Understanding Faulty Data

Methods to Capture, Report, and Score Data Cleanliness

Candidate: Nell S. Rose, GISP
Advisor: Ryan Baxter
Who am I?

Neil S. Rose, GISP

GIS Manager, City of New Braunfels, TX

- BA Sociology
- BA Applied Geography
- Grad Cert, GIS
- Grad Cert, Geospatial Programming and Web Map Development
- 8 years in local government GIS
- Employs a high level of python scripting
- Automate, automate, automate
- Professional interest in data cleansing automation
The Agenda

**Introduction and Background**
What is data? Why is it important? How does it live? What makes data good or bad?

**Goals, Objectives, Methodologies**
Identify, report, and score data cleanliness using python solutions

**Anticipated Results**
What will work? What won’t?

**Timeline**
Project completion and presentation
Introduction and Background

What is data? Why is it important? How does it live? What makes data good or bad?
What is “data”? 

- **Data quantifies or qualifies a phenomena**
  - Temperature, location, cost, distance, color

- **Data != Information**
  - Information is derived from the analysis and interpretation of data

- **GIS data isn’t special, only spatial**
  - Vector and raster data structures

- **Vector consists of points, lines, and polygons**
  - Acquired by heads up digitization or device capture

- **Raster consists of a grid of values**
  - Represents a continuous view of real-world phenomena

- **Attributes contextualize the spatial**
  - Non-spatial data that describes the spatial feature
Why is data important?

<table>
<thead>
<tr>
<th>Improve people's lives</th>
<th>Make informed decisions</th>
<th>Stop molehills from turning into mountains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get the results you want</td>
<td>Find solutions to problems</td>
<td>Back up your arguments</td>
</tr>
<tr>
<td>Stop the guessing game</td>
<td>Be strategic in your approaches</td>
<td>Know what you’re doing well</td>
</tr>
<tr>
<td>Keep track of it all</td>
<td>Make the most of your money</td>
<td>Access the resources around you</td>
</tr>
</tbody>
</table>

Council on Quality and Leadership
From planning to destruction, the lifecycle of data

<table>
<thead>
<tr>
<th>From planning to destruction, the lifecycle of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Planning</td>
</tr>
<tr>
<td>Data Generation</td>
</tr>
<tr>
<td>Data Collection or Acquisition</td>
</tr>
<tr>
<td>Data Processing</td>
</tr>
<tr>
<td>Data Storage or Preservation</td>
</tr>
<tr>
<td>Data Management</td>
</tr>
<tr>
<td>Data Analysis</td>
</tr>
<tr>
<td>Data Publishing or Sharing</td>
</tr>
<tr>
<td>Data Visualization</td>
</tr>
<tr>
<td>Data Interpretation</td>
</tr>
<tr>
<td>Continual Actions</td>
</tr>
<tr>
<td>Archiving/Destruction</td>
</tr>
</tbody>
</table>
Data Planning

The First Step

- Why is the data important?
- How is it being collected?
- How will it be stored?
- Determine data parameters
  - Schema
  - Storage
  - Access
  - Maintenance
Data Generation, Collection, or Acquisition

Data Planning

**Data Generation**

Data Collection or Acquisition

- Generating data is more applied to big data
- Large, unstructured datasets
- Doesn’t work well for GIS without data wrangling
  - *Not considered a core phase of the data lifecycle*

**Collection/Acquisition**

- Data collected through field operations, digitization, or from an external source
- Field operations and digitization draw directly from Data Planning
- Acquisition may require data wrangling
Data Processing

Data Planning
Data Generation
Data Collection or Acquisition
Data Processing
Data Storage or Preservation
Data Management
Data Analysis
Data Publishing or Sharing
Data Visualization
Data Interpretation
Continual Actions
Archiving/Destruction

<table>
<thead>
<tr>
<th>Also known as…</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data wrangling</td>
</tr>
<tr>
<td>• Data preparation</td>
</tr>
<tr>
<td>• Data munging</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How to process data…</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Transforms raw data into desired data through:</td>
</tr>
<tr>
<td>• Cleaning</td>
</tr>
<tr>
<td>• Structuring</td>
</tr>
<tr>
<td>• Enriching</td>
</tr>
<tr>
<td>• Differs from data cleansing</td>
</tr>
</tbody>
</table>
Data Storage or Preservation

- Builds data storage based on generated and wrangled data
- Can result in schema, field naming, accuracy, precision, and naming convention errors
- Not considered a core phase of the data lifecycle
## Data Management

<table>
<thead>
<tr>
<th>Data Planning</th>
<th>Jeanette Wing</th>
<th>Sanborn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Generation</td>
<td>• Optimized storage process</td>
<td>• Analysis</td>
</tr>
<tr>
<td>Data Collection or Acquisition</td>
<td>• Varies based on data generated</td>
<td>• Data updates</td>
</tr>
<tr>
<td>Data Processing</td>
<td>• <em>Not considered a core phase of the data lifecycle</em></td>
<td>• <em>Not considered a core phase of the data lifecycle</em></td>
</tr>
<tr>
<td>Data Storage or Preservation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Data Analysis
- Data Publishing or Sharing
- Data Visualization
- Data Interpretation
- Continual Actions
- Archiving/Destruction
Data Analysis

- Through analysis, derivative data is created
- Initial data input is not changed
- *Not considered a core phase of the data lifecycle (but is closely related)*
Data Publishing or Sharing

Providing access to data...

