for the next-generation BI (Mathew, 2012). Spatially-enhanced BI solutions can help improve the decisions made at the personal, local, national, or global levels, due to the fact that, as it is widely acknowledged, the presentation of data within a space helps discover patterns that otherwise would be difficult to find. For example, on a personal level, a user could determine if the land that s/he wants to buy is not close to a garbage disposal or high tension towers, or whether the area is over-populated; on a local level, municipalities could rely on this data to prepare plans for cantonal human development in order to improve living conditions and strengthen territorial planning, or business managers could evaluate the distribution of customers that generate more profit; on a national level, authorities could make better decisions to create risk prevention plans for hazardous areas (e.g., flooding, erosion, earthquake) taking into consideration population, house distribution, or road availability, or business executives could analyze performance data related to a specific business field.

Nevertheless, several issues must be addressed before considering spatial data as "first-class citizens" in the decision-making processes at different levels of management. Traditionally, spatial data

was stored in a central repository in a specific format provided by software tools developed for Geographic Information Systems (GISs). This software was perceived as a standalone technology until mid-1990s and was used mainly by experts in the field, i.e., geo-specialists, e.g., geographers, cartographers, surveyors (Yeung & Hall, 2007). This limitation was imposed due to the complexity of tools that manage spatial data and the necessary specialized knowledge (geo-knowledge) required to understand the concepts related to spatial data format, as well as functions and operations. These geo-specialists were mainly in charge of digitalizing spatial data and they understood very well the particularities of spatial data and problems that may affect its quality. Furthermore, they were aware that GISs will not help them improve spatial data quality since these systems work under the assumption that data is perfect and do not provide the capabilities to establish data quality control (Delavar & Devillers, 2010). Although, spatial DBMSs (called geo-databases by geo-specialists) currently provide some controls, e.g., declarative and dynamic (triggers), that may help improve spatial data quality, the transition from the traditional storage of spatial data to a new alternative requires geo-specialists to acquire new knowledge that is not always an easy and time-efficient endeavor.

On the other hand, the technology has changed and other kinds of users with a profile that is usually unrelated to GISs are in charge of either collecting data from existing repositories or creating it from aerial photographs and satellite images (Mäkelä, 2006). As a consequence, many users and producers of spatial data are not trained in geo-fields and may not be aware of the existence and number of problems that this kind of data may have (Delavar & Devillers, 2010; Boin & Hunter, 2005; Wang & Strong, 1996). Therefore, there is a need to increase this awareness among non-expert users and look for solutions to improve spatial data quality. However, although research on spatial data quality has a 30-year history,

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/improving-spatial-data-quality-through-spatial-etl-processes/96151

Related Content

Evaluation of Nosocomial Infection Risk Using a Hybrid Approach

José Neves, Eva Silva, João Neves and Henrique Vicente (2016). *Applying Business Intelligence to Clinical and Healthcare Organizations* (pp. 24-42).

www.irma-international.org/chapter/evaluation-of-nosocomial-infection-risk-using-a-hybrid-approach/146061

Business Intelligence in Audit

Leticia R. Webb (2012). *International Journal of Business Intelligence Research* (pp. 42-53).

www.irma-international.org/article/business-intelligence-audit/69968

Optimizing the Accuracy of Entity-Based Data Integration of Multiple Data Sources Using Genetic Programming Methods

Yinle Zhou, Ali Kooshesh and John Talburt (2012). *International Journal of Business Intelligence Research* (pp. 72-82).

www.irma-international.org/article/optimizing-accuracy-entity-based-data/62023

Data Warehouses and Business Intelligence in Croatia: Do Managers Know How to Use Them?

Kornelije Rabuzin and Darko Škvorc (2016). *International Journal of Business Analytics* (pp. 50-60).

www.irma-international.org/article/data-warehouses-and-business-intelligence-in-croatia/149155

Developing an Explainable Machine Learning-Based Thyroid Disease Prediction Model

Siddhartha Kumar Arjaria, Abhishek Singh Rathore and Gyanendra Chaubey (2022). *International Journal of Business Analytics* (pp. 1-18).

www.irma-international.org/article/developing-explainable-machine-learning-based/292058