Lowering the Barriers to Integrative Aquatic Ecosystem Science: Semantic Provenance, Open Linked Data, and Workflows

IN51C-1605

T. C. Harmon[1], A. F. Hofmann[2], R. Utz[3], E. Deelman[4], P.C. Hanson[5], P. Szekely[4], S. Villamizar[1], C.A. Knoblock[4], Q. Guo[1], D. Crichton[6], M.P. McCann[2] and Y. Gil[4]

[1]University of California Merced; [2]Monterrey Bay Aquarium Research Institute, [3]NEON Inc., [4]USC Information Science Institute, [5]University of Wisconsin Madison, [6]NASA Jet Propulsion Lab

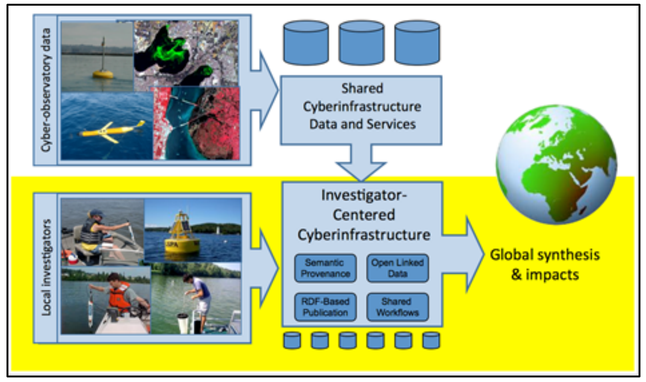
*Contact: Tom Harmon (tharmon@ucmerced.edu) or Yolanda Gil (gil@isi.edu)

Abstract

Environmental cyber-observatory (ECO) planning and implementation has been ongoing for more than a decade now, and several major efforts have recently come online or will soon. Some investigators in the relevant research communities will use ECO data, traditionally by developing their own client-side services to acquire data and then manually create custom tools to integrate and analyze it. However, a significant portion of the aquatic ecosystem science community will need more custom services to manage locally collected data. The latter group represents enormous intellectual capacity when one envisions thousands of ecosystems scientists supplementing ECO baseline data by sharing their own locally intensive observational efforts. This poster summarizes the outcomes of the June 2011 Workshop for Aquatic Ecosystem Sustainability (WAES) which focused on the needs of aquatic ecosystem research on inland waters and oceans. Here we advocate new approaches to support scientists to model, integrate, and analyze data based on: 1) a new breed of software tools in which semantic provenance is automatically created and used by the system, 2) the use of open standards based on RDF and Linked Data Principles to facilitate sharing of data and provenance annotations, 3) the use of workflows to represent explicitly all data preparation, integration, and processing steps in a way that is automatically repeatable. Aquatic ecosystems workflow exemplars are provided and discussed in terms of their potential broaden data sharing, analysis and synthesis thereby increasing the impact of aquatic ecosystem research.

Introduction and Challenges

A key observation by the 2011 Workshop for Aquatic Ecosystem Sustainability (WAES) participants was that integrative and holistic research efforts are needed to address the sustainability of aquatic ecosystems. Workshop participants agreed to several observations about the challenges faced by the community:



- The difficulty of finding datasets can be great, such that many opportunities are lost.
- Many environmental and ecological scientists are unaware of relevant advances in computer science, particularly in rapidly changing areas that are not traditionally connected with them.
- Environmental and ecological sciences would benefit greatly from faster turnaround of sensor to analysis.
- Remote sensing is positioned to have an immediately greater impact in environmental and ecological sciences if its use were better supported by infrastructure.
- The continuity of provenance and other metadata improve the usefulness of the data to individual scientists and enable the reuse of the transformed data by other scientists.
- Facilitating the creation of standardized data formats, metadata properties, and automated analytical tools would be highly beneficial.
- Workflows have been defined in several communities and are used for explicit process sharing with a number of benefits.

