

# Understanding Data Quality

IST400/600

JIAN QIN



## Why are we concerned about data quality?

2



IST400/600 Scientific Data Management

## Why are we concerned about data quality?

3

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



IST400/600 Scientific Data Management

## What is data quality?

4

### Quality Dimensions

- **Relevance**
- **Accuracy**
- **Timeliness**
- **Accessibility**
- **Interpretability**
- **Coherence**

### Strategic Goals

- **Relevance**
- **Quality**
- **Timeliness**
- **Utility**
- **Completeness**
- **Comparability**

Burns, E.M. et al. (2002). Data quality assessment methodology: A framework. In: Proceedings of the Survey Research Methods Section, ASA (2002).

<http://www.amstat.org/Sections/Srms/Proceedings/y2002/Files/JSM2002-000347.pdf>

IST400/600 Scientific Data Management

## Data quality glossary

5

- **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.

For more terminology on data quality, see:

[http://www.hanford.gov/dqo/glossaries/Glossary\\_of\\_Quality\\_Assurance\\_Terms1.pdf](http://www.hanford.gov/dqo/glossaries/Glossary_of_Quality_Assurance_Terms1.pdf)

IST400/600 Scientific Data Management

## The two sides of quality

6

### Target data system

- Background
- Frames and Sampling (if applicable)
- Data Collection
- Data Preparation
- Data Dissemination
- Sponsor Self-Evaluation
- Data Analysis Results

### Metadata for users

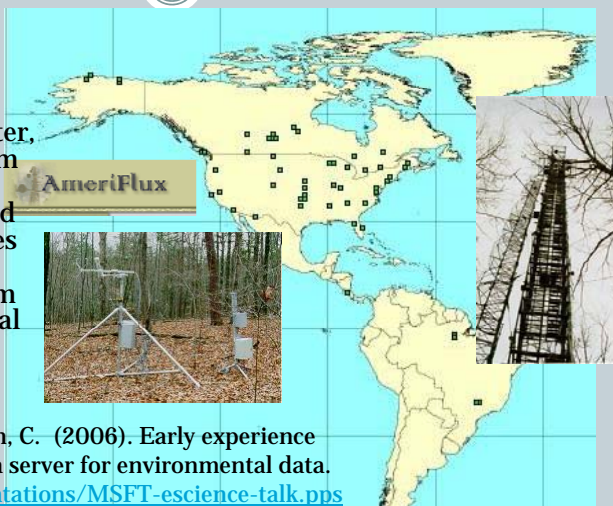
- Lineage
- Positional accuracy
- Attribute accuracy
- Logical consistency
- Completeness
- Currency
- Status

IST400/600 Scientific Data Management

## Target data systems: An example (1)

7

- AmeriFlux network provides continuous observations of ecosystem level exchanges of CO<sub>2</sub>, 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.



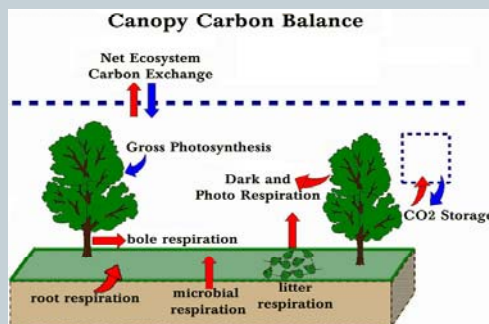
Agarwal, D. and van Ingen, C. (2006). Early experience prototyping a science data server for environmental data.  
<http://bwc.lbl.gov/Presentations/MSFT-escience-talk.pps>

