Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine Integrated Ecosystem Assessments

Stace E. Beaulieu[1]*, Peter A. Fox[2], Massimo Di Stefano[1,2,3], Andrew Maffei[1], Patrick West[2], Jonathan A. Hare[4], Michael Fogarty[4]

[1] Woods Hole Oceanographic Institution, Woods Hole, MA 02543 USA

[2] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

[3] Center for Coastal and Ocean Mapping, University of New Hampshire, Durham, NH 03824 USA

[4] Northeast Fisheries Science Center, National Oceanic and Atmospheric Administration, Woods Hole, Massachusetts 02543 USA

* stace@whoi.edu, +1-508-289-3536

20 October 2016

Revised #ESIN-D-16-00075 for *Earth Science Informatics* as a research article

Abstract

There is a growing need for cyberinfrastructure to support science-based decision making in management of natural resources. In particular, our motivation was to aid the development of cyberinfrastructure for Integrated Ecosystem Assessments (IEAs) for marine ecosystems. The IEA process involves analysis of natural and socio-economic information based on diverse and disparate sources of data, requiring collaboration among scientists of many disciplines and communication with other stakeholders. Here we describe our bottom-up approach to developing cyberinfrastructure through a collaborative process engaging a small group of domain and computer scientists and software engineers. We report on a use case evaluated for an Ecosystem Status Report, a multi-disciplinary report inclusive of Earth, life, and social sciences, for the Northeast U.S. Continental Shelf Large Marine Ecosystem. Ultimately, we focused on sharing workflows as a component of the cyberinfrastructure to facilitate collaboration and reproducibility. We developed and deployed a software environment to generate a portion of the Report, retaining traceability of derived datasets including indicators of climate forcing, physical pressures, and ecosystem states. Our solution for sharing workflows and delivering reproducible documents includes IPython (now Jupyter) Notebooks. We describe technical and social challenges that we encountered in the use case and the importance of training to aid the adoption of best practices and new technologies by domain scientists. We consider the larger challenges for developing end-to-end cyberinfrastructure that engages other participants and stakeholders in the IEA process.

Keywords

e-Science, executable workflow, indicator, IPython Notebook, open science, use case methodology

Acknowledgements

We would like to thank others in the Ecosystem Assessment Program who contributed to use case development, including G. DePiper, K. Friedland, S. Gaichas, K. Hyde, R. Gamble, M. Jones, and S. Lucey. We would like to thank our colleagues for commenting on this manuscript, including B. Lee, J. Futrelle, X. Ma, A. Shipunova, A. Voorhis, and S. Zednik, and three anonymous reviewers. Support for this research was provided by the U.S. National Science Foundation #0955649 with additional support to SB by the Investment in Science Fund at Woods Hole Oceanographic Institution.

Introduction

There is a growing need for cyberinfrastructure to support science-based decision making in management of natural resources (e.g., Acreman 2005; Reichman et al. 2011; Palmer 2012; Muste et al. 2013; Horsburgh 2015). Over the past decade the U.S. has moved toward an ecosystem-based management approach for marine ecosystems, and there is a need for development of cyberinfrastructure to support the science teams who are reporting on these ecosystems and provisioning services such as fisheries. We were motivated to develop cyberinfrastructure to provide a transparent pathway from data to knowledge to action, responding to the U.S. National Ocean Policy Implementation Plan, in particular "improving science-based products and services for informed decision-making" (National Ocean Council 2013). Here, we define cyberinfrastructure as infrastructure that comprises "*both technology and human expertise necessary to support scientific research processes and collaboration*" (Jirotka et al. 2013). Levin et al. (2009, 2014) and Samhouri et al. (2014) describe a formal process for an Integrated Ecosystem Assessment (IEA), involving natural and social scientists working together to assess a marine ecosystem with respect to management objectives (Fig. 1). Data collected, integrated, and interpreted in a marine IEA may be as diverse as climate indices, satellite-derived sea surface temperature, counts of phyto- and zooplankton from net tows, and landings data from commercial fisheries.

For any coupled natural and human system it is challenging to develop cyberinfrastructure to enable multi- and inter-disciplinary research to understand, model, and make predictions for the system as a whole. Technical challenges include handling, integrating, analyzing, and tracking provenance of very heterogeneous data (e.g., Reichman et al. 2011). In an IEA to make sense of a plethora of data, it is common practice to focus on a select subset of indicators of natural or anthropogenic drivers or ecosystem states that can be monitored for changes over time and space (Samhouri et al. 2012). Indicators tend to be derived datasets and are often “synthesized products” (term used in NOAA 2014), resulting from complex data processing workflows that integrate not only data and models but also subjective choices made by scientists based on knowledge in their domain. Social challenges include scientists of different domains using different terms to describe their data and different software and tools to work with data (e.g., Pennington 2011; Cooke and Hilton 2015). E-Science teams inclusive of scientists and information technology (IT) experts face the additional challenge that “IT experts cannot understand the needs of the scientists – and scientists cannot understand what is even possible – without conceptual integration between the scientists and IT experts” (Pennington 2011).

Here we report on the ECO-OP (an abbreviation joining ECOsystem and interOPerability) project involving fisheries scientists, oceanographers, computer scientists, information modelers, and software developers. As part of this project, we identified and conducted a use case to support the bi-annual generation of an Ecosystem Status Report (hereinafter the Report) as part of an IEA for the Northeast U.S. Continental Shelf Large Marine Ecosystem. The Report is composed of chapters, each of which is prepared by different specialists for climate forcing, physical pressures, primary and secondary production, benthic invertebrates, fish communities, protected species, anthropogenic factors, and integrated ecosystem measures (Ecosystem Assessment Program 2012). The software framework to be developed needed to enable these different specialists to process heterogeneous data and provide products for the Report. The framework would be flexible to allow for addition and subtraction of indicators from the Report and portable to accommodate assessment of marine ecosystems in other managed regions of the ocean.

The ECO-OP project addressed challenges in developing cyberinfrastructure for e-Science teams participating in marine IEAs. Following our definition of cyberinfrastructure above, our use case for the Report involved integrating *technologies* ranging from data sharing (including access and re-usability) to executable workflows and *human expertise* including knowledge and practices in multiple natural and social science domains. In the spirit of open science (Reichman et al. 2011; Nosek et al. 2015), we aimed beyond transparency toward the reproducibility standard in the U.S. NOAA Information Quality Guidelines (NOAA 2014) for indicators and other data products in the Report. Below, we describe the software prototype that we developed and how we aided its adoption by the scientists producing the Report. We discuss how to scale the prototype and other considerations for the larger cyberinfrastructure to be developed for the IEA process.

Methods

Methodology to develop cyberinfrastructure and evaluate the use case

We employed a bottom-up approach in which a small team with diverse skills worked closely to evaluate use cases with very specific goals as representative of a larger set of goals. This approach engages domain scientists directly in the collaborative development of a software solution. The use cases were iteratively developed to articulate specific goals of fisheries scientists delivering indicators and data products, capture detail on what went into reaching those goals, and the outcomes they needed to evaluate success. Computer scientists and software developers provided options for technologies which were then evaluated to determine how they could be adopted and then how they could be incorporated into a larger framework of cyberinfrastructure. In addition to engaging with fisheries scientists in the use case evaluation, informatics and software experts in the small team also regularly attended science meetings to learn more about the science, understand concepts, share ideas, and build trust. This methodology is in contrast to top-down approaches that prescribe technologies for domain scientists as end users.

The use case for the Report explored options for the portion of the IEA process including "Develop Indicators," "Monitoring of Ecosystem Indicators," and "Assess Ecosystem" (Fig. 1). We provide a diagram as an overview of the data-level and application-level mediation requirements to compile the Report (Fig. 2). We also show representative temporal and spatial indicators as derived data products in the Report (Fig. 3). We evaluated the use case through the Tetherless World Constellation (TWC) Semantic Web Methodology (hereafter, TWC Methodology), a collaborative process of rapid prototyping based on a small team including domain scientists (Fox and McGuinness 2008). Essentially, the small team was a subset of a larger e-Science team collaborating on a prototype Report. The TWC Methodology is a cycle involving ten stages (Fig. 4):

(1) 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. The initial goal of the use case for the Report was to efficiently generate figures representing ecosystem data and information products; this goal was expanded to be inclusive of generating the Report documents [portable document format (PDF) and associated webpages].

(2) The small team with mixed skills met initially to define the use case and then subsequently (in stage 10 described below) to evaluate each prototype to complete an iteration of the cycle. The authors of this paper comprise the team for the use case: facilitator (Fox, Maffei), domain experts [Hare, Fogarty, and other scientists in the Ecosystem Assessment Program at NOAA's

Northeast Fisheries Science Center], knowledge representation and information modeling (West), software engineering (Di Stefano), and scribe (Beaulieu). The larger group of fisheries scientists contributing to the Report comprises ~40 individuals working at ~10 different NOAA offices and academic institutions.

