Coping with Your Data:

WHOI Data Solutions and Resources Workshop

July 26, 2013

Slides for WHOI Ocean Informatics initiative:
Phase 1 (2009 - 2013) and Phase 2 (starting today)

WHOI Ocean Informatics Phase I

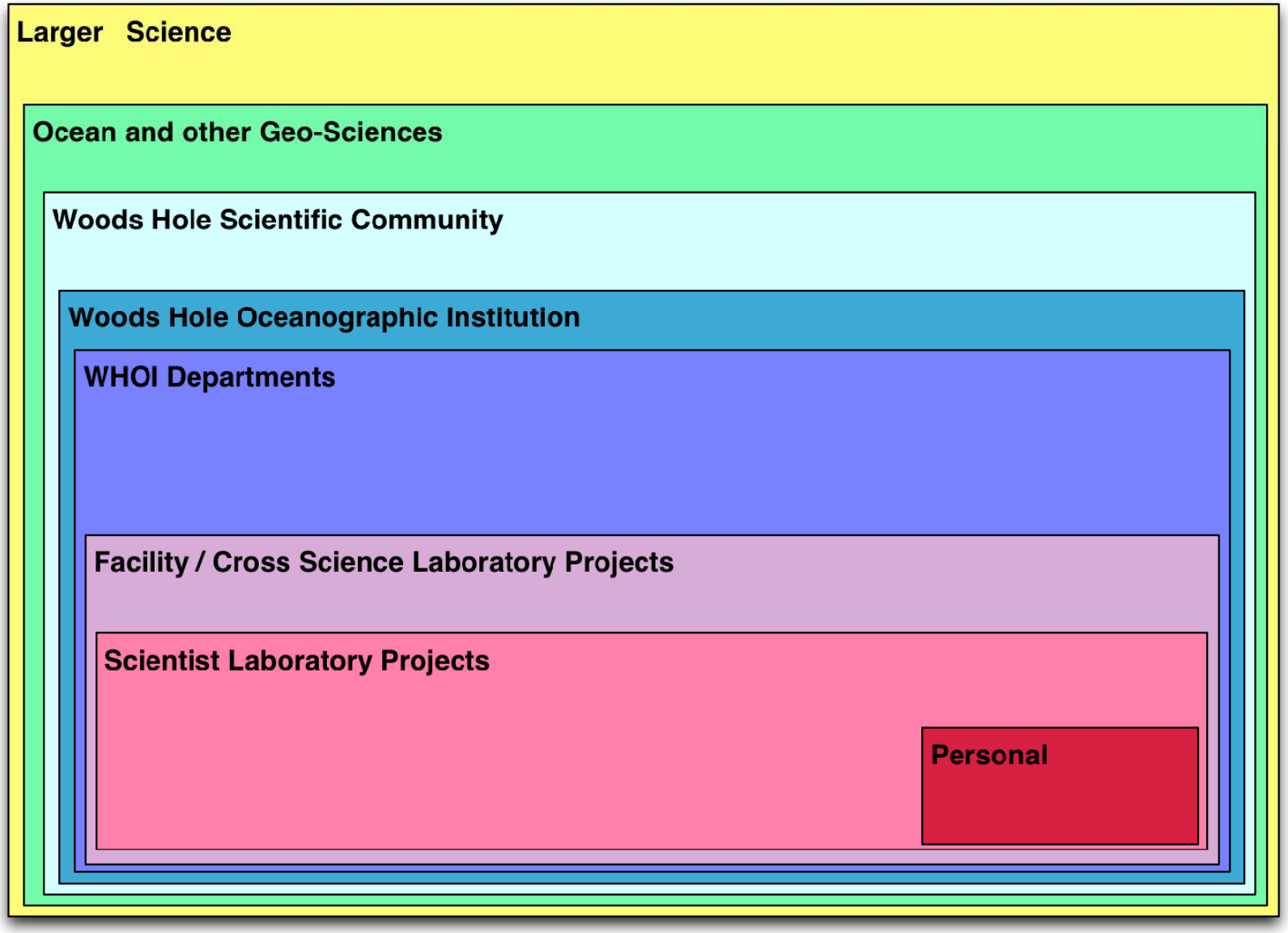
Coping With Your Data: WHOI Data Solutions and Resources Workshop

