Assessing Fitness for Purpose for Re-

'The Industry Needs to Collaborate'

Data used on the web must be constantly monitored and rigorously tested to ensure fitness for purpose. The user must be guaranteed accurate results from online queries. This means the data used for analysis and decision-making must be accurate. This article looks beyond the map and focuses on how web users might access spatial data from multiple sources. The importance of data certification and spatial data quality management will also be highlighted.

By Mike Sanderson and Steven Rammage

Figure 1: Data Quality Improvement Cycle.

Value

In 1999 The Paper Industry Research Association (PIRA) valued geographic datasets assembled over the previous 10-15 years in the then European Community at 36bn. It estimated that it was double this value in the United States of America. Recent estimates indicate the value has risen to 100bn in the E25 community. Most of this geographic data was collected before GPS became ubiquitous.
Not only that but this asset needs to be used in support of sustainable policies and development across Europe. The ability to process these data automatically in situations not envisaged by the original collection programmes (we will term it re-use) becomes an essential goal of delivering on the i2010 agenda. Already spatial data quality problems are holding back the European initiatives. There is evidence that the heavily fragmented geographic community in Europe is failing to tackle interoperability and spatial data aggregation. The 2006 eContent+ programme focussing on making digital content more accessible, usable and exploitable was only able to support 3 GI projects with a value of 3.5m against a total available budget of 50m. Something is holding back a spatial contribution to the European knowledge economy.

e-Government Initiatives

The demand for increasingly accurate data will be driven by the need to automate spatial data processing. In order to deliver on e-Government initiatives or joined-up decision-making if you prefer that term, data from different sources need to be made available across the web and aggregated without human intervention. Interoperability is the start point. In order to carry out effective data aggregation activities and tackle spatial data quality issues the industry needs to collaborate. In 2003, Laser-Scan worked with Autodesk, Intergraph, MapInfo and Oracle on an industry initiative called Open Spatial Enterprise. This was a customer driven initiative based on customer requirements at the City of Winnipeg and Thames Water. A real-world interoperable spatial data management platform started to emerge. While the applications could all share data from the same central Oracle database, the initiative aimed to standardise differences in the handling of cartographic text or the orientation of points around a given topological definition of object geometry. This resulted in an industry standard for managing spatial, annotation, and attribute information. The resulting specifications are the subject of an OGC standard and have been adopted as a table structure in Oracle Database 10g. Subsequently other vendors like Bentley and Star Informatic joined the initiative.

Data Quality Issues

The initiative mentioned above was data centric as opposed to GIS centric. What does this mean? The mainstream IT industry has been tackling these issues for several years longer than the GI community. According to Oracle Magazine (March/April, 2005) US businesses lose $600 billion per annum on business data quality issues. The smart way for the GI industry is to make use of these experiences.
Han Wammes asked for the industry to revisit spatial data quality (Geoinformatics, March 2006). It can be graphically illustrated. Reading up on the attack on the Chinese Embassy in Belgrade in 1999, it is apparent that the CIA had access to maps of the area for 1992 and 1997. The 1992 map showed the embassy in a different location in the city. On the 1997 map an unlabelled building was selected as the target by comparing parallel road street numbering based on a visual inspection. Whilst the CIA was confident it had identified the Serbian Ministry of Supply, it was the Chinese Embassy that had re-located.

'Certifying'

The supply chain for spatial data re-use, starts with discovering data in registries across the web. Data in these registries needs more than

use of High-Quality Spatial Datasets

a statement about content and temporal validity. Organisations should be 'certifying' their spatial data quality to allow re-users of the data to make a determination of its fitness for purpose. The first phase of spatial data collection was driven by the need for automated map production. The current phase of spatial data usage needs to support the process of making critical business decisions electronically, with all the relevant and correct information at our fingertips. We need to be sure that it is fit for our particular purpose: because it is tested and certified to be so. This means identifying patterns in spatial data and providing data discovery tools to determine the rules behind those patterns.