(3) Analysis of the use case included identifying the actors and source data, writing a narrative description, outlining a flow, and drawing an activity diagram (Fig. 5). Expectations ultimately were refined to the following: The framework should retrieve data, report quality assurance/quality control, conduct standard analyses, provide iterative and interactive visualization, allow for interpretation, and generate final graphics to embed into webpages and PDF. In addition, the data represented in each figure should be available. The framework should also document the specific process for each data and information product, including source data, code, and related contextual information suitable for traceability, repeatability, explanation, verification, and validation. The framework should use the same components/structure for each data and information product, thereby allowing the addition and subtraction of data and information products in future Reports.

(4) Neither an information model nor ontology was formally developed in the Report use case. However, we explored and mapped concepts that were important to document as metadata, due to different terms being used by different actors in the use case. In this project our use of "semantics" in the TWC Methodology involved "developing shared conceptualizations across disciplinary boundaries" *sensu* Pennington (2011).

(5) The TWC Methodology advocates finding and using relevant tools; thus, we tested a number of existing open source tools as we iterated the prototype including Drupal, Wt (the C++ Web Toolkit), and the IPython (now Jupyter) Notebook (Pérez and Granger 2007; Ragan-Kelley et al.

2014; Shen 2014). In particular, the IPython Notebook is an "interactive computational environment" with a web application and "notebooks, for recording and distributing the results of the rich computations" (https://github.com/ipython/ipython-website/blob/b578013e545d18deafa0f9e1567e3db5368f0cf6/notebook.rst l, accessed 17 October 2016).

(6) Science/expert reviews occurred within each iteration of the cycle as the prototype was being developed for the next major group evaluation.

(7 & 8) We adopted technologies that were available as open source and leveraged the technology infrastructure (hardware and software) that the fisheries scientists were already using to generate indicators. Cooke and Hilton (2015) provide a comprehensive list of factors to consider when selecting technologies for e-Science teams (e.g., ease of use, accessibility, security, compatibility).

(9) The initial rapid prototype acted “to glue the components together and connect them to interfaces and visualization tools. ...latter stages of the prototype must pay increasing attention to non-functional aspects of the use case, such as scalability, reliability, etc.” (Fox and McGuinness 2008).

(10) The final stage is evaluation of the prototype to determine whether/how it should be redesigned and redeployed. In practice this stage involves demonstration of the software prototype to the larger e-Science team and then an evaluation by the small team to complete the iteration of the cycle.

We developed prototypes for the Report use case during three complete iterations of the TWC Methodology. Each iteration of the cycle took a few to several months, accounting for the time to develop and test software, and demonstrate and evaluate each prototype. The fisheries scientists requested transfer of the technologies after demonstration of the third iteration prototype, which focused on the “Climate Forcing” and “Physical Pressures” chapters in the Report (Ecosystem Assessment Program, 2009). Prior to the delivery to fisheries scientists, the small team conducted three small "spin-off" use cases to further test the software prototype. These small use cases were intended to examine whether the prototype that was successful for one portion of the Report could also be adapted for indicators and data products from other chapters in the Report (Ecosystem Assessment Program, 2012). We delivered the prototype software environment to the fisheries scientists in two ways: in a virtual machine (VM) provided to individuals, and by installation on a server at the Narragansett facility with the aid of NOAA’s IT staff.

Training to aid adoption of the technologies

During each iteration of the cycle described above, the e-Science team gains some exposure to the cyberinfrastructure inclusive of technologies and others’ expertise, but it is mainly the small team that gains hands-on experience with the software prototype. Additional training and hands-on experience is desired to aid adoption of the technologies by the larger team. We provided training opportunities and technical support in groups and for individuals, as recommended by Cooke and Hilton (2015). In the first iteration prototype, fisheries scientists were introduced to several applications that were new to them: interactive programming software (IPython Notebook), version control software (Subversion), and content management systems (including Trac and Drupal). Ultimately we focused the training on IPython Notebook and changed to version control with GitHub. We offered three group training workshops, two of which were specific to ECO-OP cyberinfrastructure. The first workshop, which involved the second iteration prototype, was essentially an introduction to IPython Notebooks utilizing a shared online server that the e-Science team logged into as users. During the one-day workshop and for a few months afterward (as we were conducting the third iteration of the use case), users were provided folders on the shared server to store their notebooks and data products. The second workshop was provided after we completed the final prototype and was aimed towards learning Python programming and best practices for version control. This training involved a two-day Software Carpentry Bootcamp (Wilson 2014) held at Northeast Fisheries Science Center and was also open to other fisheries scientists. The third workshop was to assist the e-Science team in using the final prototype - i.e., ECO-OP pyecoop software library distributed within a VM - to generate data products specific to their chapters of the Report. The purpose of this final training over 2.5 days was to assist with user-specific, individual needs (we asked participants to come with their own data and code).

Results

Initial prototypes

As a first step towards developing the prototype Report, the small team sketched an activity diagram which identified the primary actors in the collaboration, including many people (e.g., data preparation reviewer, Report compiler/editor) and a software agent (Fig. 5). Pre-conditions for the use case included that source data are accessible. The basic flow for the use case may be described as: Source data are retrieved > Source data are processed into preliminary data products (which are stored) > Intermediate and final data products including indicators are calculated, analyzed, and plotted in an iterative and interactive process (and stored) > Indicators are interpreted > Text is written for context, interpretation, and synthesis > Report is compiled (and stored). Post-conditions for the use case, not explicitly addressed in the prototype, included storage and archiving of the preliminary, intermediate, and final data and visualization products and the Report itself.

During the first two iterations of the TWC Methodology, we were developing multiple software prototypes corresponding to different components of the desired cyberinfrastructure. The first iteration prototype targeted software tools for data access, data processing, metadata acquisition, and data visualization. We focused on the first two chapters in the Report, "Climate Forcing" which included climate indices [e.g., North Atlantic Oscillation; Fig. 2.1 in the 2009 Report (Ecosystem Assessment Program, 2009)] and "Physical Pressures" which included sea surface temperature anomalies [e.g., Fig. 3.5 in the 2009 Report (Ecosystem Assessment Program, 2009)]. The first iteration prototype separately considered a tool for data access and processing (IPython Notebook), tools for manual contribution of metadata in controlled vocabularies (Trac and Drupal), and other web applications for interactive display of final datasets. In practice, we utilized IPython Notebooks to output comma-separated value files for time-series indicators, we manually input metadata for these indicators to other file formats, we stored the data and metadata files at specific addresses, and the web applications called to these addresses to display one or more indicators. As a result of the evaluation of the first iteration prototype, the fisheries scientists were intrigued but not comfortable with IPython Notebook, mainly because this first demo involved converting code from one programming language (MATLAB) to another (Python) [not necessary in further iterations due to the availability of a Python-MATLAB bridge (and, now, also a Matlab kernel for Jupyter; Jupyter Team 2015)]. The fisheries scientists were not keen to learn tools to manually contribute metadata and requested that we focus on automated acquisition of metadata. They also requested that we further customize a web application for interactive display of the indicators. In response the small team sketched a Graphical User Interface (GUI) with a drop-down list to select indicators, more options for plotting, and buttons for exporting data and visualization products, viewing metadata, and saving a session.

For the second iteration prototype we built a web-app GUI using Wt that could be displayed on its own or within an IPython Notebook. We recorded a demo to show the larger e-Science team how to use the web-app GUI for interactive display of the indicators and how to log in and use both the IPython Notebook and the web-app GUI to re-calculate an indicator with the latest version of code, then store and display the final data file. To support this human-oriented process we implemented a shared server to contain the development environment and allow for easy sharing of notebook files and the output data files, images, and PDFs. Converting notebooks into PDFs was a key new development made possible with the nbconvert tool, which also handles other formats including HTML and LaTeX (Frederic 2013). We continued to focus on indicators in the “Climate Forcing” and “Physical Pressures” chapters of the Report but also performed workflows using IPython Notebooks for ecosystem indicators, including a phytoplankton abundance anomaly (Di Stefano et al., 2012) and time series of copepod abundance [Fig. 4.10 in the 2009 Report (Ecosystem Assessment Program, 2009)].

To evaluate the second iteration prototype, we distinguished three levels of users: users of an interactive PDF for the Report with hyperlinks to data and metadata (Level 1), users of the web-app GUI to access final data products (Level 2), and users interacting with IPython Notebooks (Level 3). A major result of the evaluation was that the fisheries scientists aspired to become Level 3 users and asked to have an IPython Notebook tutorial as soon as possible. The overall assessment was that the IPython Notebook technology offered the most flexibility for calculating, analyzing, and plotting indicators for the Report and would also enable the production of an interactive PDF. The fisheries scientists requested that we explore further the conversion of notebooks to HTML, as the group was considering providing the Report directly online as a website. Essentially, the IPython Notebook appeared to be a single tool that could accommodate components considered separately in the first iteration prototype.

