DIRTY DATA, DONE DIET CHEAP

Mark McCart
GIS-T 2018 | Little Rock, AR
JUST THE FACTS PLEASE

- What
- Where
- When
- Why
- How

The Dirty Data examples you are about to see and hear in this presentation are real.

Only the names have been changed to protect the innocent.
How often is data gathered only to be used once and never touched again?

How often has data been delivered per RFP requirements only to find out it didn’t meet your original objectives?

How often is the same data collected by different offices/groups in the same organization due to lack of communication?

How often do you see a cool demonstration of an application that would solve an issue in your organization only to learn your data isn’t as ‘ready’ as you thought?

How often does an organization go out and collect data but forget to budget time and resources for QA/QC because you need that data now?

How often has data been collected with resource-grade GPS equipment only to find that the geometry doesn’t match what was on the ground?
MOST COMMON REASONS FOR DIRTY DATA

- Lack of forward-thinking about data
- No organizational data governance
- Poorly written RFPs or Data Collection Plans
- No QA/QC plan in place and/or lack of resources to execute plan
- No one ‘owns’ the data / lack of data champion
- Not treating data as ‘enterprise’ data
- Letting vendors dictate how they want to deliver your data to you
- Lack of training
- Poor choice of technology
- Lack of communications and/or inter-office politics
Multiple data sets

- Multiple ‘Authoritative’ data sets
- Multiple deliverables in several iterations
- ‘Silos of Excellence’
- Not true ‘As-Builts’
HOW TO AVOID MULTIPLE Datasets

- Have a long-term vision for your data
- During RFP development, include Subject Matter Experts (SMEs) for multiple disciplines
- Clearly define a single deliverable
- A true ‘As-Built’ – only show what was built
DELIVERABLE FORMAT

- Contract/RFP states multiple versions of software as ‘acceptable’
- ‘GIS’ data is just ‘Business data with a spatial component’
- Data on incorrect table/layer/feature class
HOW TO AVOID FORMAT ISSUES

- Have a long-term plan/vision for the data
- During RFP development, include Subject Matter Experts (SMEs) for multiple disciplines
- Clearly defined RFP/Data Collection Plan
- Take a ‘Service Centric – Software Neutral’ approach to data
- Put thought into how you want to leverage data, don’t just pick the simplest format
- Take into consideration integration with other platforms: CAD, BIM, Spatial, Dashboarding, Analytics, etc.
ATTRIBUTE DATA

- Domains/Subtypes/Look-up tables not utilized
- No unique ID
- Date/Time format issues
- Missing/Null/Empty fields
- Misspellings
- Incorrect field types
HOW TO AVOID ATTRIBUTE ISSUES

- Have a long-term vision for your data
- During RFP development, include Subject Matter Experts (SMEs) for multiple disciplines
- Require domain and subtypes
- Have QA/QC plan in place and execute it
- To avoid upper/lower case issues, use all caps for data
- Use UTC Date/Time formats
GEOMETRY

- No topology rules set or followed
- Incorrect spatial projection
- Incorrect shape or location
- Too many vertices
- Captured at multiple scales
- Captured with incorrect grade of GPS equipment
HOW TO AVOID GEOMETRY ISSUES

- During RFP development, include Subject Matter Experts (SMEs) for multiple disciplines
- Use a real Datum and Projection appropriate to the scale of your project
- Think ahead on how someone else can leverage this geometry and make that your basis of how data will be collected: Survey vs. Resource vs. Digitizing
- Set up and enforce topology rules
- Have QA/QC plan in place and execute it

**Precision By Scale**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1,200</td>
<td>± 3.33 ft</td>
</tr>
<tr>
<td>1:2,400</td>
<td>± 6.67 ft</td>
</tr>
<tr>
<td>1:4,800</td>
<td>± 13.33 ft</td>
</tr>
<tr>
<td>1:10,000</td>
<td>± 27.78 ft</td>
</tr>
<tr>
<td>1:12,000</td>
<td>± 33.33 ft</td>
</tr>
<tr>
<td>1:24,000</td>
<td>± 40.00 ft</td>
</tr>
<tr>
<td>1:63,360</td>
<td>± 105.60 ft</td>
</tr>
<tr>
<td>1:100,000</td>
<td>± 166.67 ft</td>
</tr>
</tbody>
</table>
How This Is Possible?

In 1882, a land surveyor by the name of Josiah A. King, and his three-man crew, traveled 40 miles from the nearest white settlement called “the Grand Rapids of the Mississippi.”

As the November winds blew around the crew, they surveyed a six square mile area between Moose and Coddington Lakes. Perhaps it was the chilling weather, or all of the desolate swamps around them, but the crew became confused, and they ended up plotting Coddington Lake about a half mile further northwest than it was actually located. Josiah’s crew’s error is Minnesota’s great fortune.

-Chip Jones, Minnesota Fun Facts
FINAL THOUGHTS

- Don’t just collect data to have it. Have a long-term vision on how and why you want the data
- Always include the SMEs from multiple disciplines in your organization when developing RFPs/Data Collection Plans
- Think ‘Enterprise’ vs ‘Silos of Excellence’
- Plan for and budget appropriately for QA/QC Process and enforce it
- Choose appropriate data capture method (digitize vs. resource-grade vs. survey-grade)
- Organizational data governance plans are essential for success – but you have to enforce them
QUESTIONS?

mark.mccart@iowadot.us
@MarkMcCcartGIS