The next step is to check rule conformance. This process provides meaningful data certification where data conformance can be expressed as a percentage of conforming instances. This allows the supplying organisation to quantify conformance and establish EQFM frameworks (www.eiqa.com/products/quality/model/index.asp) for investing in spatial data. It also allows the supplying organisation and the re-use organisation to assess the fitness for purpose that the data can be put to. The re-using organisation may be seen as an aggregator. It will change data for its own purposes. It may report some of these changes to the originating organisations, it may not. This data improvement cycle can be visually represented, see Figure 1.

Spatial Data Quality

Spatial data quality may be described as:

- Data quality overview elements: components of data quality which can only be described subjectively such as purpose, usage and lineage (that is: the history of a dataset);
- data quality elements: components of data quality, which can be expressed in quantitative form (see below).

The following data quality elements are defined in ISO 19113 – not all of these will be applicable to every dataset:

- completeness – "presence and absence of features, their attributes and relationships";
- logical consistency – "degree of adherence to logical rules of data structure, attribution and relationships";
- positional accuracy – "accuracy of the position of features";
- temporal accuracy – "accuracy of the temporal attributes and temporal relationships of features";
- thematic accuracy – "accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships".

So we have the basis for making the assessments we need for re-use. What we need now are:

- A framework for making assessments;
- Mechanisms for presenting data to the assessment framework.

Laser-Scan has developed a framework for making the assessments. It is called Radius Studio, see Figure 2. This is based on a common language interface and makes use of mainstream standards: W3C, OWL, ISO TC/211, OGC – GML and WFS.

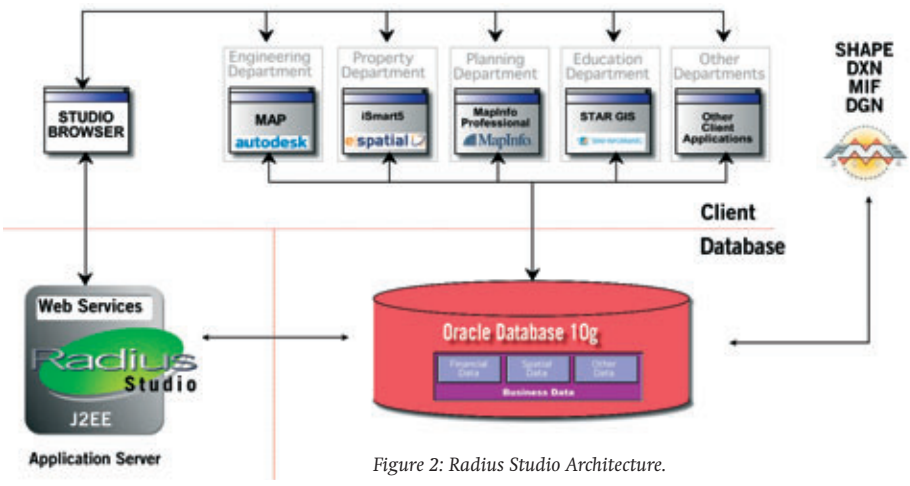

Figure 2: Radius Studio Architecture.

SWRL

There is much work still to be undertaken for data re-users in Europe to take advantage of the valuable geographic datasets already created. Further development of the semantic web rules language (SWRL) for use with spatial data constructs is necessary. Re-users will still be faced with the problem of semantic interoperability, or the difficulty in aggregating data that were collected and tagged using different vocabularies and different perspectives on the data. To achieve semantic interoperability, systems must be able to exchange data in such a way that the precise meaning of the data is readily understandable and the data itself can be converted or translated by any GIS or web mapping tool into a form that it understands. In addition for the re-users or aggregators there is a need to have access to a set of presentation standards. Web Processing Service (WPS) is still immature. It describes individual geoprocessing services accessible via the Web.

MapGuide Open Source

The recent open source announcement by Autodesk (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=6153839) and the formalization of the MapGuide Open Source site, which is hosted by the Open Source Geospatial Foundation, is an important initiative. It creates the opportunity to create a community. Is the value of public spatial datasets in Europe not worth creating a community that is focused on assessing the fitness for purpose for re-use? We believe it is.

Mike Sanderson is CEO of Laser-Scan, Steven Rammage works as a Product and Marketing Director for Laser-Scan. More information can be found at www.laser-scan.com or by sending an email to info@laser-scan.com.