Title
The Life Cycle of CENS Data

Authors
Wallis, Jillian C
Mayernik, Matthew S
Pepe, Alberto
et al.

Publication Date
2007-10-10

The Life Cycle of CENS Data

Jillian Wallis, Matthew Mayernik, Alberto Pepe & Christine Borgman
Data Management

Introduction

Understanding the Data Life-Cycle

- The success of eScience research depends not only upon effective collaboration between scientists and technologists but also upon the active involvement of data archivists.
- Archivists rarely receive scientific data until findings are published, by which time important information about their origins, context, and provenance may be lost.
- A better understanding of these processes will enable archivists to participate in earlier stages of the life cycle and to improve curation of these types of scientific data.

Research Findings

- Research reported here addresses the life cycle of data from collaborative ecological research with embedded networked sensing technologies.
- Evidence from our interview study and field research yields a nine-stage life cycle.
- Among the findings are the cumulative effect of decisions made at each stage of the life cycle; the balance of decision-making between scientific and technology research partners; and the loss of certain types of data that may be essential to later interpretation.

Problem Statement: Why is it necessary to understand the life cycle?

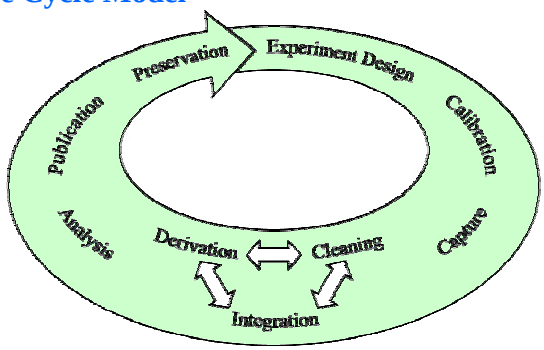
- Archivists typically receive scientific data only after the findings of a study are published or after a researcher retires.
- Neither of these archival outcomes provides access to scientific data in a timely manner.
- More importantly, by the time that archivists receive data, much of the information necessary for future interpretation may have been lost.
- Shifting the practices of archiving such as appraisal, curation, and tracking provenance into earlier stages of a given material's life cycle can increase the likelihood of capturing reliable, valid, and interpretable data and thus improve both short- and long-term access and interpretation.
- To determine how early these archiving processes might begin, it is necessary to identify the life cycle of a given type of data.
- We are not alone in this pursuit -- in keeping with the scientific data archival research agenda for the next decade, our goals are to develop:
  a) More detailed data models for each domain, including intra-domain and inter-domain commonalities
  b) Automatic processes for data and metadata capture
  c) Consistent methods of data description in this scientific and technical environment.

Proposed Model: The data life cycle, from planning to storage.

Life Cycle Stages

1. **Experiment Design**
   The beginning of the data life cycle is the design of new experiments. CENS researchers design new experiments by reusing data from prior research.
2. **Instrument Calibration & Ground-Truthing**
   Before sensors are deployed, they are calibrated to known solutions or values to identify the offset between the actual measurement and the expected measurement.
3. **Data Capture**
   Once sensors have been deployed successfully in the field, researchers begin to collect observations of physical phenomena. Some sensor measurements are direct (e.g., temperature, wind speed) and others are indirect (e.g., measure of fluorescence as an indicator of chlorophyll activity).
4. **Cleaning Data**
   After data have been captured, calibration and ground-truthing information need to be applied to the data to normalize any calibration offsets from the sensing equipment.
5. **Data Integration**
   Few of the observations and samples collected in the field can be interpreted without derivation into more meaningful data points. Data typically must be averaged into composite points before they can be used in analysis.
6. **Derivation**
   Researchers are looking for trends over time and across spatial locations. Datasets each given deployment are integrated by multiple researchers, for multiple reasons, and in multiple combinations.
7. **Analysis**
   Researchers use statistical, modeling, and visualization tools that vary by research specialty and individual preference. They test and generate hypotheses and draw conclusions about data obtained from the deployments.
8. **Publication**
   Data collected during embedded network sensor deployments culminate in scholarly publications such as journal articles, conference papers, posters, and technical reports.
9. **Preservation**
   Few, if any, of the CENS researchers interviewed had data preservation strategies commensurate with those of the archival community. It is more accurate to say that they back up their data.

Life Cycle Model



Significant Findings

- We have identified nine stages that appear to be common to the CENS deployments studied, the researchers, and to the resulting data, as shown in Figure above.
- The order of the steps is not absolute, as some stages are iterative while others may occur in parallel.
- Actions taken at each stage of the life cycle influence how the resulting data can be interpreted, hence it is important that these stages be documented and associated with the resulting dataset.
- There is a cumulative effect of decisions made at each stage of the life cycle. For example, decisions made in the experimental design stage determine what data exist for analysis, or calibration decisions are essential to interpreting the data.
- There is a balance of decision making between scientific and technology research partners.
- The engineering data may be more essential to later interpretation of application data than previously believed.

Future Research

- Our future research will continue to explore and refine the data life cycle identified here, and to build systems to support it.
- At present, much of the sensing technology is experimental, but commercial off-the-shelf sensors also are in use. Research questions about data provenance will evolve as the technology stabilizes and the scientific research questions broaden.