Andrew Maffei, July 26, 2013

Agenda

1. Overview of OIWG Phase I
2. Report
3. OIWG Accomplishments / Activities
4. WHOI's Ocean Informatics Assets
5. WHOI's Role in Informatics
6. Discussion about data challenges

Scoping of WHOI Informatics Efforts



1. Overview of OIWG Phase I

- Four members of OIWG Phase I: Art Gaylord, Peter Fox, Jen Schopf and myself.
- Focus in last 2.5 years has been working on projects and efforts setup by Ocean Informatics Working Group during 1st 2 years with goal of having concrete science lab scale informatics results to show to WHOI scientists
- Funding for efforts has been provided by GBMF OI ($1.2M), NSF/OCI ECO-OP ($252K), WHOI OIWG overhead budget ($81K), NSF/OCE BCO-DMO($???K), NSF/OCE R2R (~$564K), USGS Coop Agreement(~$42K)
- We now have a firmer concept about WHOI's potential role in national and global scale geo-science informatics initiatives
- A national-scale window of opportunity is upon WHOI scientists to contribute to national scale science geo-informatics efforts. It is time to decide our interest and commitment level.
- It's time for OIWG effort to inform WHOI scientists about what informatics, open-data, etc. can mean to their personal research and community research goals so they can evaluate next steps for WHOI in this realm

2. OIWG Report

- Report Contents
  - Summary and options for moving forward
  - App A – Ocean Informatics related projects and activities, their goals and status
  - App B – Emerging OI Toolbox for laboratory-level informatics
  - Strength, Weakness, Opportunity, Threat (SWOT) analysis
- Please review it and give us feedback!
  - printed copies available today, public version available in future.

To: WHOI Scientific Staff
Cc: WHOI Senior Technical staff
From: Andrew Maffei, Ocean Informatics Working Group Phase I Leader
Date: July 26, 2013
Subject: Report of the Ocean Informatics Working Group (OIWG) Accomplishments and Activities (2009-2013)

WHOI's Ocean Informatics Working Group (OIWG) has achieved its initial core mission of initiating and expanding a basic Ocean Informatics infrastructure at WHOI. Created in 2009 with seed funding from the Directorate, the OIWG, OI staff and partners have brought in external support, established new collaborations, and produced a suite of software tools along with other accomplishments detailed below (and in Appendix A). The question now is "What Next?". It is now timely to evaluate the progress made and choose a role and priority of Ocean Informatics for the Institution.

We encourage those scientific and senior technical staff interested in this topic to join in the next phase of this effort.. The "Coping with your Data" meeting held on July 26, 2013 will include demonstrations of the work done to date and an exploration of the appropriate priority, scope, and role of WHOI in the emerging cyber-infrastructure for Earth and Ocean sciences (geo-informatics). This meeting acts as a marker of the transition from phase I to phase II of this program as the first leadership team (Andrew Maffei, Peter Fox, Art Gaylord, Jen Schopf) hands it off to the next (Stace Beaulieu, Cyndy Chandler, Joe Futrelle, and Lisa Raymond).