Final prototype

The third prototype focused on the IPython Notebook tool and ultimately was refined to the final prototype delivered to fisheries scientists. Much of the development in the third iteration of the use case involved building a software library for processing, analyzing, and visualizing indicators in IPython Notebooks and an environment to accommodate all the dependencies. Our first "spin-off" use case was to test the conversion of an IPython Notebook to an Ecosystem Advisory webpage. We used a notebook created in the first iteration prototype for the "Physical Pressures" chapter to successfully reproduce a webpage in HTML format for long-term temperature trends in the Northeast U.S. Shelf ecosystem (Di Stefano et al., 2013). The demonstration of the third iteration prototype included this simulated Ecosystem Advisory webpage and a notebook (Fig. 6) that retrieved and processed data for two climate indicators and output an interactive PDF (Fig. 7) formatted to look exactly like a portion of the "Climate Forcing" chapter in the Report (Ecosystem Assessment Program, 2009). This notebook (Fig. 6), which requires the installation of TeX Live [TeX distribution for several Linux distributions (https://www.tug.org/texlive/)] into the environment, utilizes the pdflatex command to compile text files with image files created on-the-fly as a result of data visualization in the notebook. The interactive PDF (Fig. 7) included embedded links to data files plotted in the figures.

As a result of the evaluation of the third prototype, the fisheries scientists determined that the expectations for the use case were met. However, prior to the transfer of technologies, they requested that we address some of the challenges in reproducing other chapters of the Report. Our second and third "spin-off" use cases examined challenges in reproducing the workflows for a fisheries indicator (Fig. 3a) and a map of primary production (Fig. 3b) from other chapters in the Report (Ecosystem Assessment Program, 2012). For both of these use cases, our goal was to determine whether a complex workflow utilizing many data sources, multiple tools, and multiple programming languages could be accommodated with an executable workflow in an IPython Notebook. We worked directly with the fisheries scientists responsible for these data products in the Report to determine the earliest point at which the prototype developed for the Report use case (dashed box in Fig. 5) could apply to their respective workflows. The fisheries indicator is constructed by a natural scientist and a social scientist working together. Their workflow had a number of manual steps in accessing multiple data sources and preparing preliminary data, including the use of a manual data query extraction tool. However, the remainder of the workflow involving these preliminary data products could be conducted within an IPython Notebook with an extension for the R programming language (now, an R kernel for Jupyter; Jupyter Team 2015). The map of primary production is constructed by one scientist and involves an even more complex workflow that starts with accessing thousands of source data files. The scientist utilizes SeaDAS (http://seadas.gsfc.nasa.gov) tools and Interactive Data Language (IDL) to process data and construct the map image. At the time although SeaDAS tools could be implemented in a Python environment, there was no extension for IDL in IPython Notebook. Today, Jupyter has an IDL kernel (Jupyter Team 2015), and the scientist should be able to create a notebook to execute the complete workflow from source data retrieval to outputting a figure for the Report, without having to convert code into Python.

The final prototype was a software environment for Linux operating systems inclusive of a software library with general utility to enable the reproducibility of scientific workflows that acquire data online, process and plot data, and package text and figures into a document. Workflows are conducted within IPython Notebooks. The ECO-OP pyecoop software library is available at a GitHub repository with GNU Lesser General Public License, accessible via https://data.rpi.edu/xmlui/handle/10833/1756. The pyecoop software library, written in Python (>=2.7 , >=3.3), has several modules including a module with utility functions (ecoop.ecooputil) and a module that defines methods for data in the “Climate Forcing” chapter of the Report (ecoop.cf). Dependencies for the pyecoop code include the installation of TeX Live and RubyGems (https://rubygems.org/). Other Python libraries are required, including matplotlib (Hunter 2007), pandas (McKinney 2010), and scipy (Jones et al. 2001). The software environment includes IPython Notebook and other open source applications used in generating indicators and documents, such as Geographic Resources Analysis Support System (GRASS Development Team 2015), Octave (Eaton et al. 2014), and R (R Core Team 2013). The software environment was distributed within a VM (important for when users are not online) and by installing a single-port instance on a server at NOAA’s Narragansett facility. Ultimately the components of the delivered cyberinfrastructure included software and human resources (including training described below) but excluded hardware resources. We did not prescribe data storage or archiving, and the Report use case did not require support for high performance computing (this may be required for other use cases involving ecosystem modeling).

Results of training to aid adoption of the technologies

We provide some results for our first and third group training opportunities which were specific to ECO-OP cyberinfrastructure; however, we did not conduct surveys or interviews for a more rigorous evaluation of the training. Thirteen fisheries scientists participated at the first workshop. The most positive result was that one month after the training, one of the fisheries scientists was using IPython Notebook to develop and document new indicators, utilizing extensions to enable functionality for other programming languages. Upon seeing these new notebooks, another fisheries scientist joined the shared server (available in the second prototype) as a new user and aided the development of the notebook for the Ecosystem Advisory webpage that was part of our third prototype demonstration. Eight fisheries scientists participated at the third workshop; six did not attend the first training which placed them at a disadvantage since we assumed some familiarity with IPython Notebooks. At least one attendee was able to generate a PDF with their own data and code. All attendees left the workshop with the software requirements installed and configured in a VM on their own laptops. The environment provided to each attendee with the VM was fully compatible with the software infrastructure installed on the server at NOAA’s Narragansett facility. Comparing these two training opportunities, the first appeared to be more successful with the single shared software environment; we think that we lost users when each distribution was installed separately as a VM, not only due to challenges in the installation but also in terms of having to use email or other shared storage services to share notebooks. Importantly, the training was of benefit not just to the users, but also to the small team developing the software environment, to observe the challenges expressed by domain scientists with a range of skills. The first training session aided development during the third iteration of the use case. The third training session was conducted after deciding upon the final prototype and helped us with documentation prior to delivery.

Discussion

Solution for sharing workflows and delivering reproducible documents

Our solution for the fisheries scientists to reproduce a portion of their Report was a software environment in which IPython Notebook acted as a lightweight, flexible, re-usable, scientific workflow technology to document data processing, analyses, visualization, and reporting. The solution is in the spirit of open science in which the sharing of workflows engenders trust in the derived data products (Reichman et al. 2011; Nosek et al. 2015; Wright 2016). We recognize that the delivered prototype, which reproduced a portion of the "Climate Forcing" chapter in the Report (Fig. 7) and accommodated workflows for a variety of other ecosystem indicators, only addressed a limited set of technical and social challenges involved in preparing and compiling the Report. We addressed many challenges in terms of software required to execute the workflows (e.g., use of different programming languages, integrating with open source software libraries); however, we were not able to fully address challenges in the sharing of these workflows. We did not go so far as to enable a repository, management system, or social network for the sharing of workflows (e.g., Goble et al. 2010; Liu et al. 2015). Ultimately we were limited in implementing a shared file system in the final prototype, although this may be more straightforward to develop today due to recent developments for multi-user servers for notebooks (e.g., Wakari, JupyterHub).

We successfully reproduced a portion of one chapter and additional indicators, but an ultimate goal would be to enable a Report "on-demand" (at the time of this project, production of the Report was manually intensive and limited to every two years). Many technical and social challenges arise when considering the compilation of the entire Report as a reproducible document, a reason why we drew this step outside of the dashed box in the activity diagram (Fig. 5). A major challenge at this time would be the accessibility of source data for the many data processing workflows. For reproducibility in the future, the cyberinfrastructure would also need to account for versioning of IPython Notebooks for each data visualization product. The main technical challenge that we highlight here is sustaining a computational infrastructure for all of the e-Science team members' software environments and dependencies inclusive of repository(ies) with version control. This assemblage of very dynamic and distributed software environments is analogous to a "scientific software ecosystem" in recent publications (e.g., Howison et al. 2015). In addition, to reproduce all of the chapters, all of the fisheries scientists would need to adopt new technologies, which we address below.

Training to aid adoption of the technologies

Our experience with fisheries scientists provides a specific example of the general importance of training and professional development when selecting technologies to support multi-disciplinary e-Science teams (e.g., Cooke and Hilton 2015). We recognized with the initial prototypes that training would be central to our success in transferring the software environment to fisheries scientists. One measure of success for our delivered prototype is how the fisheries scientists used the technologies for their subsequent Report and other work conducted for the IEA process. We expected our bottom-up/user-driven approach to promote adoption of technologies based on research “finding that technical systems that were well aligned with and ready to accomplish the task scientists intended were more likely to be successfully adopted by the community” (Olson et al. 2008). Ultimately, only a few fisheries scientists utilized the prototype to produce portions of the subsequent Report. This may in part be due to technology readiness for the scientists (e.g., many had never interacted with a Linux operating system, and/or had no experience with the Python programming language). As noted by the iMarine project described in the next section, “in the domain of fisheries, marine biology and environmental sciences... users and researchers generally lack advanced IT skills” and “it is important to bear in mind the time to learn to use new tools” (iMarine 2014). Additional consultation and/or continued training was needed for fisheries scientists to build on and extend our prototype to produce chapters for the next Report. Pennington (2011) describes additional factors that influence technology adoption that may have been factors in our project, e.g., extrinsic motivation (which would be more applicable in a top-down approach).

