Rule based quality control for automated generalisation and conflation
Rules

Controlling processes via accessible business rules
Rules to assess data quality
Rules based language to drive data integration and cleaning
Rules based language to drive the generalisation process
Writing rules: an iterative process
Use case

Automated validation within the OSGB GenIE System
Core principles

- Parallel processing of data in partitions to maximize performance
- Maximum automation
- Minimum editing
- Known quality of output - ‘good enough’
A new 'street-level' dataset designed to be the most detailed open data mapping product available for Great Britain from OS.

OS Open Roads
A generalised road network product, enabling high level network queries and the sharing of data.

OS Open Rivers
A generalised network product for Great Britain's rivers and watercourses.

OS VectorMap District
A customisable 'District level' dataset designed to support a wide range of customer applications that utilise geographic information.
Automated validation

- **Rule-based test framework** used to ensure end-to-end process is not broken during development.
- **Data** is parallel processed in partitions.
- **Validation** on each partition as data is committed.
- **Rule-based validation** to identify non-conformances.
- Non-conformances are categorised as either 'critical' or 'warning'.
- **Edit jobs** are automatically created for 'critical non-conformances'. Warnings used to provide a quality statement.
- Automated feature counts are checked against baseline figures to look for processing failures.
- **Maximise production efficiency** ensuring the data is valid in this interim store minimises the need for manual editing after product specific generalisation.
- **Re-validation** following product specific generalisation.
- **Ensuring consistency and validity** in the published products removes the need for expensive manual ‘eye-ball’ of the data.

**Key point of validation**

**Errors and causes of non-conformances**

- Errors in the output data can occur due to:
  - Errors in the source data – errors are fed back to a data improvement team.
  - Scenarios in source data not previously identified – these are verified against the data specification, and referred to the development team.
  - Errors in a custom generalisation action – these are reported as defects back to the development team.
  - Failure of a standard generalisation action – these are fed back to the software supplier (1Spatial).
  - Environmental failures – these are reported back to the appropriate team.
Use Case

Rule based
Data conflation
Aim: build a thematic map from multiple sources

TOP10NL data
- Authoritative data
- Not up to date

OSMdata
- Up to date roads
- Sped limit attributes

Can we make a map based on Top10NL data, showing fast through routes in the city?
Data Conflation

Input data from multiple sources

Clean Sources → Match → Transfer → Clean Results

output data
Identify and bring missing data

Transfer all OSM motorways which haven’t got a match in TOP10NL
Clean conflated data
Results after generalisation and styling
Where next?

1. Make a reusable set of tools (rules, actions) to facilitate making thematic maps from multiple sources semi automatically

2. Machine readable description of the tools for them to be used automatically
   ⇒ Piece of the jigsaw for building an automated on demand mapping service
   ⇒ Link back to formal descriptions of map content based on scale and purpose, spatial relationships between features
Summary

Editable rules proved useful to the end user for:

• Adapting the logic used during generalisation to their needs
• Managing quality throughout the system

⇒ Delivering a flexible, efficient and reliable production system

Rules are also key for data integration

• Adapt the matching process to the data available
• Check and enforce quality at all stages of the process
Next week at the ICC

1Spatial stand at the technical exhibition

Barbara Greiner
Maps and Models - The Art of automated Map Production using flexible Tools and Workflows

Poster session:
Wednesday, July 5, 3:50 PM - 6:50 PM

Oral presentation:
Session: Map Production and GeoInformation Management
Thursday, July 6, 2017, 2:50 PM - 3:10 PM

Nicolas Regnauld, Paul Duré
Producing maps from multiple sources using automated data conflation, generalisation and advanced styling

Oral presentation:
Session: Generalisation and Multiple Representation
Thursday, July 6, 2017, 5:10 PM - 5:30 PM

G.Touya, J. Berli, I. Lokhat, N. Regnauld
Experiments to Distribute Map Generalization Processes

Oral presentation:
Session: Generalisation and Multiple Representation
Thursday, July 6, 2017, 8:30 PM - 8:50 PM
Questions?

Nicolas Regnauld
Consultant in Geospatial data processing
nicolas.regnauld@gmail.com

Derek Howland
Operations Domain Leader – Automated Generalisation
derek.howland@os.uk
+44 (0) 2380 055225