- Whether by open data, ftp, REST, or other means of sharing
- Data can be viewed, downloaded, and accessed
- Data doesn’t change
- *Not considered a core phase of the data lifecycle*
Data Visualization

Seeing the data…

- Symbolizes the data, whether on a map, chart, or graph
- The data doesn’t change
- *Not considered a core phase of the data lifecycle*
Data Interpretation

- Provides explanation to data analysis and visualization
- The data doesn’t change
- Not considered a core phase of the data lifecycle
Continual Actions

Data Planning
Data Generation
Data Collection or Acquisition
Data Processing
Data Storage or Preservation
Data Management
Data Analysis
Data Publishing or Sharing
Data Visualization
Data Interpretation

**Continual Actions**

Archiving/Destruction

---

**Repeatable processes...**

- **Describe data**
  - Metadata creation
  - Data dictionaries
  - Data discovery

- **Data cleansing**
  - “ensures that data are properly collected, handled, processed, used, and maintained at all stages” USGS

- **Data security**
  - Backups
  - Cyber security
Archiving/Destruction

The Last Step

• What to do with data when it’s being replaced or becomes irrelevant?
• None of the sources reviewed discussed this step
• Depends on organizational policies and practices

How to…

• There are several ways to archive or destroy data:
  • Stored as a read-only
  • Transferred to external media
  • Server snapshot
  • Deleted
• Census vs Municipal
The Core Data Lifecycle

Data Planning
Data Acquisition
Data Wrangling
Maintenance Operations
Data Archival and Destruction
o Having bad data results in bad analyses, maps, interpretations, and results

o This can be prevented by utilizing data maintenance operations:
  • Backups
    • Prevents loss of data caused by degradation, accidents, corruption, or malicious actions
  • Documentation
    • Metadata creation and management
    • Data dictionaries
    • Keyword data discovery
  • Quality management
    • Data cleansing
    • In-place tools
    • Scheduled operations

Garbage In, Garbage Out

Having data is good, having good data is better
Existing Tools

Topology
Rules set by the user to define a geospatial relationship

Domains
Set inputs for specified fields

Data Reviewer
Semi-automated data review to identify errors and issues

Attribute Assistant
Aids users in data entry based on intersecting features
Goals, Objectives, Methodologies

Identify, report, and score data cleanliness
Identify Known Issues

- Naming conventions
- Field name truncation
- Alias utilization
- Field attribute policies
  - NULL vs blank or 0
  - Empty data
  - Domain likeness, duplication, use
  - Field type vs data entered
- Metadata
- Bad/illegal characters
- Leading and ending spaces
- Double (or more) spaces
- Mixed coordinate systems
- Reserved words
- Hosted feature last edits
- Published feature metadata
  - Naming, summary, description, terms of use, tags, credits

Your input needed!

- Don’t see an issue you deal with on this list?
- Let me know about it!

Don’t see an issue you deal with on this list?
Let me know about it!

Your input needed!
Determine Solutions for Reporting Known Issues

Conceptual Solution Examples

**Reserved Words**
- Using python and arcpy to compare a list of field names to the list of reserved words
- Using a looped if-then statement to flag uses of any reserved words

**Naming Conventions**
- Types of cases...
  - flatcase
  - UPPERFLATCASE
  - lowerCamelCase
  - UpperCamelCase
  - snake_case
  - SCREAMING_SNAKE_CASE
  - Camel_Snake_Case
- Using python, regex, arcpy, and machine learning to identify whether the naming convention matches the user-selected convention

**Mixed Coordinate Systems**
- Using python and arcpy to describe the spatialReference property and flag datasets that don’t match the user-selected coordinate system
Scoring Cleanliness, Summary Report and Appendix

Flag and report the issues
- The tool will flag all known issues based on what the user wanted to run
- A summary report template will be filled in with the appropriate data about the issues observed

Scoring cleanliness of data
- Using the reported issues, score the cleanliness of data
- Not yet determined formula for weighting and scoring issues, errors, and inconsistencies

Provide appendix of issues
- To aid the user, an appendix of all issues will be generated
- The appendix will include feature dataset, feature, field name, and OID, grouped by issue
Anticipated Results

What will work? What won’t?

Unique Solutions for Unique Issues

Simple Solutions
- Reserved Word List Compare

Complex Solutions
- Naming Conventions

Philosophical Choice
- NULL vs Blank

Data Cleanliness Score Formula
Timeline

Project completion and presentation
Timeline
From starting proposal to potential presentation venues

- **Begin Proposal**: Aug, 2020
- **Present Proposal for Feedback**: Oct, 2020
- **SCAUG or TAGITM**: Apr, 2021
- **ESRI UC**: Jul, 2021
- **GIS Pro**: Oct, 2021
Sources


Questions?

Thank You

Neil S. Rose, GISP
nsr8@psu.edu