In the long-term, perhaps more important than training to adopt specific technologies, our training encompassed best practices that were new to many of the scientists. Because technologies change frequently it is important for training to “generalise to broader classes of technologies and the socio-technical arrangements to which they point” (Jirotka et al. 2013). Including the Software Carpentry Bootcamp our training opportunities may be considered an attempt to grow the culture of best practices for data and software management in the community in which fisheries scientists work. Our training led to the broader use of open source tools and version control by scientists at the Northeast Fisheries Science Center. However, to build e-Science teams for new applications, there needs to be continued interaction with computer scientists, software engineers, and other IT experts.

Comparing our approach to other efforts to develop cyberinfrastructure for e-Science teams in IEAs

Our project involved a bottom-up approach in which a small team addressed very specific use cases as representative of a larger body of collaborative work for marine IEAs. The approach also involved the informatics and software experts engaging with domain scientists at their regular meetings to improve understanding of concepts and to develop relationships and trust in addition to the targeted use cases. At the end of each cycle of the TWC Methodology the small team shared the latest prototype with the larger e-Science team, thus directly involving end users in the evaluation. We aspired to prototype a software environment that would enable the flexibility for these end users to also become developers, re-shaping and expanding the software environment as needed to accommodate more data and information products in the Report. This lack of "clear delineations between users and developer" has been recognized in general for the development of technologies and infrastructure for e-Science teams (Jirotka et al. 2013). Our bottom-up approach is aligned with the Computer Supported Cooperative Work "focus on the scientists' everyday work practices, with a view to enabling new collaborations" (Jirotka et al. 2013), very much focused on the individual scientist and how s/he collaborates with other scientists contributing to an IEA.

Our approach is much smaller in scale than efforts that we highlight below from the European Union and Australia that also are directed toward cyberinfrastructure for IEAs. The European iMarine project is described as "an open and collaborative initiative aimed at supporting the implementation of the Ecosystem Approach to fisheries management" (http://www.i-marine.eu/Pages/Home.aspx, accessed 31 December 2015). Many of the goals of iMarine are similar to the ECO-OP project, including “facilitated retrieval, access, collaborative production and sharing of information and tools” (http://www.i-marine.eu/Pages/Home.aspx, accessed 31 December 2015). To achieve these goals iMarine provides web-based virtual research environments (VREs) through domain-specific infrastructure built onto D4Science e-infrastructure, “a virtual aggregator of resources available in interoperable e-infrastructures” (Taconet et al. 2014). Our interpretation is that scientists are users of the platform although they may be developers of workflows incorporated into the platform. As a future research effort we recommend exploring how to incorporate the ECO-OP prototype inclusive of executable workflows in IPython Notebooks into the iMarine platform.

For Australia we highlight the eReefs project, built upon “an innovative central information infrastructure reflecting best practice in environmental information management” (http://ereefs.org.au/ereefs/platform, accessed 31 December 2015). We draw an analogy between our Report use case and the “Report Card” of the eReefs Platform (http://ereefs.org.au/ereefs/platform, accessed 15 April 2016). In our use case we explored the use of a scientific workflow tool to account for processing source observational and model data into data visualization products, similar to the eReefs pilot (however, they used a proprietary tool; Chen et al. 2011). The ECO-OP project accounted for additional heterogeneity and issues of interoperability by addressing additional “spin-off” use cases and through a provenance use case described elsewhere (Ma et al. 2017). The current eReefs project (2012 - 2017) is intended to develop an information architecture to “allow for the next generation of data interoperability by augmenting established, standardised, services and allowing for the integration of multi-service use" (Car 2013). As a future research effort we also recommend exploring how to incorporate the ECO-OP prototype into the eReefs Platform.

We recognize that some of the challenges in scaling up and out when developing cyberinfrastructure with a bottom-up approach, differ from top-down development efforts. Top-down efforts may enforce policies or encourage the removal of technical or social barriers that inhibit broad usage of collaborative tools. However, although the ECO-OP project only addressed a small portion of the overall cyberinfrastructure that would be implemented within a VRE, we see most if not all of the socio-technical issues we considered critical to the success of our use case also applying to VREs (i.e., Jirotka et al. 2013, their sxn. 4.2). Our bottom-up approach in which the scientists (as end users of the infrastructure) are participating directly in the development of the infrastructure, was a nimble and rapid means to achieve the prototype Report. Our approach aligns with the concepts of "vertical user stories" in agile software development (e.g., Pulsifer et al. 2011) and participatory design (or co-design) in socio-technical systems (Muller and Kuhn 1993). Moreover, the adaptation of a more agile and iterative, i.e., quicker, sequence of try, evaluate, and revise indicates that future efforts to develop cyberinfrastructure for e-Science teams in IEAs (but also more generally) consider incorporating an agile approach or the small team/TWC Methodology as a means to supplement the larger development process.

Toward end-to-end cyberinfrastructure for the IEA process

The work conducted by scientists in the IEA process is embedded within a larger process involving other stakeholders in ecosystem-based management (Fig. 1). An ultimate goal is to extend the cyberinfrastructure developed for e-Science teams to address challenges at the science-policy interface including "... communication and debate about assumptions, choices and uncertainties, and about the limits of scientific knowledge" (van den Hove 2007). Essentially, cyberinfrastructure for the IEA process should encompass a virtual organization (*sensu* Ahuja and Carley 1998) of diverse stakeholders including scientists, decision makers, and the public. Our work in this project is just one example of the growing need for cyberinfrastructure to support science-based decision making in management of natural resources (e.g., Acreman 2005; Reichman et al. 2011; Palmer 2012; Muste et al. 2013; Horsburgh 2015). Our vision was to facilitate the engagement of natural and social scientists in routine ecosystem assessments, yet we aspire to involve other stakeholders through presenting robust science data in forms that various end users can consume and verify. This vision is shared by others developing cyberinfrastructure for IEAs including iMarine (Taconet et al. 2014) and eReefs (Car 2013).

The ECO-OP project provided a pilot toward end-to-end transparency starting from a scientist's desktop and being shared with collaborators, to a report provided to managers, policy makers, and the public. IPython Notebooks can be used as electronic lab notebooks, whereby scientists digitally record the steps involved in their computations and ultimate data products (Shen 2014). These notebooks essentially document a provenance chain, especially useful for indicators that summarize large collections of underlying heterogeneous data. Our solution included interactive and transparent workflows of data analysis and delivery of a reproducible document, but did not represent provenance in a machine-readable standard. After completing the use case with fisheries scientists described in this paper and to respond to the Executive Order for open, accessible, and machine-readable data (Obama 2013), the ECO-OP project explored a provenance use case to adopt the W3C PROV-O standard (Ma et al. 2017). As an example of a report using the PROV-O standard, the U.S. National Climate Assessment is incorporated into the Global Change Information System (GCIS) with a knowledge base that links data products, key messages, and certainty (Tilmes et al. 2013). Future efforts could bridge the ECO-OP prototype with GCIS or other information systems to represent provenance chains from acquisition of source data to inclusion of derived data products in interpreted figures in a report. As an example of analogous efforts, we note that the eReefs project includes integration with provenance and vocabulary services (Car 2013). We also note that semantic mediation may facilitate discovery, access, and understanding of data products by diverse stakeholders and recommend further development of a knowledge network to accommodate concepts in the IEA process (Fig. 2; Fox et al. 2012).

Conclusions

Our motivation was to develop cyberinfrastructure, including technology and human expertise, to enable routine, well-documented, integrated assessments of a marine ecosystem. The small team approach with computer scientists and IT specialists working directly with fisheries scientists and oceanographers led to rapid results, with a limiting factor being sufficient training for adoption of the technologies by the larger group of domain scientists. The prototype that we delivered for the Ecosystem Status Report for the Northeast U.S. Continental Shelf Large Marine Ecosystem enabled the reproducibility of a portion of a collaborative, multi-disciplinary report with very heterogeneous data types. However, we only addressed a limited subset of the many technical and social challenges in facilitating collaboration and reproducibility for the Report as a whole. This project provided 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 in the U.S. National Ocean Policy Implementation Plan.

