OSGeo Journal

The Journal of the Open Source Geospatial Foundation Volume 2 / September 2007

In This Volume

Sponsor Perspectives

1Spatial: Spatial Data Quality and the Open Source Community

by Mike Sanderson, Graham Stickler, Steven Ramage, 1Spatial

With a specific interest in FDO and the direction of other open source technologies, 1Spatial (formerly Laser-Scan) became an official OSGeo Associate Sponsor in 2006. 1Spatial's area of expertise is data management, quality management and their Radius Studio product. The authors were invited to present some of their thoughts on these topics, including how they see these ideas being delivered using open source tools. –Editor

Introduction

The Open Source Geospatial Foundation (OSGeo) introduces itself on its website, with the words "The Open Source Geospatial Foundation has been created to support and build the highest-quality open source geospatial software." It is clear the objective is to create a peer review community to improve software quality (1). This article addresses the quality element, but from an entirely different perspective.

Spatial data are subject to different regimes of quality management, irrespective of whether they are used in an open source environment or not. There will always be issues regarding data consistency and integrity or fitness for purpose because data are constantly changing. This may be through real world change, the introduction of new technologies for capture, update or organisational change that alters the boundary conditions.

Since the mid '90s the International Standards Organisation (ISO) and the Open Geospatial Consortium (OGC) have worked on standards, creating and overseeing the Web Feature Service (WFS), Web Map Service (WMS) and Geography Markup Language (GML). Then more recently through the work of the OSGeo mechanisms now exist to access geospatial data regardless of source through Feature Data Objects (FDO).

As a result a whole new set of issues are created around spatial data quality and fitness for purpose. We want to now extend the paradigm to include these spatial data quality and reuse issues as data collected for one purpose are being accessed for increasingly diverse use, more often by a completely separate entity. As an industry, Google has raised the industry profile to the point where it will hurt us all if data quality is suspect.

Our vision is to provide a set of web-based tools to enable spatial data quality to be assessed, i.e. conformance checking. Ideally this will offer improved spatial data management in an open source environ-

ment. It will support the goal of improving operational efficiency by allowing a framework to be developed to aggregate spatial data.

We wish to offer the OSGeo community the possibility of contributing to these initiatives via our Practitioner Program and through the Open Geospatial Consortium (OGC) Data Quality Working Group.

Background Information

By way of supporting information it is necessary to understand the work that has been carried out by the International Standards Organisation (ISO). ISO provides guidelines for putting together spatial data quality management frameworks. ISO 19113 and ISO 19114 describe various mechanisms for determining and measuring data quality. These ISO standards embody principles and evaluation procedures for geographic information (for a useful summary of these principles and procedures, see Chapter 15 in Spatial Data Quality (2). The opening sentence of ISO 19113 sets the tone:

> Geographic datasets are increasingly being shared, interchanged and used for purposes other than their producers' intended ones.

The forerunner to the current ISO standards were applied by them using the Digital Chart of the World (DCW) project as a case study. The DCW[1] was produced in 1992 and there was an effort to work on "fit for purpose" issues. But in 1995, the initiative ran out of steam, because only non-quantitative (i.e. qualitative) assessments of the quality of geographic datasets could take place. Quantitative assessments were just not possible for large geospatial datasets due to the lack of processing power available at the time. It is quantitative assessment that is really valuable for assessing logical consistency and positional accuracy. And as Jakobsson said so eloquently: "Combining data sets that have no quality information can be very difficult or impossible." (2)

Quantitative Spatial Data Quality

A data quality audit is designed and implemented in order to determine the answer to the question, how bad (or good) is your spatial data? There are three high-level objectives of a data quality effort:

- Produce statistically valid data quality measurements of source or master data
- Investigate, identify, and document leading data quality causes (and exceptions)
- Create an assessment and recommendation report

However, in order to undertake quantitative assessments, there must be some guidelines to follow and a formal approach. The flowchart in Figure 1 describes such an approach for reviewing spatial data quality. This is a tried and tested approach used by 1Spatial for projects with customers and partners - it offers a route for undertaking quantitative assessments.

