GQBox: Geospatial Data Quality Assessment

Yassine Lassoued
CMRC, University College Cork, Ireland
y.lassoued@ucc.ie

Mohamed Reda Bouadjenek*
Alcatel-Lucent Bell Labs France
Centre de Villarceaux, Nozay
reda.bouadjenek@alcatel-lucent.com

Omar Boucelma
LSIS - Aix-Marseille Université
omar.boucelma@lsis.org

Fernando Lemos
PRiSM, University of Versailles
flem@prism.uvsq.fr

Mokrane Bouzeghoub
PRiSM, University of Versailles
mok@prism.uvsq.fr

ABSTRACT

In order to measure and assess the quality of GIS, there exist a sparse offer of tools, providing specific functions with their own interest but are not sufficient to deal with broader user's requirements. Interoperability of these tools remains a technical challenge because of the heterogeneity of their models and access patterns. On the other side, quality analysts require more and more integration facilities that allow them to consolidate and aggregate multiple quality measures acquired from different observations or data sources, in using/combining seamlessly different quality tools. Clearly, there is a gap between users's requirements and the spatial data quality market. This demo paper will illustrate GQBox, a geographic quality (tool)box. GQBox supplies a standards-based generic meta model that supports the definition of quality goals and metrics, and it provides a service-based infrastructure that allows interoperability among several quality tools.

Categories and Subject Descriptors

H.2.8 [**:**]: [Database Management]: Database applications - Spatial databases and GIS; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms, Design, Performance

Keywords

Geospatial Data Quality, Web services

*This work has been mainly done while being a Master student at University of Versailles.

ACM SIGSPATIAL GIS 2010, November 2-5, San Jose, CA, USA

1. INTRODUCTION

The role and importance of spatial data quality has been recognized by different institutions and organizations. Data quality problems are even worse in the internet era due, notably, to the availability of various data sources, accessible by means of standard interfaces such as OGC's WFS and WMS. Quality issues are of different kind: for instance, in the context of GIS application development with multiple sources, users face a data integration problem [4]. A more recent context has been brought by volunteered geographic information (VGI) projects such as OpenStreetMap where openness has exacerbated data quality issues because data modification (by anyone) is made easy.

To tackle data quality issues, several research, development and standardization efforts have been conducted during the last decades [1] and led to several quality metadata standards, methodologies and dimensions/metrics. This demo paper will illustrate GQBox [2], a geographic quality toolbox which supplies a standards-based generic meta model that supports the definition of quality goals and metrics. GQBox is implemented as a service oriented platform where several quality tools are considered as web services.

2. SEAL TRACKING SCENARIO

The scenario is provided by the Marine Ecology Group of the Coastal and Marine Resources Centre (CMRC), University College Cork, Ireland. Several marine biologists at CMRC and world-wide, study the behavior of marine animals, such as seals, turtles and dolphins, by tracking them and analyzing their trajectories in light of various environmental parameters recorded as part of the tracking process. In our example, seals are tagged using specifically designed devices that would record their position at predefined time intervals together with the water depth and temperature. Seal track data are sent via satellite for processing.

Figure 1 shows the seal track database schema. Table *Seals* contains information about tagged seals such as their identifiers, names and descriptions. All seals' positions are stored in the *SealPositions* feature class and linked to seals using *SealID* as a foreign key. To each position corresponds a series of 12 depth (pressure number) and temperature measurements. These measurements are stored in table *Measurements* and linked to the seal positions using the *PositionID* foreign key. Coastlines are represented by the *Coastlines* feature area class.

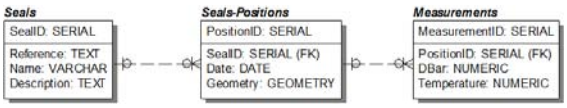

Figure 1: Seals Database Schema

Despite the simplicity of the above schema, data quality problems may arise and can be described as follows:

1. **Topological Consistency -** There are cases where recorded animal positions are inland, which leads to topological inconsistencies between these positions and the coastlines.

2. **Domain Consistency -** There are cases where temperature values are outside the typical sea temperature range of the area (5°C to 16°C), e.g. 35°C. This leads to domain inconsistencies of the temperature measurement.

3. **Completeness -** Quite so often, less than 12 depth-temperature measurements are recorded leading to data incompleteness.

