Understanding Data Quality

IST400/600

JIAN QIN

Why are we concerned about data quality?
Why are we concerned about data quality?

- Protect potential data consumers from unintended consequences resulting from
  - Misinformation
  - Mistaken assumptions about data collection methods, measurement precision, or scale

What is data quality?

<table>
<thead>
<tr>
<th>Quality Dimensions</th>
<th>Strategic Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Relevance</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Quality</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Timeliness</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Utility</td>
</tr>
<tr>
<td>Interpretability</td>
<td>Completeness</td>
</tr>
<tr>
<td>Coherence</td>
<td>Comparability</td>
</tr>
</tbody>
</table>

Data quality glossary

- **Accuracy**: the degree of agreement between an observed value and an accepted reference value. Accuracy includes a combination of random error (precision) and systematic error (bias) components which are due to sampling and analytical operations; a data quality indicator.

- **Assessment**: the evaluation process used to measure the performance or effectiveness of a system and its elements, used to denote any of the following: audit, performance evaluation, management systems review, peer review, inspection, or surveillance.

- **Comparability**: the degree to which different methods, data sets and/or decisions agree or can be represented as similar; a data quality indicator.


The two sides of quality

<table>
<thead>
<tr>
<th>Target data system</th>
<th>Metadata for users</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Background</td>
<td>• Lineage</td>
</tr>
<tr>
<td>• Frames and Sampling (if</td>
<td>• Positional accuracy</td>
</tr>
<tr>
<td>applicable)</td>
<td>• Attribute accuracy</td>
</tr>
<tr>
<td>• Data Collection</td>
<td>• Logical consistency</td>
</tr>
<tr>
<td>• Data Preparation</td>
<td>• Completeness</td>
</tr>
<tr>
<td>• Data Dissemination</td>
<td>• Currency</td>
</tr>
<tr>
<td>• Sponsor Self-Evaluation</td>
<td>• Status</td>
</tr>
<tr>
<td>• Data Analysis Results</td>
<td></td>
</tr>
</tbody>
</table>

IST400/600 Scientific Data Management
Target data systems: An example (1)

- AmeriFlux network provides continuous observations of ecosystem level exchanges of CO₂, water, energy and momentum spanning diurnal, synoptic, seasonal, and interannual time scales and is currently composed of sites from North America, Central America, and South America.


Target data systems: An example (2)

- 149 Sites across the Americas
- Each site reports a minimum of 22 common measurements.
- Communal science – each principle investigator acts independently to prepare and publish data.
- Data published to and archived at Oak Ridge.
- Total data reported to date on the order of 150M half-hourly measurements.

http://public.ornl.gov/ameriflux/
Target data systems: An example (3)

- **Data quality problems:**
  - Measurements Are Not Simple or Complete
  - Gaps in the data
    - Quiet nights
    - Bird poop
    - High winds
    - ...
  - Difficult to make measurements
    - Leaf area index
    - Wood respiration
    - Soil respiration
    - ...
  - Localized measurements – tower footprint
  - Local investigator knowledge important
  - PIs’ science goals are not uniform across the towers

Checking for data quality

- Real field data has both short term gaps and longer term outages
  - The utility of the data depends on the nature of the science being performed
  - Browsing data counts can give rapid insight into how the data can be used before more complex analyses are performed

(Agarwal and van Ingen, 2006).
Checking for data quality

- Real field data has unit and time scale conversion problems
  - Sometimes easy to spot in isolation
  - Sometimes easier to spot when comparing to other data
  - Browsing data values can give rapid insight into how the data can be used before more complex analyses are performed

(Agarwal and van Ingen, 2006).

Metadata for users (1)

- Describing data quality in metadata (FGDC metadata standard, see example record in handout)
  - Lineage
    - Methodology: Information about a single step of field and/or laboratory work.
    - Source of Data: List of sources and a short discussion of the information contributed by each.
    - Processing steps: Information about a single data processing event. Can describe process applied to an acquired data set or to raw data collected.
Data quality benefits of lineage

- Communicates suitability, reliability, accuracy, currency, redundancy
- Enhance interpretation, prevents misinterpretation, misuse of environmental data
- Enhance a user’s justification for using data
- Reduces possible false sense of data precision
- Facilitate integration of data
- Allows non-expert data user to understand processing steps
- Communicates processing steps leading to creation of scientific data product
- ...
**Lineage (3)**

- **Scientific processing benefits of lineage**
  - Records processing history for internal records, audit, quality control
  - Records computational history for judging statistical validity of future operations
  - Reduces data provider liability
  - Provides consistent documentation for distributed datasets
  - Finds the sources of faulty, anomalous processing inputs
  - Saves processing “recipes”; modifies and reruns processing sequence
  - ...

**Lineage (4)**

Metadata for users (2)

- Positional accuracy
  - Determining accuracy requires comparison of a recorded position against the actual position as defined by a known datum. Positional accuracy would be determined by how close the represented position of a feature is in relationship to its actual position on the earth.

  **Positional Accuracy**:
  **Horizontal Positional Accuracy**:
  **Horizontal Positional Accuracy Report**: As for PRISM maps, accuracy of this data set is based on the original specification of the Defense Mapping Agency (DMA) 1 degree digital elevation models (DEM). The stated accuracy of the original DEMs are 130 m circular error with 90% probability.

  **Quantitative Horizontal Positional Accuracy Assessment**:
  **Horizontal Positional Accuracy Value**: 130 m with 90%
  **Horizontal Positional Accuracy Explanation**: The broad DMA production objective for 1-degree DEM's.
Metadata for users (3)

- Attribute accuracy
  - Determining accuracy requires comparison of recorded entry against the actual as defined by predefined standards.

Metavist interface for data quality

Explanation of the accuracy of the identification of entities and assignment of attribute values in the data set.
The report can reference more extensive descriptions in other documents.
Metadata for users (4)

- Logical consistency
  - How well does the data fit within logical rules of data structure.
  - Attribute logical consistency entails the testing of two or more functionally related attributes. The value for one attribute determines the valid values for its related attributes. (If X then Y)
  - Feature logical consistency is the testing for feature to feature relationships that are consistent with known or expected rules.

Metadata for users (5)

- Completeness
  - Completeness of spatial coverage tests expected spatial coverage against actual coverage either as areas missed or stations covered
  - Completeness of temporal coverage is for time series data when there are gaps in the time recordings
  - Completeness of classification examines how exhaustive is the classification system and are there generalizations
  - Completeness of verification examines the verification method for the data
  - Completeness of attribution examines if each record is complete
Metadata for users (6)

- Other metadata elements for data quality
  - Currency
    - Beginning date
    - End date (processing time between collection and storage if ongoing collection)
  - Status
    - Maintenance and update frequency
    - Progress
Assessment of data quality

Scientific data quality assessment

- Records information about data quality
- Metadata
  - Information about metadata creator and maintainer
  - Updates metadata records
  - Corrects errors and update metadata records

Quality dimensions

<table>
<thead>
<tr>
<th>For scientific data</th>
<th>For metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Completeness</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Provenance</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Conformance to expectations</td>
</tr>
<tr>
<td>Interpretability</td>
<td>Logical consistency and coherence</td>
</tr>
<tr>
<td>Logical consistency</td>
<td>Timeliness</td>
</tr>
<tr>
<td>Coherence</td>
<td>Accessibility</td>
</tr>
<tr>
<td></td>
<td>Proper context</td>
</tr>
<tr>
<td></td>
<td>Use of standard vocabularies</td>
</tr>
</tbody>
</table>

IST400/600 Scientific Data Management