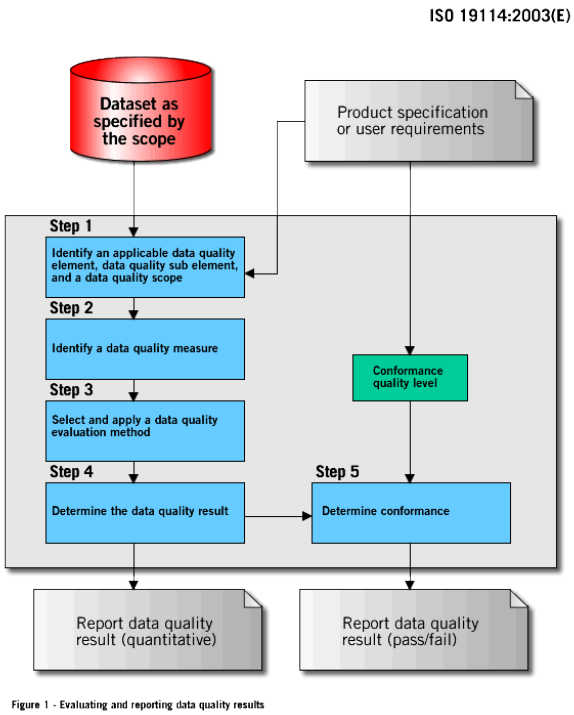

Figure 1:

What is important in this approach is contextual analysis of the data quality records or spatial data holdings. This involves working through steps 2 to 5 using a rules-based approach to determining the conformance of spatial data against the specifications or business rules. In an Open Source environment this would be accessed using an XML interface to Feature Data Objects (FDO). There are two key elements to this approach. Firstly, it provides a quantitative assessment of spatial data quality, i.e. data conformance % against business rules. Secondly, it is an

[1] Digital Chart of the World: http://www.nlh.no/ikf/gis/dcw/

independent verification, i.e. not using the GIS tools that create or edit the data – these remain unchanged.

Data quality problems are often widespread and originate in the source data itself. They can be geometric, topological or attribute-based. In order to combat these problems, data custodians, analysts and stakeholders must understand their source data. This understanding can come from data profiling or what can be thought of as self-describing metrics. The analysts and data managers must understand how the data profile fits the business requirements. This is where a business rules-based audit is not only useful, but critical.

Data Quality Rules!

In order to establish the fitness for purpose for spatial data, it is clear that it is first necessary to understand the business rules relating to those data and how they should be interpreted. Conducting a business rules-based audit can be critical to combating data quality problems.

Figure 2 highlights an example of identified measures for assessing spatial data quality. It refers to five key elements that change according to the type of spatial data or its application. The objective at this stage is to define the business rules that the data should obey. It is often difficult to obtain this information from the user. The original data model specification is often unavailable or hasn't been updated since inception. Rather than start from a blank piece of paper a potential source of rules is the data itself. Rules may be discovered using an analysis of dominant statistical patterns in the data.

The current challenge is that there are limited toolsets available in the open source community and the wider geospatial market place as a whole to carry out such tasks and quantitatively measure spatial data quality. While we may look to open architectures and standards as defined by Service-Oriented Architectures (SOA), the World Wide Web Consortium (W3C) semantic Web framework and the Web Ontology Language (OWL), it is clear that the semantic Web rules language doesn't support the geospatial needs. Work done recently by the OGC as part of their OWS-4 test bed on a Topology Quantitative Assessment Service (TQAS) supports this view.