List of abbreviations

ECO-OP, abbreviation joining ECOsystem and interOPerability;

GCIS, Global Change Information System;

GUI, Graphical User Interface;

IDL, Interactive Data Language;

IEA, Integrated Ecosystem Assessment;

IT, information technology;

NOAA, National Oceanic and Atmospheric Administration;

PDF, portable document formats;

TWC, Tetherless World Constellation;

VM, virtual machine;

VRE, virtual research environment;

References

Acreman M (2005) Linking science and decision-making: features and experience from environmental river flow setting. Environmental Modelling & Software 20:99-109. doi: 10.1016/j.envsoft.2003.08.019

Ahuja MK, Carley KM (1998) Network structure in virtual organizations. Journal of Computer-Mediated Communication 3:0. doi: 10.1111/j.1083-6101.1998.tb00079.x

Car NJ (2013) The eReefs information architecture. 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1-6 December 2013, p. 831-837. http://www.mssanz.org.au.previewdns.com/modsim2013/C7/car2.pdf. Accessed 23 December 2015

Chen Y, Minchin SA, Seaton S, Joehnk KD, Robson BJ, Bai Q (2011) eReefs – a new perspective on the Great Barrier Reef. 19th International Congress on Modelling and Simulation, Perth, Australia, 12-16 December 2011, p. 1195-1201. http://www.mssanz.org.au/modsim2011/C4/chen.pdf. Accessed 23 December 2015

Cooke NJ, Hilton ML (eds) (2015) Enhancing the Effectiveness of Team Science. The National Academies Press, Washington, D.C. doi: 10.17226/19007

Di Stefano M, Fox P, Beaulieu S, Maffei A (2012) The integrated ecosystems assessment initiative - enabling the assessment of impacts on large marine ecosystems: informatics to the forefront of science-based decision support. 2012 ICES Annual Science Conference, Bergen, Norway. http://tw.rpi.edu/media/2012/10/08/cd52/ICES_2012.pdf. Accessed 23 December 2015

Di Stefano M, Fox P, Maffei A, West P, Hare J (2013) An open source approach to enable the reproducibility of scientific workflows in the ocean sciences. American Geophysical Union Fall Meeting, San Francisco, CA. http://tw.rpi.edu/media/2014/02/23/b139/AGU2013-IN51A-15330-MDS.pdf. Accessed 23 December 2015

Eaton JW, Bateman D, Hauberg S, Wehbring R (2014) GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent

Publishing Platform. ISBN 1441413006. http://www.gnu.org/software/octave/doc/interpreter/

Ecosystem Assessment Program (2009) Ecosystem Assessment Report for the Northeast U.S. Continental Shelf Large Marine Ecosystem. U.S. Department of Commerce, Northeast Fisheries Science Center Reference Document 09-11, 61 pp. http://www.nefsc.noaa.gov/publications/crd/crd0911/. Accessed 23 December 2015

Ecosystem Assessment Program (2012) Ecosystem Status Report for the Northeast Shelf Large Marine Ecosystem - 2011. U.S. Department of Commerce, Northeast Fisheries Science Center Reference Document 12-07, 32 pp. http://nefsc.noaa.gov/publications/crd/crd1207/. Accessed 23 December 2015

Fox P, Batchelder H, Lawrence S, Maffei A, Young O (2012) Information models for development and evolution of complex multi-scale knowledge networks for marine ecosystems. Ocean Sciences Meeting, Salt Lake City, UT. https://tw.rpi.edu//web/doc/OSC2012_139_ecoop_poster. Accessed 23 December 2015

Fox P, McGuinness DL (2008) TWC Semantic Web Methodology. http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology. Accessed 23 December 2015

Frederic J (2013) Nbconvert refactor. Final 1.0. http://digitalcommons.calpoly.edu/physsp/85. Accessed 17 October 2016.

Goble CA, Bhagat J, Aleksejevs S, et al. (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research 38:W677-W682. doi: 10.1093/nar/gkq429

GRASS Development Team (2015) Geographic Resources Analysis Support System (GRASS) Software, Version 7.0. Open Source Geospatial Foundation. http://grass.osgeo.org

Horsburgh JS (2015) Hydrology domain cyberinfrastructures: Successes, challenges, and opportunities. American Geophysical Union Fall Meeting, abstract #H42A-07. https://agu.confex.com/agu/fm15/meetingapp.cgi/Paper/66729. Accessed 15 April 2016

Howison J, Deelman E, McLennan MJ, Ferreira da Silva R, Herbsleb JD (2015) Understanding the scientific software ecosystem and its impact: Current and future measures. Research Evaluation 24:454-470. doi: 10.1093/reseval/rvv014

Hunter, JD (2007) Matplotlib: A 2D graphics environment. Computing in Science & Engineering 9:90-95. doi: 10.1109/MCSE.2007.55

iMarine (2014) Executive Summary: "iMarine data platform for collaborations" workshop, 7 March 2014, FAO, Rome, Italy. http://uripreview.i-marine.eu/be0c89a7-6eca-4ae1-ac87-9a52d8800641.pdf. Accessed 31 December 2015

Jirotka M, Lee CP, Olson GM (2013) Supporting scientific collaboration: Methods, tools, and concepts. Computer Supported Cooperative Work 22:667-715. doi: 10.1007/s10606-012-9184-0

Jones E, Oliphant E, Peterson P, et al. (2001) SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/. Accessed 18 October 2016.

Jupyter Team (2015) Jupyter Documentation. Kernels (Programming Languages). http://jupyter.readthedocs.io/en/latest/projects/kernels.html. Accessed 18 October 2016.

Levin PS, Fogarty MJ, Murawski SA, Fluharty D (2009) Integrated Ecosystem Assessments: Developing the scientific basis for ecosystem-based management of the ocean. PLoS Biology 7:e1000014. doi: 10.1371/journal.pbio.1000014

Levin PS, Kelble CR, Shuford RL, Ainsworth C, deReynier Y, Dunsmore R, Fogarty MJ, Holsman K, Howell EA, Monaco ME, Oakes SA, Werner F (2014) Guidance for implementation of Integrated Ecosystem Assessments: a US perspective. ICES J. Mar. Sci. 71:1198-1204. doi: 10.1093/icesjms/fst112

Liu J, Pacitti E, Valduriez P, Mattoso M (2015) A survey of data-intensive scientific workflow management. Journal of Grid Computing 13:457-493. doi: 10.1007/s10723-015-9329-8

Ma X, Beaulieu SE, Fu L, Fox P, Di Stefano M, West P (2017) Documenting provenance for reproducible marine ecosystem assessment in open science. In: Diviacco P, Leadbetter A, Glaves HM (eds) Oceanographic and marine cross-domain data management for sustainable development. IGI Global, Hershey, PA, pp. 100-126, doi: 10.4018/978-1-5225-0700-0.ch005

McKinney W (2010) Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference:51-56.

Muller MJ, Kuhn S (1993) Participatory design. Communications of the ACM 36:24-28. doi: 10.1145/153571.255960

Muste, M, Bennett, D, Secchi, S, Schnoor, J, Kusiak, A, Arnold, N, Mishra, S, Ding, D, Rapolu, U (2013) End-to-end cyberinfrastructure for decision-making support in watershed management. Journal of Water Resources Planning and Management 139:565-573. doi: 10.1061/(ASCE)WR.1943-5452.0000289

National Ocean Council (2013) National ocean policy implementation plan. https://www.whitehouse.gov//sites/default/files/national_ocean_policy_implementation_plan.pdf. Accessed 23 December 2015

NOAA (2014) National Oceanic and Atmospheric Administration Information Quality Guidelines. Issue date of this revision: 30 October 2014. http://www.cio.noaa.gov/services_programs/IQ_Guidelines_103014.html. Accessed 15 April 2016

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015) Promoting an open research culture. Science 348:1422-1425. doi: 10.1126/science.aab2374

Obama B (2013) Executive order -- Making open and machine readable the new default for government information. The White House, Office of the Press Secretary, May 09, 2013. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-. Accessed 23 December 2015

Olson JS, Hofer EC, Bos N, Zimmerman A, Olson GM, Cooney D, Faniel I (2008) A theory of remote scientific collaboration. In: Olson GM, Zimmerman A, Bos N (eds) Scientific collaboration on the internet. MIT Press, Cambridge, MA, pp 73-99

Palmer MA (2012) Socioenvironmental sustainability and actionable science. BioScience 62:5-6. doi: 10.1525/bio.2012.62.1.2

Pennington D (2011) Collaborative, cross-disciplinary learning and co-emergent innovation in eScience teams. Earth Science Informatics 4:55-68. doi: 10.1007/s12145-011-0077-4

Pérez F, Granger BE (2007) IPython: A system for interactive scientific computing. Computing in Science and Engineering 9:21-29. doi: 10.1109/MCSE.2007.53

Pulsifer PL, Collins JA, Kaufman M, Eicken H, Parsons MA, Gearheard S (2011) Applying agile methods to the development of a community-based sea ice observations database. American Geophysical Union Fall Meeting, abstract #IN54A-08. http://adsabs.harvard.edu/abs/2011AGUFMIN54A..08P. Accessed 15 April 2016

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/Ragan-Kelley M, Pérez F, Granger B, Kluyver T, Ivanov P, Frederic J, Bussonier M (2014) The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. American Geophysical Union Fall Meeting, abstract #H44D-07. http://adsabs.harvard.edu/abs/2014AGUFM.H44D..07R. Accessed 15 April 2016

Reichman OJ, Jones MB, Schildhauer MP (2011) Science challenges and opportunities of open data in ecology. Science 331:703-705. doi: 10.1126/science.1197962

Samhouri JF, Haupt AJ, Levin PS, Link JS, Shuford R (2014) Lessons learned from developing integrated ecosystem assessments to inform marine ecosystem-based management in the USA. ICES Journal of Marine Science 71:1205-1215. doi: 10.1093/icesjms/fst141