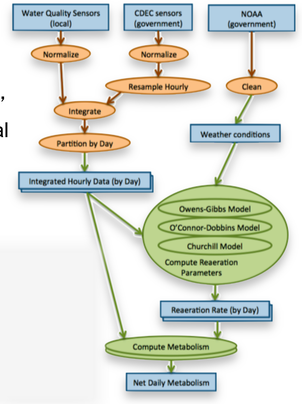
Current Approach

- Multiple, separate tools
- High learning costs
- Ad hoc, by-hand movement of data and tool invocation
- Data do not "flow" across tools

For example, stream metabolism estimates

•Oxygen reaeration
•Gross Primary Productoin (GPP)
•Community Respiration (CR24)
•Net Daily Metabolism = GPP – CR24

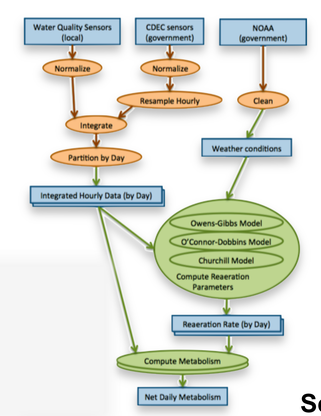
Investigator's conceptual workflow

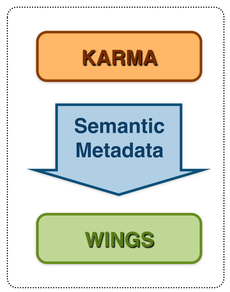


Manual Data Preparation

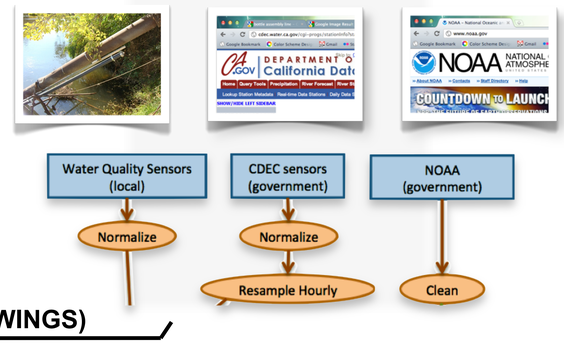
Custom Scripts

Proposed Approach → Data Integration (here, KARMA)

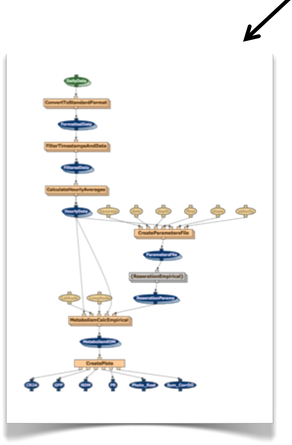




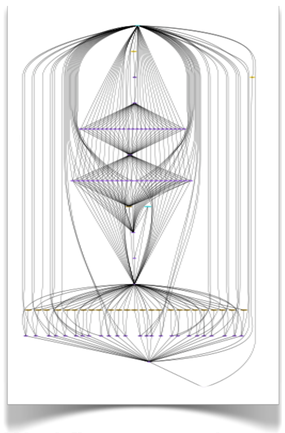
KARMA cleans and integrates heterogeneous data streams, and automatically generates semantic metadata for the resulting input files



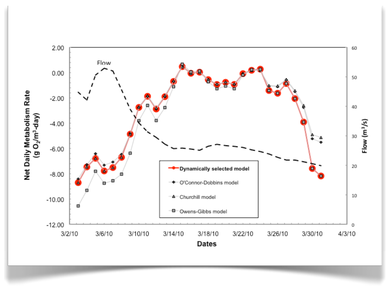
Scientific workflow (here, WINGS)



Scientific workflow



Workflow execution



Results

Source data lead to NDM estimates for changing flow rate; semantic metadata enable autonomous adaptation of algorithms (here, primarily the reaeration algorithm)

Workshop Recommendations (see WAES 2011 Final Report at http://water.isi.edu/)

- Data sharing approaches that reduce publication cost and provide immediate benefits to the scientist are important in order to rescue a lot of data in that may otherwise be lost.
- Workflow systems that manage and automate data processing steps are crucial to enable efficient processing of the volumes of data required to address multi-scale, grand environmental challenges.
- More efficient processing of environmental data would improve data collection by enabling rapid adjustment of sampling and sensor configurations.
- Pervasive provenance recording would improve reuse and productivity by facilitating the flow of data and processes across research groups and disciplines.
- Workflow sharing can drive and facilitate collaborative science projects that represent higher-impact science efforts.