Data Quality – Impact on Pipeline Integrity Management

GeoFields, Inc.

October 16, 2001
Rationale (and Disclaimer)

- Too many projects go south due to data issues
  - “Silver bullet solutions” won’t make data issues go away
- Presentation based on experience – not academic
- Some interaction with Pipeline industry members – not exhaustive
- All points open to criticism
Components of Pipeline Integrity Management

- Corporate Commitment
- Personnel
- Methods & Models
- Data
<table>
<thead>
<tr>
<th>Category</th>
<th>Context</th>
<th>Integrity Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate</td>
<td>Commitment</td>
<td>Token - Best Practice</td>
</tr>
<tr>
<td>Personnel</td>
<td>Capability</td>
<td>Inexperienced - Competent</td>
</tr>
<tr>
<td>Methods &amp; Models</td>
<td>Confidence</td>
<td>“Black Box” - Understood</td>
</tr>
<tr>
<td>Data</td>
<td>Quality</td>
<td>Unknown - Excellent</td>
</tr>
<tr>
<td>Pipeline</td>
<td>Operational</td>
<td>Liability - Asset</td>
</tr>
</tbody>
</table>
Data Quality

- Quality – “A degree or grade of excellence”
- Directly impacts the “value” of derived information
- Data quality is impacted by:
  - Data Accuracy
  - Data Completeness / Timeliness
  - Data Organization / Usability
  - Data Error
Data Accuracy

• **Accuracy** refers to data that are the result of measurement
  – Instrument limitations, limitation of recording media, limitation of recording method

• **Pipeline data set accuracy** will generally be affected by:
  – Spatial (geographic) accuracy
    • GPS Survey, Basemap (USGS, Aerial photography, satellite imagery, etc.)
  – Linear (stationing) accuracy
    • Linear surveys (CIS, DOC, etc.), ILI, Centerline stationing
  – Real Time measurement (SCADA)

• **Data accuracy** can be measured and controlled
  – Control Points and Instrument Calibration
Data Completeness / Timeliness

• Data set may not exhibit error however
  – If the data set is not complete, it does not accurately represent the entire state of the facility
  – If the data set is not timely, it does not accurately represent the current state of the facility

• Decisions based on less than appropriate completeness & timeliness of data may result in less than judicious action
  – “Appropriate” provides the latitude for phased and periodic data collection
Pipeline integrity management & risk modeling utilizes significant volumes of data

- Number of data set types, Number of pipelines
- Raw data sets, derived data sets, etc.

If data is not systematically organized and managed the result is:

**CHAOS !!**

Pipeline Integrity Management based on chaotic data structures is not defensible
Data Error

• Includes primarily “bad” data
  – Error that is unrecoverable but recognizable
  – Error that is undetected (valid – but wrong – values)

• Includes aspects of data accuracy
Types of Data Error

• Legacy Data Error
  – Data from source such as Alignment Sheets that is in error

• Transcription Error
  – Typing mistakes

• Context Error
  – “Meta data” for data acquisition or conversion recorded incorrectly or not recorded

• Specification Error
  – Data specification for acquisition or conversion not followed, or non existent
Types of Data Error

• Quantified Error
  – Data source specifies accuracy (USGS Quad - +/- 40 foot, etc.)

• Unquantified Error
  – Data source is known to have error, but error is unquantified

• Centerline Station Control Error
  – Pipe centerline has widely dispersed or poorly defined station control features (road crossings, section line crossings, etc.)
  – Station control features may have quantified error
Types of Data Error

• Linear “Alignment” Error
  – Data set has widely dispersed or poorly defined linear control features
  – Distance between control features on data set is significantly different from distance on centerline – how should difference be distributed?

• Derived Data Error
  – Data sets that are the result with more than one data set as input
    • HCA Impact Segment is result of intersection of CL with HCA Area (possibly buffered)
    • CL position – Quantified Error; HCA Area – Unquantified error
  – Error of all input data sets must be considered when determining error of derived data sets
Simple Risk Model – Weighted Summation

Coating
CP
Soils
Population

Risk Ranking

0.15
0.30
0.25
0.40
Handling of Data Error

Weighted Summation

• Weighted summation is a weighted linear overlay
  – Each data set is a set of points or linear segments representing a characteristic over a section of the centerline
  – Point & linear characteristics form a weighted “stack”
  – Resulting segments from “merging the stack” carry all input characteristics
  – Resulting segment length is the “lowest common denominator” of all input segments (including points)
Handling of Data Error
Weighted Summation

• Error is integrated by “extending” data set elements by the known inaccuracy
  – Example
    • 100 foot segment with linear accuracy of +/- 40 feet
    • 40 feet added to each end of the segment
    • Becomes a 180 foot linear segment
  – Example
    • Point “event” with linear accuracy of +/- 40 feet
    • 40 feet added to each side of point
    • Becomes an 80 foot linear segment

• Linear overlay is performed using “error extended” data elements
Data Classification

• Thresholding of Data
  – “Continuous” data - CP Potential readings, Pit depth, etc. must be set to discreet values (1-10, Good, Medium, Poor) in order to “feed” the Risk Model

• Qualitative to Quantitative Transform
  – Data such as “coating type” must be transformed from qualitative information (asphalt, FBE, PE) to a numeric value

• Matrix Classification
  – Two or more variables may be incorporated into the classification process
Managing Data Quality

• Rule #1 – Your data is likely in worse shape than you thought it was
• Understand integrity management objectives
  – Regulatory satisfaction or “Dig here”
• Understand integrity data requirements thoroughly
  – Fundamental data sets, optional data sets (fine tuning)
• Each data set should be documented
  – Data Source
  – Accuracy / Control
  – Error Issues
  – Processing Methods
  – Data structure requirements
• Remember – computational result reflects the worst error from input data sets