Samhouri JF, Lester SE, Selig ER, Halpern BS, Fogarty MJ, Longo C, McLeod KL (2012) Sea sick? Setting targets to assess ocean health and ecosystem services. Ecosphere 3:art41. doi: 10.1890/ES11-00366.1

Shen H (2014) Interactive notebooks: Sharing the code. The free IPython notebook makes data analysis easier to record, understand and reproduce. Nature 515:151-152. doi: 10.1038/515151a

Taconet M, Ellebroek A, Castelli D, Pagano P, Caumont H, Garavelli S, Parker S (2014) Sustaining iMarine: a public partnership led business model. The iMarine Sustainability White Paper, final release November 2014, 65 pp. ftp://ftp.fao.org/FI/DOCUMENT/FIGIS_FIRMS/2015/Inf11e.pdf. Accessed 23 December 2015

Tilmes C, Fox P, Ma X, McGuinness DL, Privette AP, Smith A, Waple A, Zednik S, Zheng JG (2013) Provenance representation for the national climate assessment in the global change information system. IEEE Transactions on Geoscience and Remote Sensing 51:5160-5168. doi: 10.1109/TGRS.2013.2262179

van den Hove S (2007) A rationale for science-policy interfaces. Futures 39:807-826. doi: 10.1016/j.futures.2006.12.004

Wilson G (2014) Software Carpentry: lessons learned. F1000Research 3:62, Version 1, 19 Feb 2014. doi: 10.12688/f1000research.3-62.v1

Wright DJ (2016) Toward a digital resilience. Elementa: Science of the Anthropocene 4:000082. doi: 10.12952/journal.elementa.000082

Figure captions

Fig. 1. Diagram of the Integrated Ecosystem Assessment (IEA) process, driven by the goals and targets of Ecosystem-Based Management (EBM; image available online at: http://www.noaa.gov/iea/loop.html).

Fig. 2. Schematic for data interoperability in the Ecosystem Status Report for the Northeast U.S. Shelf Large Marine Ecosystem. The data sources (lower layer), applications (middle layer, including a blank field for new tools), and the resulting integrated data products and indicators for the Report (upper layer) reflect the key elements in the use case. The two gray layers indicate mediation and the potential for semantic interoperability.

Fig. 3. Representative data products and indicators in the Ecosystem Status Report for the Northeast U.S. Shelf Large Marine Ecosystem. (a) Time-series indicator: Mean trophic level of landings by commercial fisheries [from Fig. 8.2 in Ecosystem Assessment Program (2012)]. (b) Spatial data product: Mean (1998-2010) daily primary production [from Fig. 4.2 in Ecosystem Assessment Program (2012)].

Fig. 4. Diagram of TWC Methodology, an iterative use case development methodology [modified from Fox and McGuinness (2008)].

Fig. 5. Activity diagram for the Ecosystem Status Report use case, indicating actors, entities (i.e., data files, image products, and the Report), and activities (arrows). Note the data retriever and processor is represented as a software agent (square head). The dashed box contains the activities for which we built the prototype.

Fig. 6. Screen grab of a portion of the executed Climate Forcing Notebook, showing: opening a document, importing text files, accessing a source data file, processing data, and plotting and

saving derived data products (to view details, please refer to the notebook at the GitHub repository accessible via https://data.rpi.edu/xmlui/handle/10833/1756).

Fig. 7. Screen grab of the PDF document that results from the executed Climate Forcing Notebook (to view details, please refer to the PDF at the GitHub repository accessible via https://data.rpi.edu/xmlui/handle/10833/1756).

(Figures submitted separately)

Fig. 1.

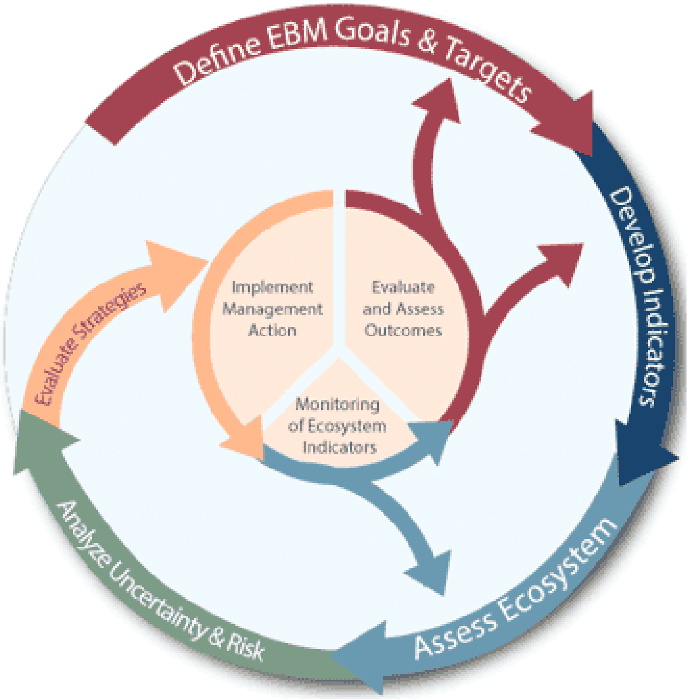

Define EBM Goals & Targets
Develop Indicators
Assess Ecosystem
Analyze Uncertainty & Risk
Evaluate Strategies
Implement Management Action
Evaluate and Assess Outcomes
Monitoring of Ecosystem Indicators

Fig. 2.

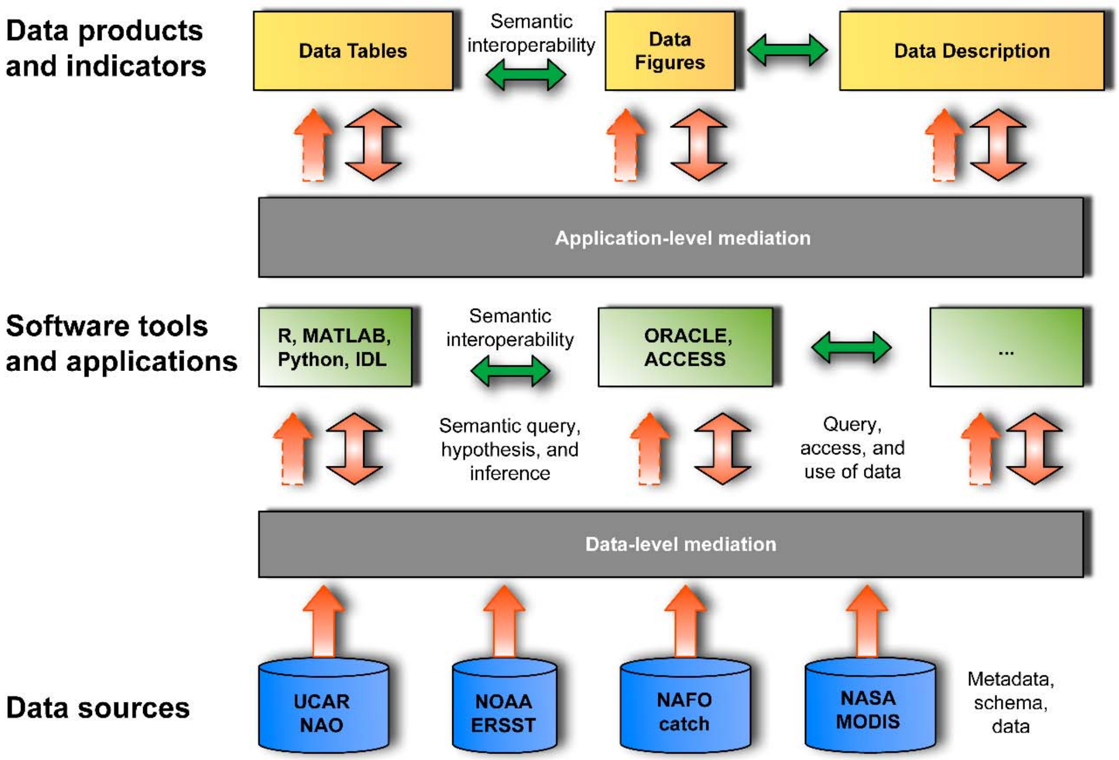

Data products and indicators
Data Tables
Semantic interoperability
Data Figures
Data Description
Application-level mediation
Software tools and applications
R, MATLAB, Python, IDL
Semantic interoperability
ORACLE, ACCESS
...
Semantic query, hypothesis, and inference
Query, access, and use of data
Data-level mediation
Data sources
UCAR NAO
NOAA ERSST
NAFO catch
NASA MODIS
Metadata, schema, data

Fig. 3. (a) (b)