Perspective and Outlook

During phase I of this effort we tried to direct the thrust of WHOI's several

3. OI-related Projects / Activities (23)

- Projects at various (scopes)
  - **BCO-DMO Ocean Data Ontology (ODO) / Advanced Search (Science Project)**
  - **WHOI Data Library & Archive Informatics (WHOI Institution)**
  - **RPI/TWC Semantics Methodology and usecases (WH Scientific Community / Project / WHOI Institution)**
  - **Underwater Ocean Imaging Informatics Program (Project / Institution / Ocean Community / Larger Science)**
    - **HabCam and IFCB**
  - **ECO-OP: Employing Cyber Infrastructure Data Technologies to Facilitate Integrated Ecosystem Assessments for Climate Impacts in NE & CA LME's (#3 & #7) (Project / Ocean Community)**
  - **R2R Eventlogger Development (National Ocean Community)**
  - **NSF EarthCube Initiative (Geo-science National Community)**
  - Rolling Deck to Repository (R2R) Repository (National Ocean Community)
  - Ocean Informatics Working Group Phase I Planning (WHOI Institution)
  - R2R Ship Operator Cruise Planning Website (National Ocean Community)
  - R2R CTD Data Quality Assessment Software (National Ocean Community)
  - R2R Ocean Data Interoperability Program (International Ocean Community)
  - PO Department-level Data Catalog (Department)
  - USGS Coastal and Marine Spatial Planning (CMSP)-related RPI/TWC Semantic Methodology (National Ocean Community)
  - Drupal Website Hosting and Support (WHOI Institution)
  - Interridge Drupal-7 "Hotvents" Database (International Hydrothermal Vent Community)
  - NDSF Data Management Subcommittee (National Ocean Community)
  - Research Data Alliance Initiative (International Science)
  - Information Modeling Services for Science Projects (WHOI Institution)
  - Marine Metadata Initiative (Internaional Ocean Community)
  - Database and Data Access for Microbial Metabolism Products (Science Project)
  - NMFS Use-case Workshop facilitation (WH Science Community)
  - Next Collaborative Informatics Proposal (WHOI Institution)

App A: OI-related projects

Appendix A – Table of WHOI Informatics Related Projects

The following table presents the informatics-related projects the Ocean Informatics Working Group has been working on since 2009. Project name, project goals pertinent to WHOI's ocean informatics work , and current status are given.

Project Name and Goals pertinent to WHOI Ocean Informatics	Status
BCO-DMO Ocean Data Ontology (ODO) and advanced search development *(BCODMO_ODO)* - Work with RPI-TWC to create a high-quality ontology as a centerpiece of an informatics infrastructure for oceanographic data that contains concepts and relationships that pertain to a broad range of repositories that contain oceanographic data. - Present the Ocean Data Ontology (ODO) to the broader oceanographic community for adoption as a community tool for documenting relationships between those oceanographic data held in repositories. - Help to spearhead a new era of oceanographic data access via the development of an "open linked-data network" of oceanographic data.	Funded, by NSF OCE BIO, semantically enabled faceted search has been deployed.
Ocean Informatics Working Group planning *(WHOI_OIWG):* - Explore ocean informatics strategies for WHOI researchers and the institution as a whole. - Educate science staff on what Ocean Informatics is - Establish funded informatics–related projects at WHOI that can be used to further demonstrate informatics advantages.	Funded by WHOI internal, "phase 1" complete.
WHOI/RPI-TWC Collaboration *(WHOI_TWC)*:	MOU signed.

4. WHOI's Ocean Informatics Assets

- Strong Informatics Collaborations (esp. P. Fox @ RPI)
- Informatics Bridging Team Concept, along with a growing set of informatics tools for various scopes
- Emerging interest and support from WHOI scientists beginning to see value in an informatics approach
- Novel ideas for informatics research and application emerging out of WHOI science laboratories (somewhat unexpected)
- A growing reputation for WHOI's Informatics expertise at a national level (the new OIWG team members)
- An emerging Ocean Informatics Toolbox

Emerging Laboratory Scope Informatics Toolbox Tools (12)

- Ocean Data Ontology
- RPI/TWC Use-case Development Methodology
- Local Data Resolver
- Openstack Environment for Science Lab Scope VMs
- Redmine Lightweight Project Management
- Science Source Code Repositories
- Actionable URL Coding
- Customizable Ocean Image Browser
- Customizable Ocean Image Manual Annotator
- Workflow Webservices to Accelerate Ocean Image Data Analysis
- Laboratory-scale Informatics Partnering Methodology
- iPython for visualizing Provenance and Publishing code

App B: Emerging OInformatics Toolbox

Appendix B – Emerging OI Toolbox for Project and Laboratory Level Informatics Partnering

- **Ocean Data Ontology** – WHOI/NSFs BCO-DMO project has worked closely with the staff at RPI/TWC over the past 3+ years to develop a high quality ontology representing concepts that are important for oceanographic data. The ODO ontology will be announced at the December 2012 AGU meeting. The BCO-DMO advanced search capability can now demonstrate to WHOI researchers the advantages of mapping to community vocabularies and employing ontologies. We plan to use the BCO-DMO example as a way of explaining these advantages to WHOI researchers.

