Synthesis quality assurance and data management best practices

Draft December 2019

Interagency Ecological Program (IEP) Synthesis Coordination Committee

Synthesis involves integration of data and results from multiple studies to answer questions in new ways. Often, investigators conducting synthesis projects feel that standards for quality assurance, quality control (QAQC) and data management are not relevant to their project because they are not collecting "new data". However, collecting data for a synthesis project can be approached much like a field project. Data is being "collected" from databases or published literature instead of from nets/sondes/microscopes, but analogous protocols, standards, and procedures should be applied. These steps should align with the IEP Data Utilization Workgroup's (DUWG) Data Life Cycle (Figure 1).

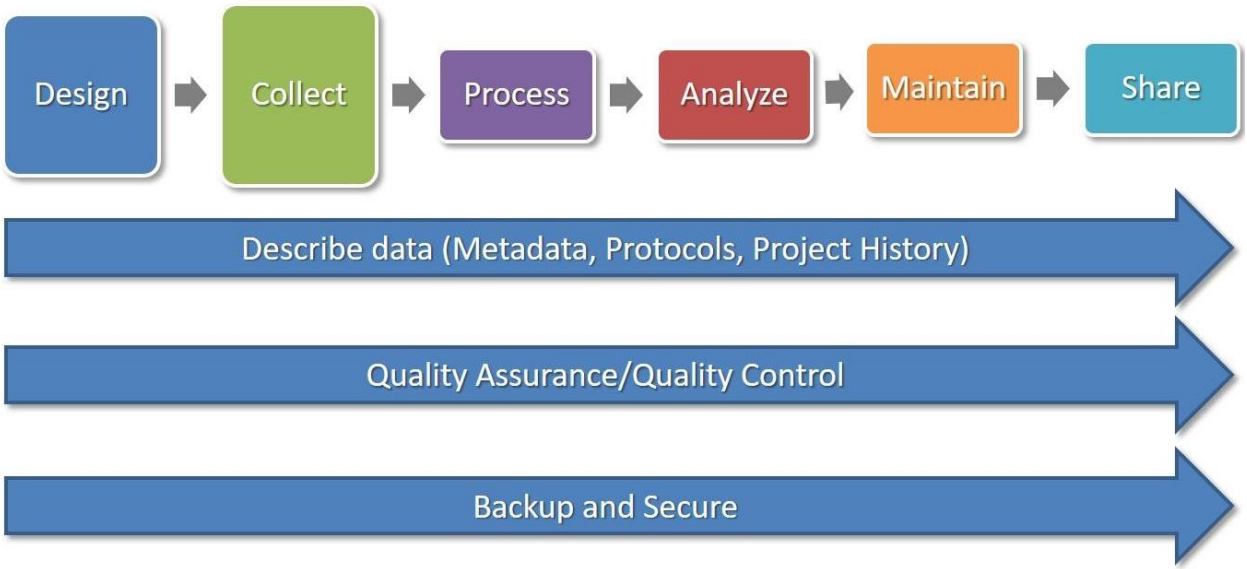

Figure 1 – The IEP Data Utilization Work Group (DUWG) Data Life Cycle model. The major steps (Plan, collect, process, analyze, maintain, and share) may be sequential or iterative, but every step in the process should include metadata, quality assurance, and backups. Adapted from (Faundeen et al. 2013).

