Introduction to Research Data Management

Quality control

http://hdl.handle.net/1969.1/164594

LIBRARIES
TEXAS A&M UNIVERSITY

Workshops

1. Build an overview
2. Collect and document data
3. Store digital data
4. **Work with data**
5. Share and preserve data
6. Plan ahead

Introduction

Focus on actions that prevent errors from entering or remaining in a dataset.

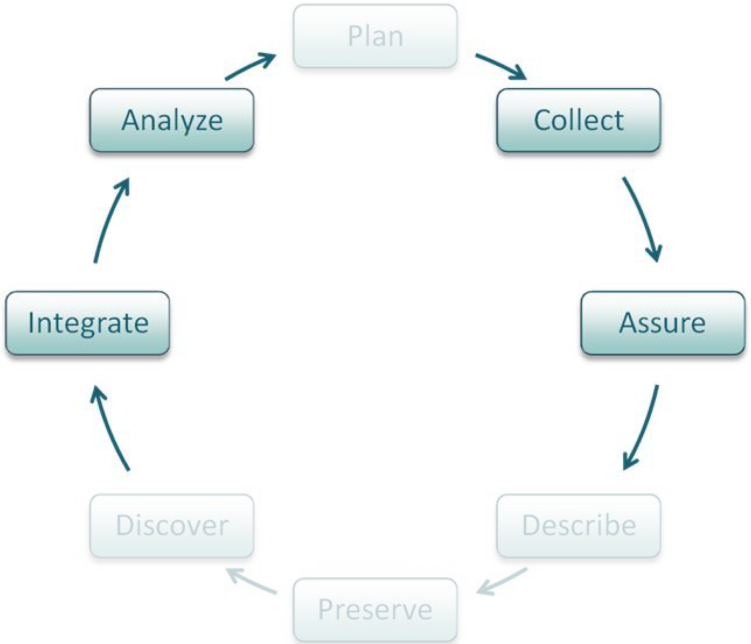
The goal is to ensure the quality of the data you plan to analyze, and to identify potential problems with data that could affect use.

Discussion

What are some actions you take or expect to take to maintain the quality of your data before, during, and after collection?

QA / QC

Quality assurance and quality control describe a range of activities that aim to prevent errors from entering or staying in a data set.



Reducing errors

Errors in data are common.

- Omission: data or metadata not recorded.
- Commission: incorrect or inaccurate data entered.

Collecting and entering data

- Assign responsibility.
- Use validation.
- Double check for mistakes.

Combining datasets

Get to know the data:

1. Is the dataset from a reliable source?
2. How and for what purpose were the data collected?
3. How have the data been manipulated?
4. Have others drawn conclusions from the data?

After data collection

- Check data formatting.
- Check for valid data.
- Check for anomalies and outliers.

Check formatting

- Tabular data include headers, rows for observations, and columns for variables.
- Consistent data type (integer, character, date-time) for each variable.

Check values

- Ensure standard dates and codes.
- Find and identify missing values with codes.
- Flag values that have been estimated or gap-filled.
- Find and examine anomalies and outliers.

missing_text	missing_number
N/A	99
None	-99
Null	999.99
Not Applicable	-999.99

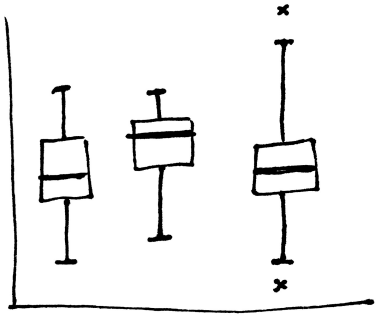
day	avg_temp_c	flag
1	31.2	actual
2	32.3	actual
3	33.4	estimated
4	35.8	actual

Find anomalies and outliers

- Make visual determinations.
- Compare to related observations.
- Use statistical tests.

Visual determinations

- Scatter plots: identify outliers when there is an expected pattern.
- Box plots: identify extreme values in the tails of a distribution.
- Mapping: indicate anomalies in geographic coordinates.
- Others?



Related observations

- Examine reference data.
- Create difference plots for co-located data streams.
- Compare two parameters that should covary.

Statistical tests

- Dixon's test
- Grubbs’ test
- Tietjen-Moore test
- Generalized extreme Studentized deviate (ESD) test
- Others?

Examine outliers

- Investigate how these values entered the dataset.
- Flag values identified as outliers or anomalies.
- Remember, they may be valid values or errors, and are worth keeping.

Documentation reminder

- Document procedures and standards followed.
- Document the checks made to data.
- Mark data with quality control flags.

Conclusion

- Reviewed the purpose of quality control actions.
- Identified common quality control practices.

References and resources

- DataOne. “Develop a quality assurance and quality control plan” [website](https://www.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan)
- DataOne. “Ensure basic quality control” [website](https://www.dataone.org/best-practices/ensure-basic-quality-control)
- NIST/SEMATECH. “e-Handbook of Statistical Methods” [ebook](http://www.itl.nist.gov/div898/handbook/)