SPATIAL DATA INFRASTRUCTURES: THE CHALLENGES AHEAD

M. Sanderson, M. Gregory, C. Tagg, C. Rose
1Spatial, Cavendish House, Cambridge Business Park, Cambridge, CB4 0WZ -
(michael.sanderson, martin.gregory, chris.tagg, claire.rose)@1spatial.com

KEY WORDS: Ontologies, SDIs, Master Data Assets

ABSTRACT:

INSPIRE is a given so attention turns to implementation. Of particular interest is the creation of Spatial Data Infrastructures (SDIs). In Great Britain (GB) we have been attempting to create a SDI for over 20 years. These trials have been driven by the need to co-ordinate activities around the maintenance of underground assets and the damage that occurs to plant and more importantly over the last decade communications disruptions. This paper reviews those experiences. The review considers data that have been collected by various departments within organisations and the use of topographic base map data against the current need to disseminate data to third party stakeholders by the internet. We consider the master data referencing needed to manage reuse in the new internet paradigm. Based on our experiences we propose a series of mechanisms that allow sensible reuse of the spatial data against fitness-for-purpose criteria.

INTRODUCTION

At present all of the GB utilities possess a considerable amount of detailed data on their assets, usually in the form of lines on maps. This statement has been modified from a publication over two decades old (Evins, Thomson & Ainsworth, 1984). In the regulated environment that exists today the main benefit from organising information is be able to report on the asset base. Operational activities, rehabilitation planning and communication with other parties are value added benefits (*ibid,* 1984). Interoperability is therefore not the main business driver for asset intensive businesses. The demand for the exchange of information driven by street works still requires seed funding (http://www.eswrac.org). The DTI and UK Water Industry Research are supporting associated research in the UK (http://www.ukwir.org/site/web/content/programme/current-programme?CatId=71#Project271).

Much work has been undertaken on SDIs in Europe over the last 5 years (Fullerton & Toth, 2006). The aggregation of local and regional datasets will become a key INSPIRE activity. Significant work has been undertaken on standards since the middle 1990s. The Open Geospatial Consortium has led to the specification of a Web Feature Server standard giving rise to a capability of viewing other organisation's data. The challenge is now to be able to reuse these data to make decisions in areas that the data were not originally collected for. In other words the data have to be assessed and quantified as to their fitness-for-purpose. This paper describes work related to the logical consistency of data sets. Matters related to assessing data as fit for purpose in the context consistency with real world objects are discussed in (Busch et al, 2004).

SDI work in the UK has been going on in excess of 20 years. Original work was carried out by the utilities at a number of trail sites. The trial sites have allowed some momentum to be maintained during changes to the topographic map base of Great Britain (Higgs & Malcolmson, 2005). For land parcel applications two basic Master Data[1] (McGuffog, 2004) components exist in the TOID™ and the municipality's property reference number (the LLPG; see Figure 1). The changes in the topographic map base evident in MasterMap™ (GB), TOP10NL (The Netherlands) and AAA (Germany) have created much larger databases in which the ability to load changes (as opposed reloading the entire map base) is a prerequisite for effective data management. Although there are similarities between topographic and linear master data based on the street works initiatives cited above it is believed that there has been little work undertaken on the interoperability of utility asset data as part of an SDI.

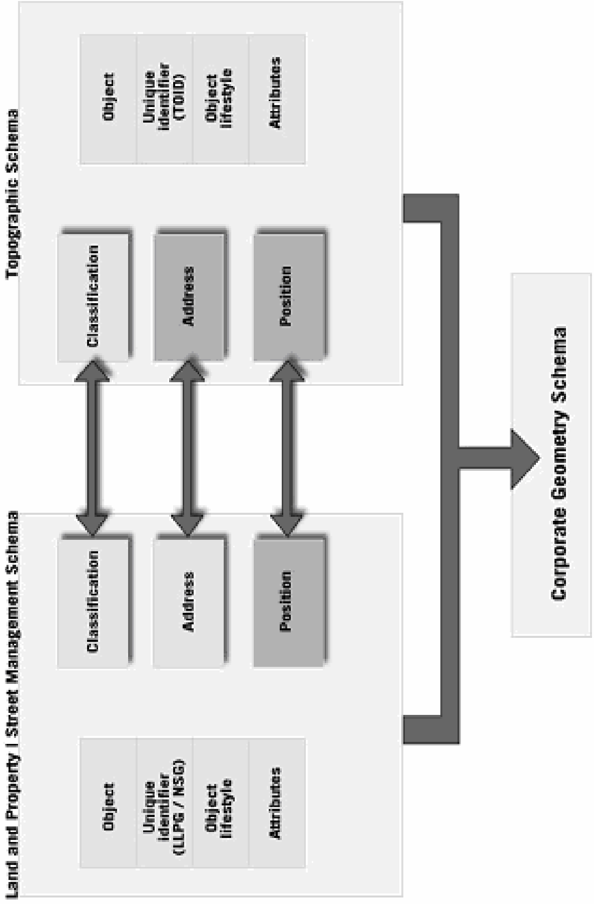


Figure 1. Land Parcel Master Data

[1] The basic premise behind a Master Data identity is that a de facto numbering system exists which guarantees uniqueness and the numbering has no inbuilt meaning.

