Final Report: Workshop to define quality management standards for data completeness in derived data products

Wade Sheldon (GCE), Don Henshaw (AND) and Ken Ramsey (JRN)

Introduction

Integration of data sets from multiple sources is an important prerequisite for network level and cross-site synthesis studies in LTER, and scaling or re-sampling of primary data is often performed as part of this process. In the EcoTrends project, for example, fine-scale (e.g. hourly) measurements collected by LTER sites are rescaled to monthly or yearly time steps for analysis and display of long-term trends. However, discussions within the LTER IM community and in ASM workshops elucidated problems with both handling and reporting of missing and flagged measurements in source data when derived data sets are produced by and for the EcoTrends project and also by the LTER ClimDB/HydroDB database. We believe it is essential that LTER establish standards for re-sampling and temporally scaling incomplete and flagged measurement data used in synthesis projects, and for documenting this process to provide quality control information for the derived data products. In addition to benefiting LTER synthesis projects, these standards will inform development of data integration tools by LTER sites, LNO, and other projects and organizations such as NCEAS, SEEK, the Canopy Databank and CUAHSI.

In order to address these issues, we conducted an initial 2-day workshop on Jan. 31-Feb. 1, 2007, in Las Cruces New Mexico. This workshop was attended by 13 people with broad expertise in environmental monitoring, data synthesis, and information management. Participants represented 9 LTER sites, the LTER network office, CUAHSI, NCEAS, the Canopy DataBank Project, as well as various standing committees and working groups in LTER (NISAC, IM-Exec, EcoTrends). Additional information about the workshop, major findings and products is presented below.

Participants

Workshop Leaders

Wade Sheldon (GCE, NISAC)
Don Henshaw (AND, NISAC)
Ken Ramsey (JRN, NISAC)

LTER Representatives

Todd Ackerman (NWT, IMExec)
Corinna Gries (CAP, IMExec, SEEK)
Tim Kratz (NTL)
Christine Laney (JRN, EcoTrends)
Margaret O'Brien (SBC, IM-Exec)

John Porter (VCR)
Inigo San Gil (LNO)
Mark Servilla (LNO)

Other Program Representatives
Judy Cushing (Evergreen College/Canopy Databank Project)
Rick Hooper (CUAHSI)
Mark Schildhauer (NCEAS/SEEK Project)

Workshop Presentations

- Sheldon: Survey of Current Practices for Reporting Missing, Qualified Data
- Henshaw: Characterization of quality flag use in LTER ClimDB records
- Sheldon: Synthesis of Incomplete and Qualified Data using the GCE Data Toolbox
- Laney: Trends approach to re-sampling, integrating incomplete, qualified primary data
- Cushing/Ramsey: Grasslands ANPP Data Integration (JRN, SEV, SGS)
- Servilla: Provenance tracking in PASTA framework
- Schildhauer: Provenance in Scientific Workflows on SEEK
- Hooper: CUAHSI Hydrologic Information System – Features of the Observational Data Model

Major Findings

- We identified numerous problems with current practices for encoding data quality and completeness information in metadata and distributed data products:
    - Missing value encodings are frequently used that prevent computer-mediated data parsing or introduce errors in automated workflows (e.g. -999, *, M in numeric columns)
    - Inconsistent, undefined qualifiers are common in data provided online
    - Descriptions of why values are missing are often omitted, which prevents proper statistical decision making (i.e. whether values are missing completely at random, missing at random, or non-ignorable/biased)
- Quality and completeness of primary data is a critical issue for data synthesis, but is not currently handled consistently by major synthesis projects and tools under development (e.g. Eco-Trends, Kepler, PASTA, CUAHSI HIS)
- There is a major need for standardization of:
    - An expressive set of qualifiers for primary data that convey complete information on why values are missing or compromised (i.e. sufficient for use by statistical analysts to check model assumptions, choose data filling approaches)
    - A reduced set of qualifiers for derived data that primary data qualifiers can be mapped to (e.g. questionable, estimated, invalid) to simplify integration

- For integrated data products, summary information on quality/completeness should be provided in the metadata and in derived attributes (e.g. percent missing values columns for each measurement attribute), with links back to the primary data so data users can fully investigate the quality issues if the summary metrics exceed the tolerances for their study or program.

Follow-up Activities

- Two working group sessions were held at the 2007 LTER Information Managers meeting in San Jose, Ca:
    - QA/QC for Real-time Data (Sheldon and Henshaw)
    - QA/QC for Metadata and Data Synthesis (Sheldon and Henshaw)
- An ad hoc QA/QC IMC working group and mailing list was established to follow up on issues raised in these workshops
- Findings from this workshop were discussed in a subsequent Annual Net Primary Productivity Database workshop lead by Cushing to inform synthesis approaches used in this activity
- Enhancement to the EML metadata schema are being proposed to support explicit qualifier attributes and/or cross-attribute referencing to improve the structure of QA/QC information in metadata documents
- A draft controlled vocabulary of recommended qualifiers and definitions created at the workshop is being finalized, and will be vetted by LTER IMs for use across the network
- A review paper on encoding of QA/QC information in archived environmental data sets is planned for 2008/2009 (will be submitted to Ecological Informatics)
- Peter McCartney (NSF OCI) recognizes that this is a ripe area for standards development, and suggested we purse an NCEAS working group proposal or other NSF workshop proposal to engage with a larger group, both of which are under consideration
- NISAC is including several follow-on activities related to standardizing approaches and vocabularies for QA/QC as part of the CI Implementation Plan, which is currently under development

Workshop Products

- Documents
    - Revisions to the EML Best Practices document to include recommendations from this workshop are underway
    - Recommendations from the workshop, along with the report and presentations, are being archived on the LTER IMC community web site for reference
- Publications
    - Sheldon, W.M. 2008. Dynamic, Rule-based Quality Control Framework for Real-Time Sensor Data. Proceedings of the 2008 Ecological Information Management Conference, Albuquerque, NM. (paper accepted)
- Software Products

    - GCE Data Toolbox for MATLAB (http://gce-lter.marsci.uga.edu/public/im/tools/data_toolbox.htm) – workshop recommendations for QA/QC of derived data implemented (e.g. inclusion of attributes summarizing % missing and % flagged in re-sampled data, ability to exclude data exceeding QA/QC tolerances from analyses)
- Proposals based on or including findings from the workshop
    - prRAMID – Python Application for Real-Time Access to Metadata Integrated Datasets. Proposal to the Marine Ecosystem-Based Management Tool Innovation Fund by Adrian Burd and Wade Sheldon. (status: pre-proposal accepted, final proposal declined)