3. WHAT WILL BE DEMONSTRATED

We illustrate the following functionalities:

- **Goal and Questions Creation.** Figure 2 depicts the GQBox interface for creating a quality goal and is related questions. First of all, the user (a business/quality analyst) has to supply a name and a description of a goal. Second, she has to refine this goal in defining a number of questions pertaining to this goal. Third, she has to link each of these questions to a set of corresponding personalized quality factors and an IS object (object stored in the Information System). Finally, she has to associate for each of these quality factors a quality service that measures it (the Geo quality services are defined in a registry) and execute the global quality goal.

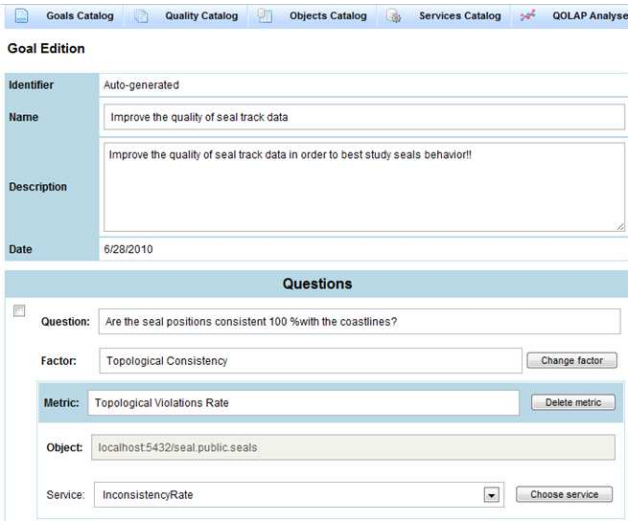

Figure 2: Screenshot of goal creation in the GQBox services

- **Execution and Analysis of Quality Goal and Results.** The quality analyst has to define the goals periodicity execution. Indeed, quality services may be periodically executed or invoked on demand to collect quality information at a certain time.

  The execution of a quality goal leads to reporting results which illustrate values returned by appropriate (web) services that correspond to various metrics.

  Figure 3 depicts the results related to the goal and questions presented in section 2. It shows that the topological violation rate at 01/07/2010 was 12,2% while the range violation for the pressure measures was 69,4%. All the temperatures were in the range validation and the values for seals' references were equal to 0%. Figure 3 also shows inconsistent objects w.r.t the coastlines: those objects are stored in GQBox database in XML format.

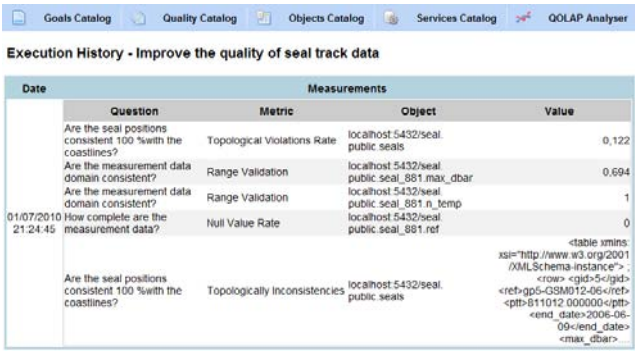

Figure 3: Screenshot of the Goal Execution Results

- **Visualization of Inconsistent Data.** We are using GeoServer which allows us to connect our spatial data to Virtual Globes such as Google Earth and NASA World Wind as well as to web-based maps such as Google Maps and Bing Maps.

Note that experiments performed on different kinds of applications (data warehousing, CRM, medical data) have shown the relevance and the usefulness of the previous versions of the QBox [2, 3, 5], in particular its ability to characterize quality goals with multidimensional factors, to reuse basic measurement process and to aggregate measurement values along defined time intervals.

4. REFERENCES

[1] R. Devillers and R. Jeansoulin. *Fundamentals of Spatial Data Quality*. Wiley, April 2006.

[2] L. Etcheverry, V. Peralta, and M. Bouzeghoub. QBox-Foundation: a Metadata Platform for Quality Measurement. In *EGC-2008, 4ème Atelier Qualité des données et des Connaisances*, 2008.

[3] L. González, V. Peralta, M. Bouzeghoub, and R. Ruggia. Qbox-Services: Towards a Service-Oriented Quality Platform. In *ER Workshops*, pages 232–242, 2009.

[4] Y. Lassoued, M. Essid, and O. Boucelma. A Metadata-Driven Geographic Data Mediator. *Ingénierie des Systèmes d'Information*, 12:137–161, February 2007.

[5] F. Lemos, M. R. Bouadjenek, Z. Kedad, and M. Bouzeghoub. Using the qbox platform to assess quality in data integration systems. *Ingénierie des Systèmes d'Information*, 2010.