IST400/600 Scientific Data Management

## Target data systems: An example (2)

8

- 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/>

IST400/600 Scientific Data Management

# Target data systems: An example (3)

9

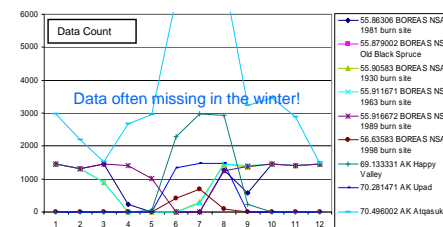
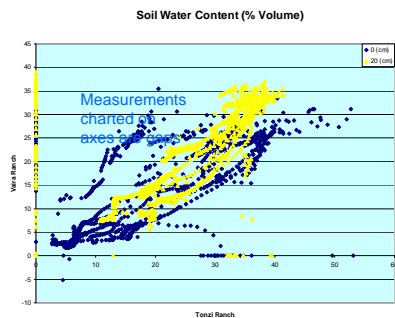
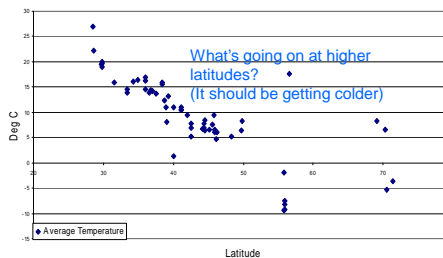
- **Data quality problems:**
  - Measurements Are Not Simple or Complete
  - Gaps in the data
    - Quiet nights
    - Bird poop
    - High winds
    - ....
  - Difficult to make measurements
    - Leaf area index
    - Wood respiration
    - Soil respiration
    - ...
  - Localized measurements – tower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



IST400/600 Scientific Data Management

## 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



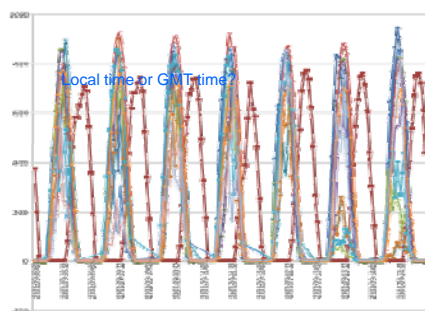
(Agarwal and van Ingen, 2006).

IST400/600 Scientific Data Management

10

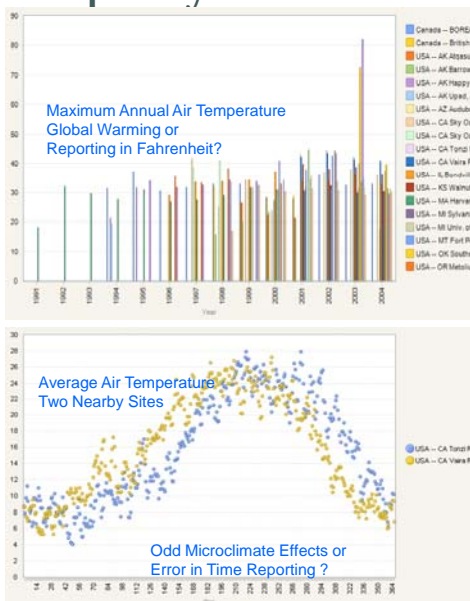
## Checking for data quality

- 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).

IST400/600 Scientific Data Management



11

## Metadata for users (1)

12

- 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.

IST400/600 Scientific Data Management

## Lineage (1)

13

Identification | Data Quality | Spatial Data Org | Spatial Reference | Entity & Attribute | Distribut

Attribute Accuracy | Consistency & Completeness | Positional Accuracy | Lineage | Cloud Cover

**Methodology**

Information about a single step of field and/or laboratory work.

Add Edit Delete

**Source Information**

National Atlas: Attribute and geospatial data

List of sources and a short discussion of the information contributed by each.

Add Edit Delete

**Process Step**

The data were converted to an ARC/INFO coverage using the '...'  
The following steps were performed by ESRI: Downloaded the A...

Information about a single data processing event.  
Can describe process applied to an acquired data set or to raw data you collected.

Add Edit Delete

IST400/600 Scientific Data Management

## Lineage (2)

14

- **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
  - ...