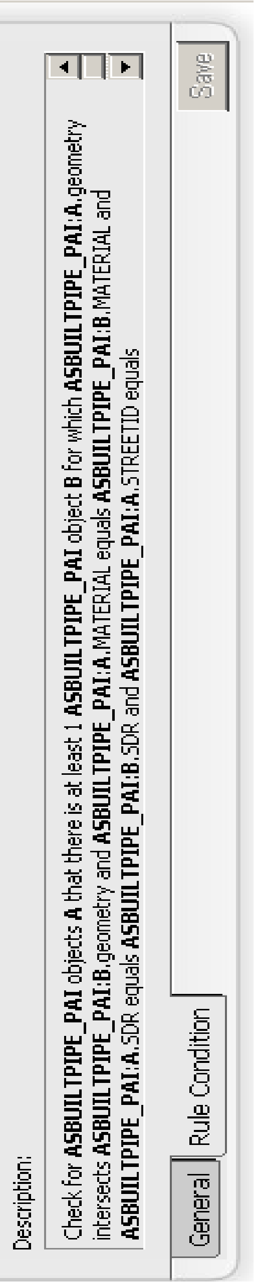
MAIN BODY

In the GB regulated utility businesses the upgrade to the large-scale topographic map base has coincided with a positional accuracy improvement programme. The utilities are content with a point and line topographic approach as they can assert their data to be logically consistent in relation pavements (kerb edges). The exact position in three dimensions is in any event largely unrecorded. The asset base is however used by the regulator to assess charging regimes. In other words, lengths of main or cable cannot be altered by more than a certain percentage as a result of positional changes; otherwise the charging formulae are impacted. In addition the cost of the original collection (PIRA, 2000) and the positional correction run into millions of euros for each utility. It follows therefore that additional business benefits must be delivered as the asset data are reused against the new topographic map base. Otherwise the utilities are content to use a frozen temporal version of the map base. The first major challenge is therefore not a technical challenge but an institutional one. This may be addressed by identifying benefits. Suggested benefits are:

- the data cannot be changed materially such that there is an effect on the regulatory framework
- a backlog of ‘as built’ records cannot be allowed to accumulate (in other words the data must remain in situ; some data are commercially sensitive and are not allowed offsite in any event)
- some aspects of data quality improvement as part of the re-alignment process should be delivered (the most obvious being related to connectivity)
- an ability to communicate the position of the assets with other utilities and the highways authority who may or may not have repurposed their data.

In order to reuse the asset data, correction must be automated in order to avoid excessive costs. The issues that need to be addressed are both geometric and alphanumeric. In addition the problem space gets more difficult as the asset data almost certainly have to be transformed from their original GIS toolset to another. These benefits are associated with network connectivity and ontologies (defined in Genesereth & Nilsson, 1987). The ontologies can be used in corrective actions on the data and also used in conjunction with external applications that describe what is represented as an example to billing applications. The reuse of spatial data within the positional change space therefore requires a more comprehensive approach then just using shift vectors to move the asset data. The challenges we investigated are described below.

The first challenge is to identify the ontologies and the business rules. 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. Rules may be discovered using an analysis of dominant statistical patterns in the data. This was accomplished by 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. In order to address the second challenge, we have chosen to store the rules, not as XML or as program code, but as a common language interface that incorporates OGC spatial operators (see Figure 2).



Description:

Check for **ASBUILTPIPE_PAI** objects **A** that there is at least 1 **ASBUILTPIPE_PAI** object **B** for which **ASBUILTPIPE_PAI:A**.geometry intersects **ASBUILTPIPE_PAI:B**.geometry and **ASBUILTPIPE_PAI:A**.MATERIAL equals **ASBUILTPIPE_PAI:B**.MATERIAL and **ASBUILTPIPE_PAI:A**.SDR equals **ASBUILTPIPE_PAI:B**.SDR and **ASBUILTPIPE_PAI:A**.STREETID equals

Figure 2. Common language Rule Definition

Geometric Checks

A number of geometric checks can be run on the source data to identify where digitising errors have been introduced. Such errors as spikes and kick-backs can be readily identified using standard GIS functions. Simply, a spike is defined to be 3 consecutive points (A, B, C) such that:

1) The distance AB is less than the distance BC
2) The sine of the angle ABC is less than a maximum value, which may be specified by the second parameter.

3) (Optionally) the distance AB is less than a maximum "length" value specified by the third parameter.

Once identified a spike is removed based on a number of parameters:
Parameter 1: The input geometry.
Parameter 2: (Optional) The maximum value for the sine of the angle in the spike. If omitted, this defaults to the sine of 1 degree.
Parameter 3: (Optional) The maximum length of the spike.
Return value: The resulting geometry with spikes removed.