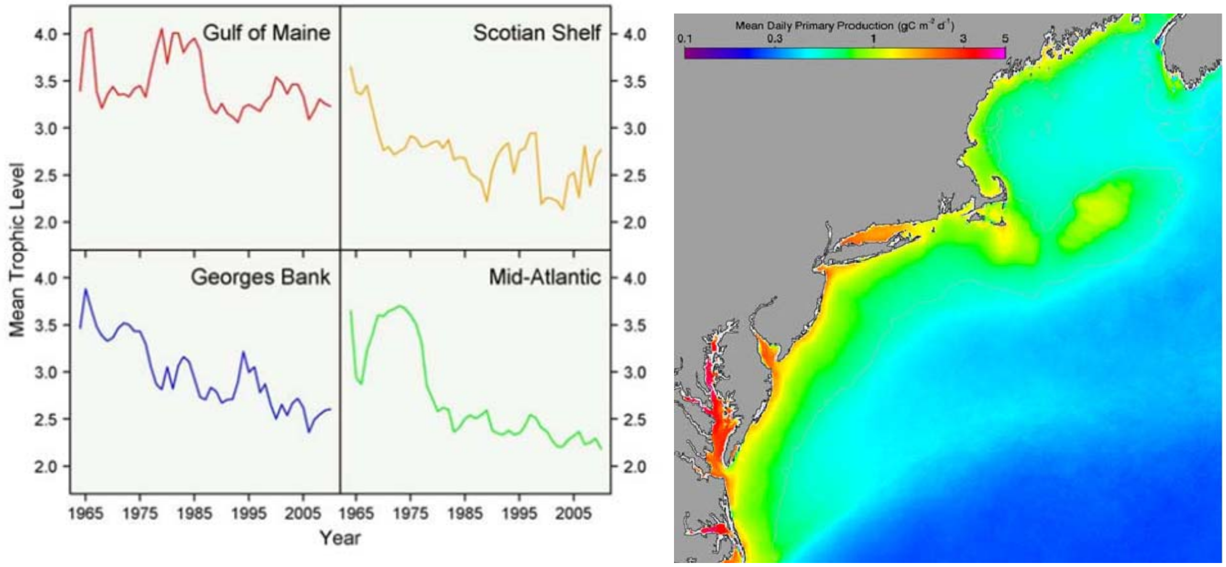

Gulf of Maine
Scotian Shelf
Georges Bank
Mid-Atlantic
Mean Trophic Level
Year
4.0
3.5
3.0
2.5
2.0
1965
1975
1985
1995
2005
Mean Daily Primary Production (gC m-2 d-1)
0.1
0.3
1
3
5

Fig. 4.

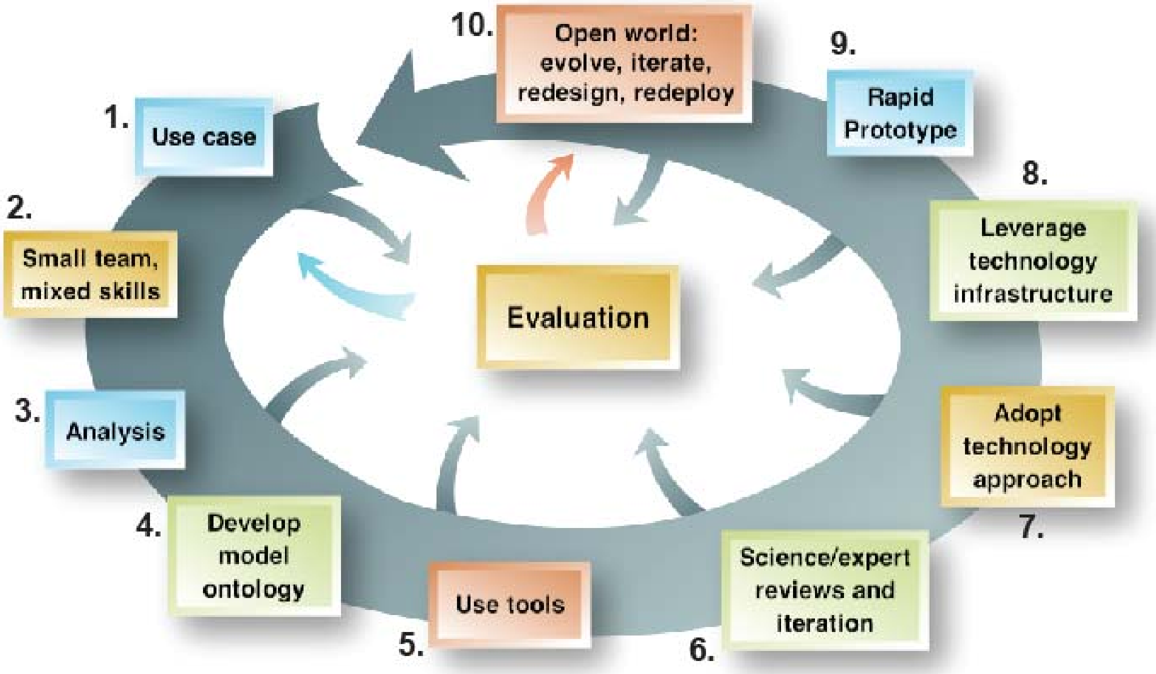

10. Open world: evolve, iterate, redesign, redeploy
9. Rapid Prototype
1. Use case
2. Small team, mixed skills
8. Leverage technology infrastructure
Evaluation
3. Analysis
7. Adopt technology approach
4. Develop model ontology
5. Use tools
6. Science/expert reviews and iteration

Fig. 5.

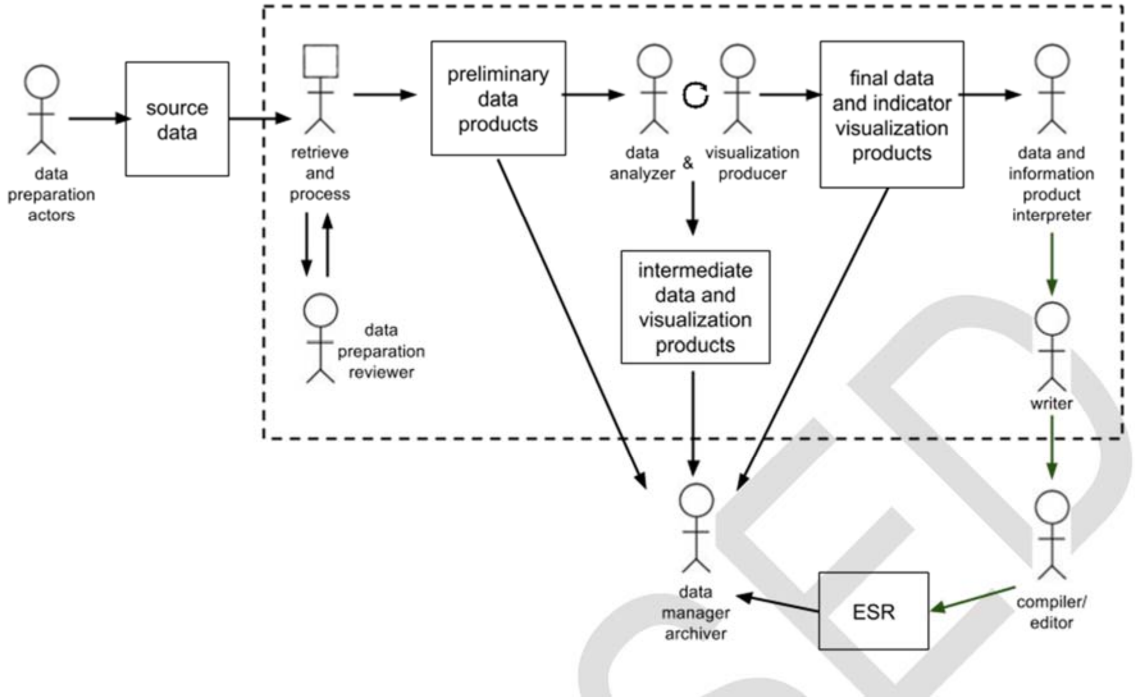
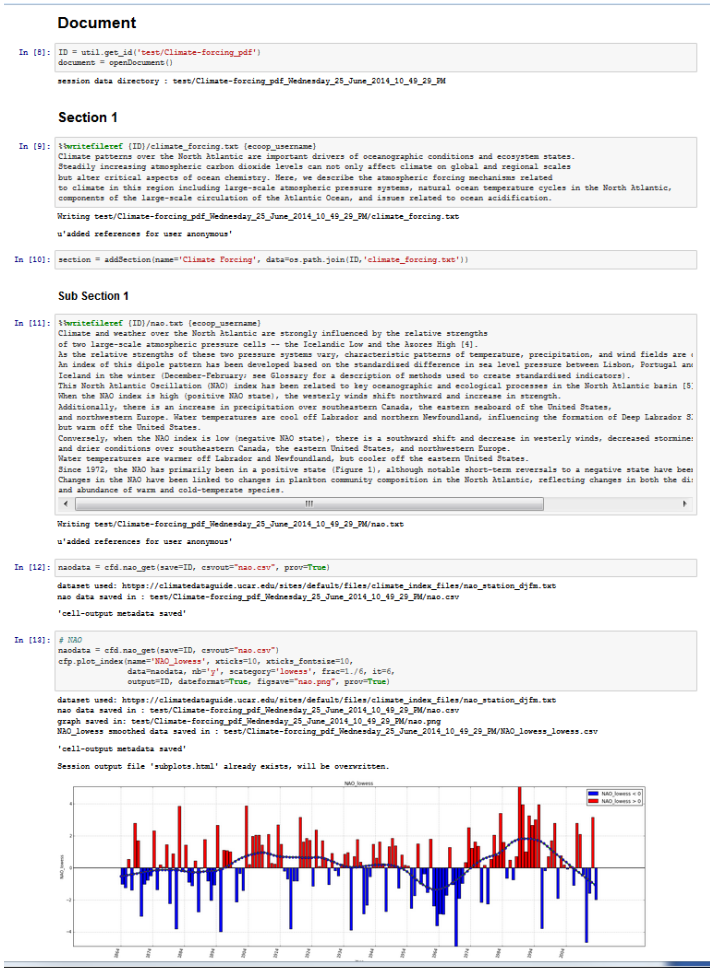

data preparation actors
source data
retrieve and process
data preparation reviewer
preliminary data products
data analyzer
&
visualization producer
intermediate data and visualization products
final data and indicator visualization products
data and information product interpreter
writer
compiler/ editor
ESR
data manager archiver

Fig. 6.