ISO 19114:2003(E)

Data quality element	Data quality subelement	Relevant?
completeness	commission	yes
	omission	yes
logical consistency	conceptual consistency	no
	domain consistency	yes
	format consistency	no
	topological consistency	yes
positional accuracy	absolute or external accuracy	yes
	relative or internal accuracy	no
	gridded data position accuracy	no
temporal accuracy	accuracy of a time measurement	no
	temporal consistancy	no
	temporal validity	no
thematic accuracy	classification correctness	no
	non-quantitative attribute correctness	no
	quantitative attribute correctness	no

Figure 2 : Summary of relevant quantitative quality information

Figure 2:

1Spatial have been addressing this problem and have developed a web-based tool, Radius Studio, based around this quantitative rules-based data quality paradigm. Using a variation of an artificial intelligence boosting algorithm adapted to spatial data mining, it works by initially considering a small sample of objects taken at random from the data store and then works outwards from these objects, considering nearby objects. An initial set of spatial rules are proposed and subsequently enhanced to include non-spatial elements such as attribute joins, equalities and inequalities, correlated to the spatial relationships between the objects sampled. The final set of rules is converted to a form that allows them to be stored in a rules repository, which is based on a common language interface that incorporates OGC spatial operators.

It is clear that such a tool, when combined with FDO and MapGuide Open Source, can enable an assessment of spatial data against business rules. This conformance checking approach can be invaluable in providing a quantitative assessment of spatial data quality.

As a result 1Spatial has already commenced a Practitioner Program to enable their end customers to validate spatial data across the Web by running spatial data audits. Through remote access to Radius Studio, Practitioners are able to use the rule building and assertion capabilities to carry out data certification audits on behalf of customers. In return 1Spatial receives feedback on the product and 1Spatial tools are recommended for any fix up or ongoing repurposing, reuse or re-engineering requirements.

Open Source Opportunity

There is an opportunity here for the open source community to engage with Radius Studio to develop this capability and make it widely available.

By lodging this assessment service within OSGeo

could enhance its reach, providing the basis for co-operation between parties to allow the harmonization of geographic information to take place, and to build a community (3).The opportunity will then exist for the open source community to extend the range of FDO providers and open up more formats that can be subject to a quantitative assessment of spatial data quality.

Another reason for placing an assessment service within OSGeo is to create a community that will work on defining the rules expression language for creating the quality measures. The Semantic Web Rules Language (SWRL) is currently not mature enough, but as history has shown, we in the industry can create a standard. Once we have this, rather than having to accept caveat emptor, we can all make our own assessments of whether the data are fit for purpose. This has to be an automated assessment, and unlike in 1995, the tools and computing power are now available to do this. It may need a 14-day free trial to make an assessment, if the data are not free, but once the assessment is made, it will then be possible to decide whether or not to pay, or how to much to pay (value) for those data. Suddenly the free vs. licensing debate becomes irrelevant.

If the rules expression language can't be created fast enough, we can move to the geographer's solution, the "pseudo-quantitative expression." Following the Amazon tradition of peer group review, the user community could assess the nominal value of a spatial data set for completeness, logical consistency, and positional, temporal and thematic accuracy. The Open Source community works well under the same model of peer group review and so we believe it should be interested in this approach.

If you are interested in extending the OSGeo into spatial data quality assessment or playing a role in standards development relating to spatial data quality we urge you to contact 1Spatial direct or through the OGC Working Group on Data Quality at:
http://www.opengeospatial.org/projects/groups/dqwg

Biobliography

1. The Cathedral & The Bazaar, Raymond, E.S., (2001) ISBN 0-596-00108-8
2. Spatial Data Quality, Wenzhong Shi, Peter Fisher, Michael Goodchild, (2002). ISBN:0415258359
3. Improving Operational Efficiency with Geographic Information, Finnish Ministry of Agriculture & Forestry, (2006). ISBN 952-453-301-4

Mike Sanderson
CEO, 1Spatial

Graham Stickler
Product and Marketing Director, 1Spatial

Steven Ramage
Business Development Director, 1Spatial
http://www.osgeo.org
info AT 1spatial.com

Acknowledgements

Various reviewers & the GRASS News Project

The *OSGeo Journal* is a publication of the *OSGeo Foundation*. The base of this journal, the LaTeX 2ε style source has been kindly provided by the GRASS and R News editorial board.

Journal online: http://www.osgeo.org/journal

OSGeo Homepage: http://www.osgeo.org

Mail contact through OSGeo, PO Box 4844, Williams Lake, British Columbia, Canada, V2G 2V8

ISSN 1994-1897