Metadata Management of Geospatial Data for Earth System Engineering Research

Newcastle University
School of Civil Engineering and Geosciences

Summer Vacation Scholarship 2011 **Gavin Montgomery**

Aim

To develop a database schema capable of processing vast spatial datasets whilst present lineage/provenance information regarding the history of the data. All using free to use open source software.

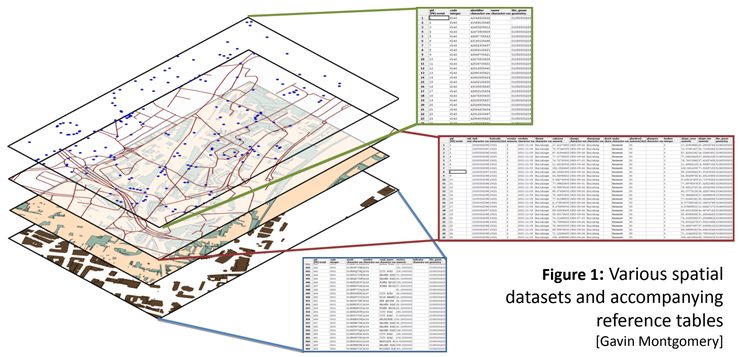

Figure 1: Various spatial datasets and accompanying reference tables [Gavin Montgomery]

Introduction

The Geomatics Data Server has been created within the school of CEG to handle all Geomatics data held by staff and researchers. The data is an assortment of various types including **satellite imagery, aerial photography, laser scans** and **vector datasets**. The importance of metadata requires that all the data uploaded to the server is correctly referenced allowing an uninitiated user to access and utilise the service.

Objectives

- Create a database to support multiple spatial tables for the Tyne and Wear region
- Develop infrastructure to automatically create metadata for newly added geometry
- When processing datasets extract (from a .csv file) theoretical lineage/provenance information
- Construct framework to associate related spatial data

The Metadata Table

The first task was to construct a metadata table which was capable of holding descriptive information regarding the data that was being held on the server. Spatial data is stored with a **geometry** column which holds coordinate data as well as various other attribute columns; for example the 'a_roads' dataset could hold the road name as well as entities regarding traffic usage. All these attributes can be held within the table; All table titles were referenced by the 'metadata' table.

The table was constructed and scripted to automatically update when a new spatial dataset was imported and insert data values.

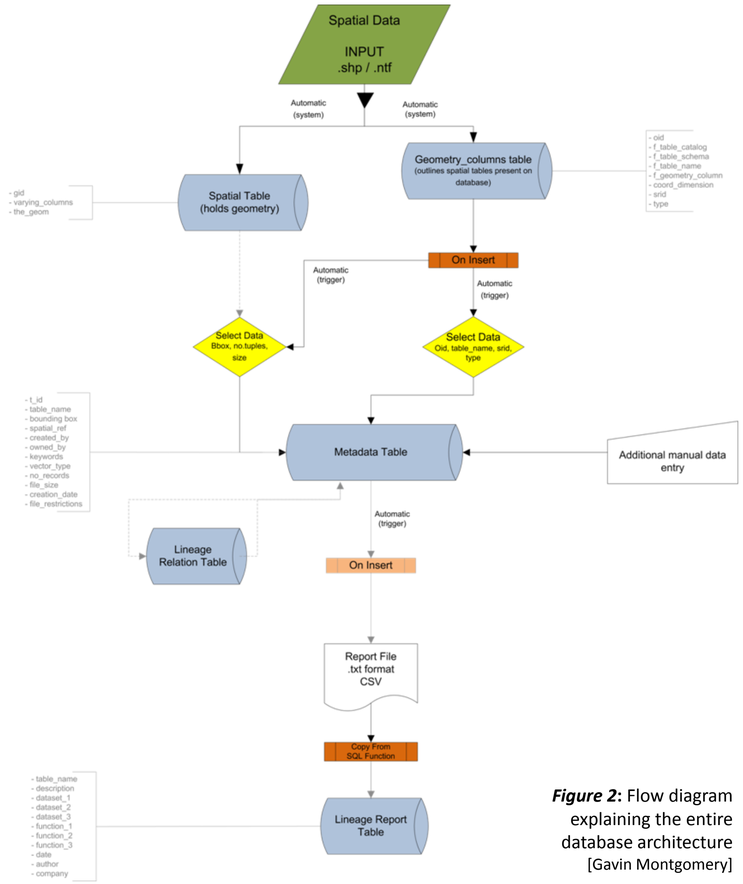

***Figure 2*:** Flow diagram explaining the entire database architecture [Gavin Montgomery]

Lineage & Provenance Methods

The term data provenance broadly refer(s) to a description of the origins of a piece of data and the process by which it arrived in a database (Buneman 2001). To relate various spatial data tables an ideal scenario was established which would be to load the new geometry with an accompanying .csv file.

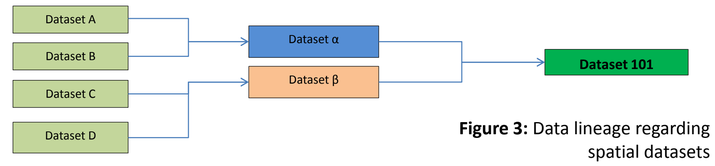

Figure 3: Data lineage regarding spatial datasets

It may be desirable to look back upon completion of 'Dataset 101' and understand what processes or types of geometry were used to create the final completed product.[Figure 3]

Connections between datasets were established through a relational table which referenced the 'oid' of the tables. One dataset can be fashioned by many tables so the lineage table can hold multiple relations; these are foreign keys referencing the original metadata table.

Conclusion

Design of a spatial database has produced many challenges facing inter-table relationships and automatic data creation. Metadata management is a vital portion of data storage; automatic methods are possible though the table still lacks various manually curated items.

This project has however proven that a large database can be produced and managed using a comprehensive open source platform, the design schema should allow for the geomatics department process spatial data on the server.

I would like to thank Dr Stuart Barr for his support & guidance and David Alderson for his ever committed technical assistance

Reference: "Data Provenance: Some Basic Issues" Buneman *et al (2001)*

gavin.montgomery@ncl.ac.uk

Geomatics Data Server: ce-gs402.ncl.ac.uk