MSDI Data Quality Matrix

Troy Blandford & Michael Fashoway

Business Plan to Improve Geospatial Data Value

- MSL completed a *GIS Coordination Strategic Plan* in 2022. The strategic plan included a goal for MSL to, “Continue to improve the collection, maintenance, and dissemination of authoritative geospatial information."

- The value of statewide geospatial data and MSDI hinges on maintaining and improving data quality and ensuring use.
  - How can MSDI Theme Leads increase the value or quality of their data?
    - Better policies/standards, more informed decisions, consistency, reaching more Montanans.

Data Quality Dimensions (measures)

- Precision & Accuracy
- Consistency
- Completeness
- Integrity
- Timeliness
- Relevance
- *Others to consider? (Themes may differ)*

This exercise is about **assessing a dataset(s) suitability for use**.

Business Plan to Improve Geospatial Data Value

Tasks for each Theme

1. Convene a working group (may happen at beginning of process or may happen later)
2. Confirm generic data quality dimensions, add or refine as appropriate for your Theme
3. Assess your Theme's dataset(s) using the defined list of quality dimensions and **populate the data quality matrix**
4. Develop a **data value improvement plan** for the data theme based on the data quality matrix
   - Identify actions and outcomes
5. Use the data value improvement plan to guide development of the MSDI Theme Plan (improving data value may be a section of the Plan or the overall guide throughout).
6. Incorporate improvement actions into appropriate staff responsibilities and work plans
7. Track progress and achievements

Data Quality Dimension	Description	Example
Precision & Accuracy	Accuracy is whether data values correctly represent the phenomenon observed (subject to usage needs and measurement technologies). Precision is how the value is stored and expressed. Is the accuracy documented? Does it meet end user's needs?	• Each address represented by a point is within 2 meters of the structure it represents (accuracy) and coordinates are stored and expressed as integer meter x-y pairs (precision). • 1-meter resolution elevation with 4" vertical accuracy in nonvegetated areas, as measured with 95% confidence interval. Repeatability (precision) is 2.2 inches.

Data Quality Dimension	Description	Example
Consistency	Similarities and differences between data stored in multiple datasets or databases	• A statewide address database having the same set of addresses as each county's address databases is consistent • Lidar point cloud classification is inconsistent (e.g, sometimes vegetation is included in the classification; most of the time it is not). Some DEMs are hydro-enforced; most are not.

Data Quality Dimension	Description	Example
Completeness*	A measure of whether the data are comprehensive with all phenomena represented and all required observations populated. Are required or priority fields populated?	• An address database that does not contain all addresses is incomplete • A 1-m DEM that covers the entirety of the state (all pixels have a value) is complete; a DEM dataset that covers an entire state but has some gaps or lacks values in some cells or doesn't meet all standards/products everywhere is incomplete;

*Easy to assess for some datasets, difficult for others. Example easy: Elevation (DEM) or imagery is complete when all pixels over the area of interest have a value. Example difficult: New addresses are added frequently. There is not a perfectly "known" population of data.

Data Quality Dimension	Description	Example
Timeliness	The update cycles of a dataset or set of datasets in a theme, supporting internal and external production processes, reporting and usage needs. This dimension includes the ability to use historical versions of a dataset (archiving) and the frequency with which updates are performed, and older data are archived. Also consider temporality to internal and external deadlines and reporting needs.	• Counties provide new and edited address datasets on a irregular schedule and statewide data are updated monthly (based on whatever counties have been updated). • Elevation (lidar) has a long turnaround from start of acquisition to data availability. Data collected in 2024 may not be delivered until 2026. What is the impact, can this turnaround be shortened, and what is ideal?

Data Quality Dimension	Description	Example
Integrity	An assessment of whether the data's structure, schema, and maintenance workflows maintain consistency, completeness, accuracy, precision, to meet user needs and interoperability	• Address data are verified and processed to a common standard so that all addresses have equal validity. • Lidar data are transformed using standardized techniques with consistent documentation of issues for users to consider.

Data Quality Dimension	Description	Example
Relevance	Whether data are available, accessible, and having the completeness, consistency, precision/accuracy, integrity, and timeliness for a given business use. Is the data available to inform business systems and answer business questions?	• Address data support NG9-1-1 systems; • Lidar/DEM data are used for modeling flood risk and vegetation; therefore, past DEM datasets need to be available to study flood or land cover changes over time (change detection). Lidar is being used for vegetation structure mapping (e.g., tree density, canopy), but most lidar is currently primarily classified for bare-earth (vegetation is not part of the classification)…meaning the lidar is not in the ideal condition for some partners.

Steps to completing the Matrix

- Assess the Theme's dataset(s) using the defined list of quality dimensions (modify/add as necessary for your theme).
  - Use the Dataset Assessment Workbook Template (Appendix D).
  - Identify a facilitator/advisor to help assess the dataset(s). This may be You, MSL staff that have gone through the process, or an outside consultant.
  - Assign parts of the assessment to working group participants or work together as a group to complete the assessment. Or, make a first attempt yourself, then seek working group review.
  - Consider and document the datasets using:
    1. **Current status**
    2. **Near-term quality improvement**
    3. **Long-term quality improvement**

The Data Quality Matrix

(Example for Structures & Addresses and Elevation)

After the data quality assessment, **determine the actions** that must be taken to address deficiencies and make improvements

Identify the technical means to make it happen:

Quality Control

Performing consistent (i.e., repeatable) data profiling analyses to determine missing values, outliers, inconsistencies, and illegal values in constrained or domain-based value sets.

Quality Assurance

Implementing proactive data validation in entry and editing workflows and software.

Documentation

Using a data management plan for each dataset that specifies who, when, and how datasets are updated and incorporated into a data theme in easily understood formats like workflow charts and checklists.

Standards

Establishing and enforcing standards for a dataset, including constraints, domains, naming conventions, and data formats.

Training

Promulgating standards and educating data creators and maintainers to build a culture of data quality awareness.

And many more…

Data Quality Matrix > Actions > **Data Improvement Plan**

Develop a data value improvement plan for each Theme.

- Use the Template Data Value Improvement Plan (Appendix E).
- Define overall goals for selected dataset/theme based on the assessment.
- Based on the assessment and deficiencies identified, define the outcomes and actions needed to achieve the goal.
- Highlight repeatable improvement actions (e.g., data validation, cleansing, quality monitoring, quality checks, standards, etc.).
- Identify resource needs for each plan outcome and action (staff time, licensing, consultant support, funding . . .)

Data Value Improvement Plan

Example Structures & Addresses

Example Elevation

The Data Value Improvement Plan informs the MSDI Theme Plan.

Questions

Thank you!

mfashoway@mt.gov

tblandford@mt.gov