1. Step 1 – Plan your project
    a. Decide what criteria you will use to choose datasets or studies for inclusion in the synthesis work. Document these criteria and how they apply to each data set you use. This documentation may take the form of a Quality Assurance Project Plan. Include:
        i. Spatial and temporal scope.
        ii. Data quality indicators (precision, accuracy, completeness, representativeness, comparability, bias)
        iii. Specific methods or sampling procedures you require for the data you'll be obtaining (if applicable).
        iv. What level uncertainty or variability (quantitative, qualitative) in procedures, methods, or measures is acceptable.
        v. Decide if you want to use raw data or QA'd data. If you are assembling data from multiple sources, you may want to apply your own QA procedures instead of dealing with a complex matrix of different QA protocols.
    b. Assess data availability in relation to the goals and objectives of your study. Are enough data available the meet your criteria to adequately address your research questions?
    c. Write a data management plan, including backup, storage, and publishing for any derived datasets and code, along with consideration for access and storage of the original data set.
    d. Collaborate with other IEP groups to ensure greatest use of resources
        i. Bring your plan to the synthesis coordination team to identify data sets and potential for integration with other projects.
        ii. Workshop your plan with applicable Project Work Teams
        iii. Use templates and recommendations made by the DUWG, which are available on the IEP GitHub Open-Data-Workshop site.
2. Step 2 – Collect the data
    a. Document which data sets were chosen, how and when they were accessed, and where their metadata and QA information are available.
    b. Fully read and understand all metadata and quality assurance documentation provided with data. In particular:
        i. What were the data originally collected for?
        ii. Will you have to make any additional quality checks on the data?
    c. Consider the compatibility of the data you are integrating. Document how you deal with *incompatibilities* across datasets.
        i. Understand the methods used to collect the data and the

comparability of methods across data sets.

ii. Understand the QA procedures applied to each dataset and how they vary among data sources or monitoring programs.

iii. Understand each study design and how study design, and differences between study designs, may or may not introduce biases into your analyses.

iv. Consider the relationship between your synthesis questions and the study designs from your source data sets. Are any variables confounded with study identity?

d. Communicate with the investigators who collected the original data set (if possible).

i. Discuss how you want to use the data.

ii. Share your data management plan

iii. Consider any compatibility concerns they may have and see whether there are any caveats to their data that are not describe in their metadata.

iv. Discuss appropriate attribution for their data in any resulting publications.

e. Backup the derived data set in a secure location. For most PIs, this will be on their agency's server, though cloud storage locations are also used frequently.

f. Update your data management plan, as needed

3. Steps 3 – Process the data

a. Document the steps used in data processing. Data processing steps include any changes to the dataset to create the new, derived data set. These steps might include reformatting to allow different datasets to be integrated, changing variable names, converting units, subsetting the data, and other non-statistical changes to the data.

b. Identify and document any additional QA/QC steps

i. Were the data filtered or manipulated in any way? Why?

c. Retain code and code documentation used for data processing. This code should be backed up with the data or in a code repository such as GitHub. The code should then be archived with the derived dataset for long-term storage.

i. Note: IEP projects are encouraged to use the IEP GitHub site

4. Step 4 – Analyze the data

a. Document the data analysis steps. Data analysis includes and statistical tests, data visualizations, or aggregation of the data to answer specific research questions.

b. Take the original study design into account when conducting analysis.

c. Back up the analysis code and results in a secure location (departmental

server, or GitHub).

d. Consider publishing the code along with the results of the analysis.

5. Steps 5 and 6 – Archive and share the data
    a. Obtain permission from investigators who collected the original data (if possible)
    b. Identify and document the original dataset within the derived dataset (dataset provenance).
    c. Include all code or documentation used to process the data.
    d. The DUWG recommends publishing data to the Environmental Data Initiative and obtaining a DOI for your data. https://environmentaldatainitiative.org/
    e. Follow all guidance in the IEP Publication Guidelines document.

This basic structure is designed to provide guidance when filling out a Data Management Plan or a quick check to make sure a PI has taken everything into account before starting a synthesis project. Every synthesis project will have different approaches, and not all steps will be relevant to all projects. For additional information on data management and quality assurance, refer to the USGS Data Management website and the DataOne data management best practices site.

Literature Cited

Faundeen, J. L., T. E. Burley, J. A. Carlino, D. L. Govoni, H. S. Henkel, S. L. Holl, V. B. Hutchison, Elizabeth Martín, E. T. Montgomery, C. C. Ladino, S. Tessler, and L. S. Zolly. 2013. The United States Geological Survey Science Data Lifecycle Model, Open-File Report 2013-1265. U.S. Department of the Interior, U.S. Geological Survey.