- **RPI/TWC Use-case Development Methodology** – Many of the researchers and technical staff mentioned above have been trained in the use-case centered methodology taught as a core of the RPI/TWC graduate program. It has proven itself to be very effective in identifying and solving real problems of scientists and provides a common language and set of tools for these staff to employ in the informatics arena. New projects are begun by customizing the methodology to meet the unique needs of a project/science laboratory, keeping it's core principles intact.

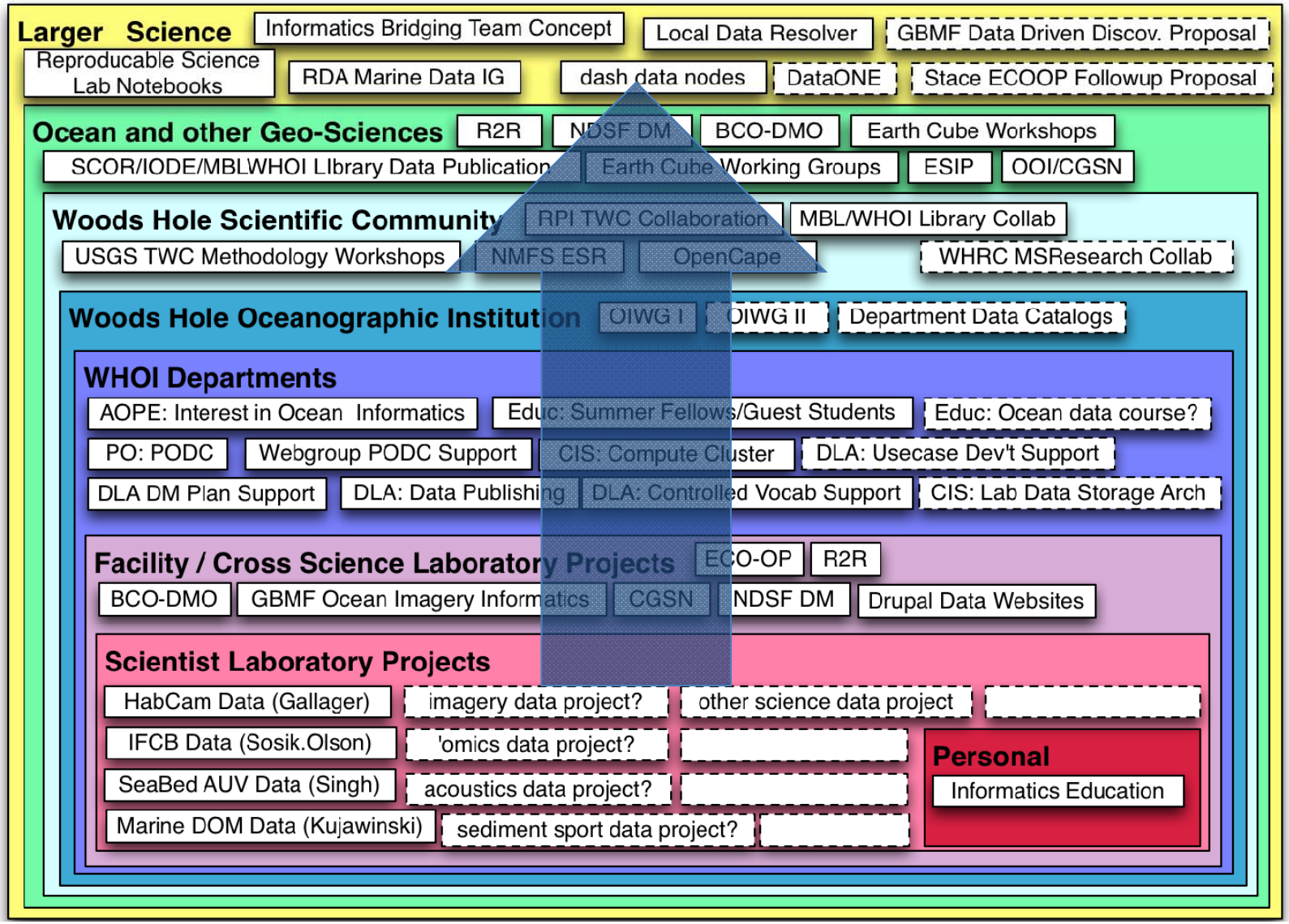
- **Local Data Resolver** – Joe Futrelle has developed a "middleware" package so that scientific code can be more easily written to access scientific data via URLs rather than static files system based names. The package is designed to eventually support data sitting on scientist's individual laboratory data servers, the WHOI FTP Server, Dropbox, larger science project servers, and even larger discipline-specific repositories. This is part of the "actionable URL" approach described below

