Collaborating with computer scientists, informaticists*, and software developers for Integrated Ecosystem Assessments

Peter Fox 1, Heidi Sosik, Stace Beaulieu (stace@whoi.edu), and Joseph Futrelle 2, David Mark Welch 3, Jon Hare and Michael Fogarty 4

1 Tetraphase World Constellation (TWC), Rosenthaler Polytechnic Institute; 2 Woods Hole Oceanographic Institution; 3 Marine Biological Laboratory; 4 Northeast Fisheries Science Center

**ECO-OP Use Case: Ecosystem Status Report**

*What is an informaticist?*

More team members: M. Di Stefano, P. West 1; A. Ma 2; G. DePiper 3, K. Friedland, S. Gaichas, K. Hyde, R. Gamble, M. Jones, S. Lucey 4

**ECO-OP**: An abbreviation of ECO system and interOPerability

**Goal**: to develop and deploy a software environment to generate a portion of the Ecosystem Status Report for the Northeast U.S. Continental Shelf Large Marine Ecosystem, retaining traceability of derived datasets including indicators of physical pressures and ecosystem states.

Diagram for TWC Methodology. The use case defines the interactions between people, hardware, software, and desired products and can be adjusted or refined after each iteration of the cycle.

Outcomes:
- A pilot toward end-to-end transparency from scientists' desks to a report provided to policy makers and the public, important for science-based decision making.
- Small team with computer scientists and IT specialists working directly with fisheries scientists led to rapid results, with a limiting factor being sufficient training for adoption of technologies by the larger group of domain scientists.
- Manuscript under review in Earth Science Informatics: Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine integrated Ecosystem Assessments.

**ECO-OP Use Case: Linked Data Provenance**

More team members: X. Ma, L. Fu 1

**Goal**: to provide standardized provenance as metadata for data products, so that a human (and, ultimately, in the future, a machine) could trace back to the source observational data and models used to compile an indicator.

For the community standard, we chose the PROV Ontology for representing and exchanging provenance information as Linked Data in the Semantic Web.

Diagram for the three top classes in PROV-O and the properties that relate them.

Proposed implementation of a workflow using Python Notebook to generate a fisheries indicator.

Entities: 1 Python Notebook; 2 Cell; 3 Datasets; 4 script written in other programming language (R) that was split into five Cells; 5 other software environments.

Different agents are identified as contributing source datasets.

Activities: 1 CellRun; 2 other activities performed in other software environments.

**Outcomes**:
- Extended the PROV Ontology for capturing provenance in the IPython Notebook, a software platform that enables transparent workflows (https://git.hub.com/5/meyers/eco-ops/tree/master/pyco-op).
- Applied the prov-ecoop ontology to case studies that included the Climate Forcing Chapter, a regional map of primary production, and a fisheries indicator in the Ecosystem Status Report.
- Book chapter in press for Oceanographic and Marine Cross-Domain Data Management for Sustainable Development: Documenting provenance for reproducible marine ecosystem assessment in open science

**Work in Progress**

- Project website: https://twpi.edu/web/projectMBVL/
- Focusing on Objectives 1, 2, and 4 in this first year:
  1) developing data access and computational infrastructure for the MBVL;
  2) generating derived data products;
  3) producing traceable product workflows.
- We will work this summer with Matthew Ball, undergraduate student in computer sciences from Bowie State University, in the PEP program.

One can think of informatics as the steps and skills involved to make sense out of data – some of this is domain-specific (scientists are informaticists in their domains), and some of this is general to information processing or to the engineering of information systems.