Document

```
In [8]: ID = util.get_id('test/Climate-forcing_pdf')
        document = openDocument()

        session data directory : test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM
```

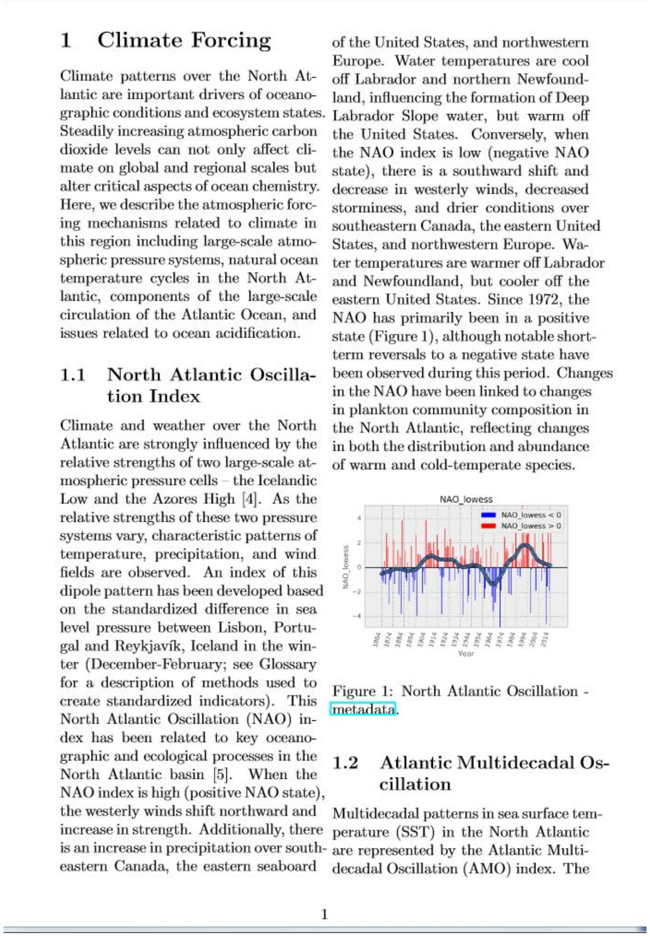
Section 1

```
In [9]: %%writefileref {ID}/climate_forcing.txt {ecoop_username}
        Climate patterns over the North Atlantic are important drivers of oceanographic conditions and ecosystem states.
        Steadily increasing atmospheric carbon dioxide levels can not only affect climate on global and regional scales
        but alter critical aspects of ocean chemistry. Here, we describe the atmospheric forcing mechanisms related
        to climate in this region including large-scale atmospheric pressure systems, natural ocean temperature cycles in the North Atlantic,
        components of the large-scale circulation of the Atlantic Ocean, and issues related to ocean acidification.

        Writing test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/climate_forcing.txt

        u'added references for user anonymous'

In [10]: section = addSection(name='Climate Forcing', data=os.path.join(ID,'climate_forcing.txt'))
```

Sub Section 1

```
In [11]: %%writefileref {ID}/nao.txt {ecoop_username}
         Climate and weather over the North Atlantic are strongly influenced by the relative strengths
         of two large-scale atmospheric pressure cells -- the Icelandic Low and the Azores High [4].
         As the relative strengths of these two pressure systems vary, characteristic patterns of temperature, precipitation, and wind fields are (
         An index of this dipole pattern has been developed based on the standardized difference in sea level pressure between Lisbon, Portugal and
         Iceland in the winter (December-February; see Glossary for a description of methods used to create standardized indicators).
         This North Atlantic Oscillation (NAO) index has been related to key oceanographic and ecological processes in the North Atlantic basin [5]
         When the NAO index is high (positive NAO state), the westerly winds shift northward and increase in strength.
         Additionally, there is an increase in precipitation over southeastern Canada, the eastern seaboard of the United States,
         and northwestern Europe. Water temperatures are cool off Labrador and northern Newfoundland, influencing the formation of Deep Labrador S
         but warm off the United States.
         Conversely, when the NAO index is low (negative NAO state), there is a southward shift and decrease in westerly winds, decreased stormine
         and drier conditions over southeastern Canada, the eastern United States, and northwestern Europe.
         Water temperatures are warmer off Labrador and Newfoundland, but cooler off the eastern United States.
         Since 1972, the NAO has primarily been in a positive state (Figure 1), although notable short-term reversals to a negative state have bee
         Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the di
         and abundance of warm and cold-temperate species.

         Writing test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/nao.txt

         u'added references for user anonymous'

In [12]: naodata = cfd.nao_get(save=ID, csvout="nao.csv", prov=True)

         dataset used: https://climatedataguide.ucar.edu/sites/default/files/climate_index_files/nao_station_djfm.txt
         nao data saved in : test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/nao.csv

         'cell-output metadata saved'

In [13]: # NAO
         naodata = cfd.nao_get(save=ID, csvout="nao.csv")
         cfp.plot_index(name='NAO_lowess', xticks=10, xticks_fontsize=10,
                        data=naodata, nb='y', scategory='lowess', frac=1./6, it=6,
                        output=ID, dateformat=True, figsave="nao.png", prov=True)

         dataset used: https://climatedataguide.ucar.edu/sites/default/files/climate_index_files/nao_station_djfm.txt
         nao data saved in : test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/nao.csv
         graph saved in: test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/nao.png
         NAO_lowess smoothed data saved in : test/Climate-forcing_pdf_Wednesday_25_June_2014_10_49_29_PM/NAO_lowess_lowess.csv

         'cell-output metadata saved'

         Session output file 'subplots.html' already exists, will be overwritten.
```

Fig. 7.



1 Climate Forcing

Climate patterns over the North Atlantic are important drivers of oceanographic conditions and ecosystem states. Steadily increasing atmospheric carbon dioxide levels can not only affect climate on global and regional scales but alter critical aspects of ocean chemistry. Here, we describe the atmospheric forcing mechanisms related to climate in this region including large-scale atmospheric pressure systems, natural ocean temperature cycles in the North Atlantic, components of the large-scale circulation of the Atlantic Ocean, and issues related to ocean acidification.

1.1 North Atlantic Oscillation Index

Climate and weather over the North Atlantic are strongly influenced by the relative strengths of two large-scale atmospheric pressure cells – the Icelandic Low and the Azores High [4]. As the relative strengths of these two pressure systems vary, characteristic patterns of temperature, precipitation, and wind fields are observed. An index of this dipole pattern has been developed based on the standardized difference in sea level pressure between Lisbon, Portugal and Reykjavík, Iceland in the winter (December-February; see Glossary for a description of methods used to create standardized indicators). This North Atlantic Oscillation (NAO) index has been related to key oceanographic and ecological processes in the North Atlantic basin [5]. When the NAO index is high (positive NAO state), the westerly winds shift northward and increase in strength. Additionally, there is an increase in precipitation over southeastern Canada, the eastern seaboard of the United States, and northwestern Europe. Water temperatures are cool off Labrador and northern Newfoundland, influencing the formation of Deep Labrador Slope water, but warm off the United States. Conversely, when the NAO index is low (negative NAO state), there is a southward shift and decrease in westerly winds, decreased storminess, and drier conditions over southeastern Canada, the eastern United States, and northwestern Europe. Water temperatures are warmer off Labrador and Newfoundland, but cooler off the eastern United States. Since 1972, the NAO has primarily been in a positive state (Figure 1), although notable short-term reversals to a negative state have been observed during this period. Changes in the NAO have been linked to changes in plankton community composition in the North Atlantic, reflecting changes in both the distribution and abundance of warm and cold-temperate species.

Figure 1: North Atlantic Oscillation - metadata.

1.2 Atlantic Multidecadal Oscillation

Multidecadal patterns in sea surface temperature (SST) in the North Atlantic are represented by the Atlantic Multidecadal Oscillation (AMO) index. The