IST400/600 Scientific Data Management

## Lineage (3)

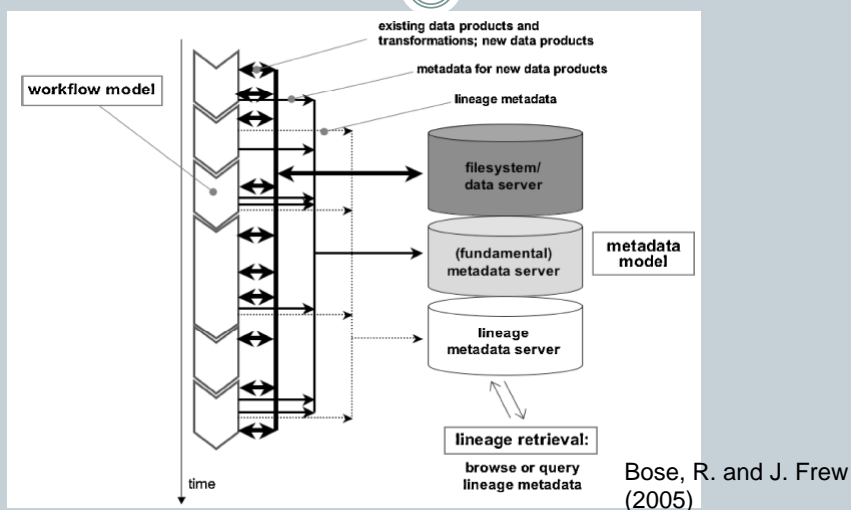
15

- **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
  - ...

IST400/600 Scientific Data Management

## Lineage (4)

16



IST400/600 Scientific Data Management



## Metadata for users (2)

17

### ○ 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.

IST400/600 Scientific Data Management

The screenshot shows a software interface with a top navigation bar containing tabs for 'Attribute Accuracy', 'Consistency & Completeness', 'Positional Accuracy', 'Lineage', and 'Cloud Cover'. The 'Positional Accuracy' tab is active.

Under the 'Positional Accuracy' tab, there are two main sections:

- Horizontal Positional Accuracy:** This section has a title circled in red. Below it is a text area containing a report: "The geospatial part of the data set was calculated from latitude/longitude coordinates. The positional accuracy is unknown. Note that some locations are the center point of broad volcanic fields, and that even at individual volcanoes the coordinates given do not necessarily match the eruption site. Tens of kilometers may separate eruptive centers of a single volcano, particularly in large caldera complexes and rift settings." To the right of the text area is a field for "Quantitative Horizontal Positional Accuracy Assessment (optional)" and three buttons: "Add", "Edit", and "Delete". Below the text area is a smaller text box: "Explanation of the accuracy of the horizontal coordinate measurements and a description of the tests used."
- Vertical Positional Accuracy:** This section has a title circled in red. It contains a large empty text area for a report, a field for "Quantitative Vertical Positional Accuracy Assessment (optional)", and "Add", "Edit", and "Delete" buttons. Below the text area is a smaller text box: "Explanation of the accuracy of the vertical coordinate measurements and a description of the tests used."

## Metadata for users (3)

19

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

IST400/600 Scientific Data Management

## Metavist interface for data quality

20

[Identification](#) | [Data Quality](#) | [Spatial Data Org](#) | [Spatial Reference](#) | [Entity & Attribute](#) | [Distribution](#) | [Metadata](#)  
[Attribute Accuracy](#) | [Consistency & Completeness](#) | [Positional Accuracy](#) | [Lineage](#) | [Cloud Cover](#)

**Attribute Accuracy Report**

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.

*Quantitative Attribute Accuracy Assessment (optional)*





IST400/600 Scientific Data Management

## Metadata for users (4)

21

- 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.

IST400/600 Scientific Data Management

## Metadata for users (5)

22

- **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

IST400/600 Scientific Data Management

## Metavist interface for data quality

**Logical Consistency Report**

No duplicate features are present. The shapefile is exported using Avenue request ExportClean. This request verifies and enforces the correctness of shapes.

Explanation of the fidelity of relationships in the data set and tests used.

The report can reference more extensive descriptions in other documents.

If creating a Logical Consistency Report is not logical, use the default "not applicable"

**Completeness Report**

After processing, the data set is checked for drawing display and number of records and file sizes compared with source materials.

Information about omissions, selection criteria, generalizations, definitions used, and other rules used to derive the data set.

The report can reference more extensive descriptions in other documents.

## Metadata for users (6)

24

- **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

