Data Validation
Online References

Submitted To: Program Manager
GeoConnections
Victoria, BC, Canada

Submitted By: Jody Garnett
Brent Owens
Refractions Research Inc.
Suite 400, 1207 Douglas Street
Victoria, BC, V8W-2E7
jgarnett@refractions.net
Phone: (250) 885-0632
Fax: (250) 383-2140

TABLE OF CONTENTS

1 INTRODUCTION

Spatial data validation and GIS Quality Assurance tools is an important and vibrant area of development. In this document we will examine online references for validating attribute data using existing tools.

2 DATA VALIDATION TOOLS

2.1 Automated QA

This is a representative tool of the vast majority of the validation tools available online – these tools perform a number of specific tests and have no facilities for extension.

This tools checks:

- Structural integrity of the spatial data set
- Spatial features and attributes
- Database integrity
- Cartographic annotation
- Edge-matching across tiles

2.2 Dog Creek QC Pro

QC Pro is targeted at validates coverages. Several consulting companies offer services based around its use.

This tool offers several innovations worthy of note:

- Configuration is specified by prototype. The tool is loaded with a "template" coverage data set that is used to create a Design Registry.
- Tests are arranged into test suites, called exams, for management and reporting purposes.
- The results are generated as html web pages.

The tool is also willing to use ARC/INFO tests in exam and report specifications.

Capabilities:

- Cleanliness
- Cover spec: topology matching, table matching, projection matching, precision matching
- Existence
- Table Spec: attribute checks

- Consistency using Arc/Info CONSIST command
- Duplicates
- Physical Consistency: Intersect Errors, Label Errors, Node Errors, Regional Errors, Extents
- Referential Integrity
- Validity: valid values, null zeroes, range, equality, unquity

Limitations:

- A bulk data tester, does not test per transaction

2.3 GeoData Sentry

This tool is similar in design to QC Pro (it is produced by the same company) with a focus on geodatabases (rather than coverages).

Tests are organized into test suites. Test can be generated based on geodatabase domain, or composite relationship information.

Has a similar set of capabilities for checking attribute validity, referential integrity and spatial relationships:

- Single Column Attribute Tests: unique values, coded domains, coded ranges, single values, null values, non-standard values, column format, column length
- Multiple Column Attribute Tests: composite column values, custom SQL query
- Referential Integrity Tests: general table relationship
- Spatial Relationship Tests: distance

Limitations:

- ArcGIS geodatabase backed by Oracle and SQLServer

The tool is modular in approach allowing the release of Runtime, Professional and a Community Edition and a "Test Inventory" of testing plug-ins.

2.4 ESRI's GIS Data ReViewer

The GIS Data ReViewer is a data quality control management application that simplifies many aspects of automated and visual spatial quality control tasks.

Capabilities:

- Make and log corrections to data
- Verify corrections made to data
- Locate errors in data capture/attribution
- Log error information accurately
- Coordinate the data review effort

- Perform batch validation of a geodatabase
- Eliminate the paper trail associated with error files by storing data error information in a database.

Liabilities:

- Limited for use with ArcInfo and requires ArcView and ArcEditor.

2.5 ESRI's ArcGIS 8.3 Geodatabase

ESRI's ArcGIS 8.3 Geodatabase includes toponymy based validation checks.

Capabilities:

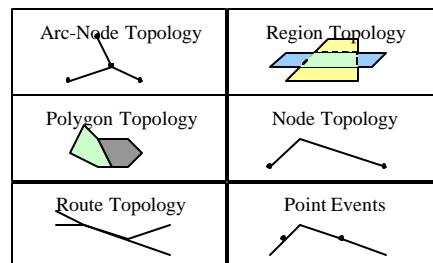- Declare and place limitations on how features to share geometry

Figure 1 - Sharing Geometry

- Create features from unstructured geometry using snapping and aggregation.
- Constraint Support:
  - Relationships between features
  - Validation rules
  - Logical Networks

All of these capabilities are expressed as an integrity rule set defining the topology. Topologies can extend across Feature Types with the same spatial reference system. Feature Types can only be included in one Topology or Geometric Network.

Topology is defined in Geodatabase using:

- **Cluster Tolerance:**

  The behavior of snapping operations is controlled using Cluster Tolerance. As an example snapping points together that are 20 cm apart is acceptable if the data accuracy is 2 meters.