Connectivity Checks

A sample of fixed-line telecommunications data can be used to explain the approach adopted. The following assumption was made: that all cable features should be connected to at least one other cable feature and that terminal features should be connected a cable feature. Rules can then be implemented (as shown in Figure 2). Once implemented breaches can be examined and rectified. Improving the connectivity yields an improved shifting capability and enhances the data for network modelling purposes.

Asset Length Checks

Given that a business benefit is to minimize changes in the overall asset value, it is necessary to establish the correspondence of alphanumeric data to geometric data before any shifting is undertaken. The following rules were applied to the telephony data set:

Rule: Check that the CABLE_LENG attribute for a Cable feature is within 10% of the calculated length of the digitized cable.

District Id	Feature	Rule Check	# Checked	% Pass	% Fail	# Features Failed
AN_ID 2154	Cable	Rule D – Cable length	3447	31.01	68.99	2378

Table 1. Results of Cable Rule

From the results above, we can see that the stored length attribute for cable features does not match the actual geometric length for nearly 70% of the data.
Similarly missing values may occur which will distort the asset length and a rule may be put in place to check for missing values or arbitrary values assigned to indicate that the length is missing. These are part of the fundamental quantitative assessments and measures for spatial data quality. Such statements are required to successfully implement Spatial Data Infrastructures irrespective of the domain

The use of ontologies

If we wanted to check for spikes in the linear features, a "check for spikes" rule could be written for each linear feature (cables, fiber cables, air pipes). This is not efficient. However, using the ontology support a logical model is created that groups features together: Thus a class called linear assets is developed encapsulating cables, air pipes, cables, mains, sewers. This adds significant value when a mutli-utility operation owns more than one asset in a region. The spike rule is then applied to the linear asset class where the linear asset class is the route node of the rule.

Since all of the data stores we have been investigating are external data stores we sought to provide support for externally defined ontologies, which describe the structure of the data. in a specific data store. This was achieved by interfacing with the open source Jena ontology library (see http://jena.sourceforge.net/), allowing ontologies in various formats such as RDF and OWL to be read.

Shifting

Where shift vectors are supplied we have found that a Delaunay triangulation algorithm is the best approach. It allows for real world change and copes better where ground features are sparse. It has the advantage of being computationally light and so allows fast processing. The more interesting case is where shift vectors do not exist in moving from one topographic map base to another. On the assumption that the new map base is more up to date and more accurate, then a suitable methodology, which starts to build an SDI for utility data can be developed from the following steps:

For the asset base, topology is created on the original map base. Tolerances of 1m were utilised. Topology is also created on the new map base. All the nodes in each topology set that are connected to 3 or more edges were given a unique identifier. For each qualifying node in the new map base the nearest qualifying node in the old map base was selected to create a node pair. A restriction of 75m radius was put on this nearest neighbour search. Specialist routines were written to cope with densely packed nodes.

For each pair of matching intersections a shift vector was created starting at the old map base intersection and ending at the new map base intersection. This network of vectors became the initial surface for shifting the assets. The first results were not very good because not enough points were captured. Additional shift vectors were added between road elements – based on the positions where vectors had already been created.

Post Migration Checks

The connectivity checks are re-run. In addition geometric checks are run to ensure that linear features crossing (or not crossing streets) in the old map base behave in the same manner against the new road network. Other additional checks to ensure that assets do not cross water features (unless via a bridge) can be added. After manual alteration of the remaining the unique identifiers from the asset base can be indexed to the street intersections.

CONCLUSION

Typically 90% of the repurposing can be carried out automatically. The metrics derived during the experiments showed that approximately 14 features per hour could be shifted manually. Once the rules base had been created 14,000 features an hour can be repurposed. There is a set-up task to create the rules base. The work highlights the challenge to quantitatively assess logical consistency (data quality) before trying to establish SDIs in this domain and others where geometric consistency and data integrity, such as connectivity are important. The benefit from developing the ontology goes beyond the shifting exercise. This is the area where more work is required to determine where else these rules can be used inter-enterprise to create further value. The rules can then stand-alone from the shifting exercise and be used for guaranteeing data quality over a longer time horizon than the migration exercise demands.

References

Busch, A. ; Gerke, M. ; Grünreich, D. ; Heipke, C. ; Liedtke, C. E. and Müller, S.(2004). Automated Verification of a Topographic Reference Dataset: System Design and Practical Results: IntArchPhRS. Band XXXV, Teil B2. Istanbul, S. 735-740

Evins, C ; Thomson, R.W. and Ainsworth, R.G. (1984). Organising Information on Water Distribution Networks. WRc, Swindon, Wiltshire.

Fullerton, K. and Toth, K. (2006). ESDI: From Inspiration to Implementation. 12th EC_GIS Workshop, Innsbruck, Austria. ISBN 92-79-02083-8

Genesereth, M. R. and Nilsson, N. (1987). Logical Foundations of Artificial Intelligence. Morgan Kaufmann Publishers: San Mateo, CA.

McGuffog, T. (2004) A General Theory of Value Chain Management Data.UKP.eb.

PIRA International. (2000). Commercial Exploitation of Europe’s Public Sector Information (Final Report). For the European Commission, Directorate General for the Information Society.