- **Redmine Lightweight Project Management** – Early in the Ocean Imaging

5. A Proposed role for WHOI in field of Informatics for your consideration

- ***GROUNDING of "cyber-infrastructure" efforts by maintaining a strong focus on solving specific laboratory-scoped data challenges. When this is perfomed "thoughtfully" we see that this approach leads to solutions for larger scopes of science need.***
  - **Thoughtful partnering** between Joe F. and PIs he serves expose the latent benefits in this approach at broader scopes -- if we can learn how to make it scale and be sustainable
  - WHOI's scientist-engineer partnering legacy of ***going to sometimes heroic efforts to make things work before a ship leaves port for a cruise*** – is part of WHOI's culture and we have seen this attitude benefit WHOI"s scientist-informatics partnering efforts
  - The **RPI semantic development methodology** is proven to work well in providing a useful way to define and focus on "real" laboratory scope use-cases.
  - Several **science projects are ripe for informatics** partnering efforts **–** if we can find a way to make this approach sustainable.
  - **WHOI scientists could choose to partner in a national scope** effort to develop an informatics research environment that promotes "grounding" as a way to create better science tools .

Scoping of WHOI Informatics Efforts



6. Discussion about data challenges

- What are the most difficult challenges in your lab when it comes to working with your data?
  - Earlier input from meeting participants
  - 8x5 index cards at tables

What's next?

Ocean Informatics at WHOI: Phase II

Contacts:

Stace Beaulieu

Cyndy Chandler

Lisa Raymond

Joe Futrelle

Others interested??

Question card:
Who or how can I get help to cope with my data?

4 goals for Phase II

Goal 1: Strengthen connections to expertise within WHOI and our local science community

Goal 2: Become aware of broader community efforts and external opportunities that may lead to new funding

Goal 3: More student involvement

Goal 4: Help with data and information management plans

Goal 1: Strengthen connections to expertise within WHOI and our local science community

Email responses to:
What are the biggest barriers for you to work with your data?

Many responded with specific examples of projects that could benefit from technical expertise, for example:
"We would like to organize storage of those images, and streamline access and cataloging of particle-tracking files and other analyses associated with each set of images."

Referring in general to help with technical expertise:
"I worry about how to make this kind of help more easily accessible here at WHOI."

Goal 2: Become aware of broader community efforts and external opportunities that may lead to new funding

Email responses to:
What are the biggest barriers for you to work with your data?

"And of course the elephant in the room: How do we FUND proper and successful data management?"

"knowing of any funding sources for data preservation, or conversion, or conservation would be very helpful to me."

Goal 3: More student involvement

Email responses to:
What are the biggest barriers for you to work with your data?

"My time - too many things going on."

- JP student training (e.g., upcoming Software Carpentry workshop)
- Summer Student Fellows and guest students

Goal 4: Help with data and information management plans

Email responses to:
What are the biggest barriers for you to work with your data?

"…my list of what drives me crazy would be:
1. What data formats should be used to submit data …at the end of NSF projects?
…4. What metadata should be included?
…6. What stage should the data be in? I edit out data collected when the instrument was out of the water… Is this enough editing or is it too much editing?
…9. Lastly, where do we put our data once it is ready? I think I would put my data at the NODC website, but how is that done?"

We may revise our goals based on your input today:

- Google doc
- Chart paper on easels around the room
- Email (stace@whoi.edu)

Notes contributed to Google doc, linked at:

http://www.whoi.edu/DoR/page.do?pid=123176

Did you select your table based on the question card?

What file formats should I use to manage my data?

Where can I get funding for data management, migration, and preservation?

How should I migrate/preserve my data?

What metadata standards/controlled vocabularies should I use?

What tools are available to organize and make my data accessible?

Should I embargo my data or make it open access?

How do I cite data and assure proper attribution?

Who or how can I get help to cope with my data?