- **Ranks:**

  When merging shared points, edges or areas, a concept of Feature Type rank is used as a tiebreaker. Feature Types with the same rank are averaged.

- **Rules:**

  Rules define allowable relationships in a topology. Rules serve as the definition for topological integrity.

Topology Rules	Domain
Area boundary must be covered by boundary of	Polygon/Polygon
Boundary must be covered by	Polygon/Line
Contains Point	Polygon/Point
Endpoint must be covered by	Line
Must be covered by	Polygon/Polygon
Must be covered by boundary of	Line/Polygon, Point/Polygon
Must be covered by endpoints of	Point/Line
Must be covered by feature class of	Line/Line, Polygon/Polygon
Must be properly inside polygons	Point/Polygon
Must be single part	Line
Must cover each other	Polygon/Polygon
Must not have dangles	Line
Must not have gaps	Polygon
Must not have Pseudo-nodes	Line
Must not intersect	Line
Must not intersect or touch interior	Line
Must not overlap	Line, Polygon
Must not overlap with	Line/Line, Polygon/Polygon
Must not self intersect	Line
Must not self overlap	Line
Point must be covered by Line	Point/Line

Geodatabase takes a user centered approach to toponymy, focusing on the editing process rather than data consistency. Rule violations are flagged for later correction.

Refractions RESEARCH

GeoConnections

2.6 ArcFM Energy

A class of validation tools exists that are targeted at specific markets. A representative example of this is ArcFM Energy extension for ArcInfo that is customized for use by the Energy sector.

Capabilities:

- Database schema definitions

- range and domain value for attributes

- Default values for attributes

- Connectivity checking

- Support for split and merge operations

- Checking of Specific relationship: high voltage line can only connect to a low voltage line through a transformer

- Network Completeness: for an electric network or gas distribution network

Liabilities

- Not a general solution

- An to ArcFM / ArcCatalog

- Unclear if transaction based validity checks are supported

Similar products exist for meeting defense GIS standards.

2.7 Production Line Tool Set GIS Data Reviewer

This ESRI tool supports manual and batch data review of GIS data. The tool maintains a Valid Value Tables and Condition Tables in the database on a per feature class basis. This configuration information is then maintained under the same transaction controls as the data itself.

The tool performs a number of topology checks (closed polygons, overlap, duplicate vertices) and spatial analysis checks (size, duplicate points w/ tolerance, overlaps between feature types).

The focus is on flagging "failed" data for manual review, to this end the tool can be used interactively by AQ Staff and in batch mode. The calculation of summaries on a random sample of data is also supported.

2.8 ENC Analyzer

www.sevencs.com

ENC Analyzer takes a very specific domain, hydrographic charts, and is able to focus on providing a high value tool.

- Provides support for attribute validation.

- Generates warnings for suspect data

Refractions RESEARCH

GeoConnections

This tool serves as finally tuned example of what is possible for a specific domain.

2.9 Geostatistical Analyst (ArcGIS 8.1 Extension)

The Geostatistical Analyst extension offers the following capabilities:

- Spatial Regression – establish relationships between layers

- Spatial Interpolation – reconstructing data

- Smoothing - looking for patterns in complete data

- Grouping - classification

- Prediction

Of specific interest to us is using these techniques for prediction – we can compare incoming transaction to our predicted expectations and provide warnings (or even errors) when outliers are encountered.

3 REFERENCES

Automated QA,
http://www.farragut.com/automate.htm

Dog Creek QC Pro, GeoData Sentry
http://www.laurelhillgis.com/dogcreek.htm

ESRI ArcNews Reprint, Vol. 24 No. 2, Summer 2002
http://www.esri.com/library/reprints/pdfs/arcnews_arcgis83-brings.pdf

Working With Geodatabase Topology,
http://www.esri.com/library/whitepapers/pdfs/geodatabase-topology.pdf

ArcGIS 8.3 Geodatabase Topology Rules,

http://www.islem.com.tr/haberpic/pdf/TopoPoster2_92363.pdf

GIS Data ReViewer,
http://www.esri.com/software/reviewer

ArcFM Energy
http://www.esri-portugal.pt/mercados/documents/arcfmenergy.pdf