INFORMATICS SOLUTIONS FOR LARGE OCEAN OPTICS DATASETS

Heidi M. Sosik
Biology Department, Woods Hole Oceanographic Institution
Woods Hole, MA 02543-1049, USA
hsosik@whoi.edu

Joe Futrelle
Applied Ocean Physics and Engineering Department, Woods Hole Oceanographic Institution
Woods Hole, MA 02543-1049, USA
jfutrelle@whoi.edu

ABSTRACT

Lack of observations that span the wide range of critical space and time scales continues to limit many aspects of oceanography. As ocean observatories and observing networks mature, the role for optical technologies and approaches in helping to overcome this limitation continues to grow. As a result the quantity and complexity of data produced is increasing at a pace that threatens to overwhelm the capacity of individual researchers who must cope with large high-resolution datasets, complex, multi-stage analyses, and the challenges of preserving sufficient metadata and provenance information to ensure reproducibility and avoid costly reprocessing or data loss. We have developed approaches to address these new challenges in the context of a case study involving very large numbers (~1 billion) of images collected at coastal observatories by Imaging FlowCytobot, an automated submersible flow cytometer that produces high resolution images of plankton and other microscopic particles at rates up to 10 Hz for months to years. By developing partnerships amongst oceanographers generating and using such data and computer scientists focused on improving science outcomes, we have prototyped a replicable system. It provides simple and ubiquitous access to observational data and products via web services in standard formats; accelerates image processing by enabling algorithms developed with desktop applications to be rapidly deployed and evaluated on shared, high-performance servers; and improves data integrity by replacing error-prone manual data management processes with generalized, automated services. The informatics system is currently in operation for multiple Imaging FlowCytobot datasets and being tested with other types of ocean imagery.

INTRODUCTION

New optical technologies promise to fill a critical gap in observing relevant space and time scales in the ocean, but they also raise challenges that must be met to effectively use the observations to advance understanding. As ocean observatory infrastructure matures, opportunities grow for extended deployments of new sensors that produce data of unprecedented quantity and complexity. This is especially true for automated optical imaging systems that characterize individual organisms and produce data at a pace that threatens to overwhelm the capacity of individual researchers who must cope with large high-resolution datasets, complex, multi-stage analyses, and the challenges of preserving sufficient metadata and provenance information to ensure reproducibility and avoid costly reprocessing or data loss. We have developed approaches to address these new challenges in the context of a case study involving very large numbers (~1 billion) of images collected at coastal observatories by Imaging FlowCytobot (IFCB) (Olson and Sosik 2007; Sosik and Olson 2007; Sosik et al. 2011), an automated submersible flow cytometer that produces high resolution images of plankton

and other microscopic particles at rates up to 10 Hz for months to years (Figure 1). With appropriate information systems in place, these datasets have the potential to provide novel insights into coastal ecosystem dynamics, including characterization of biological responses to environmental change and timely early warning of harmful algal bloom events (Campbell et al. 2010).

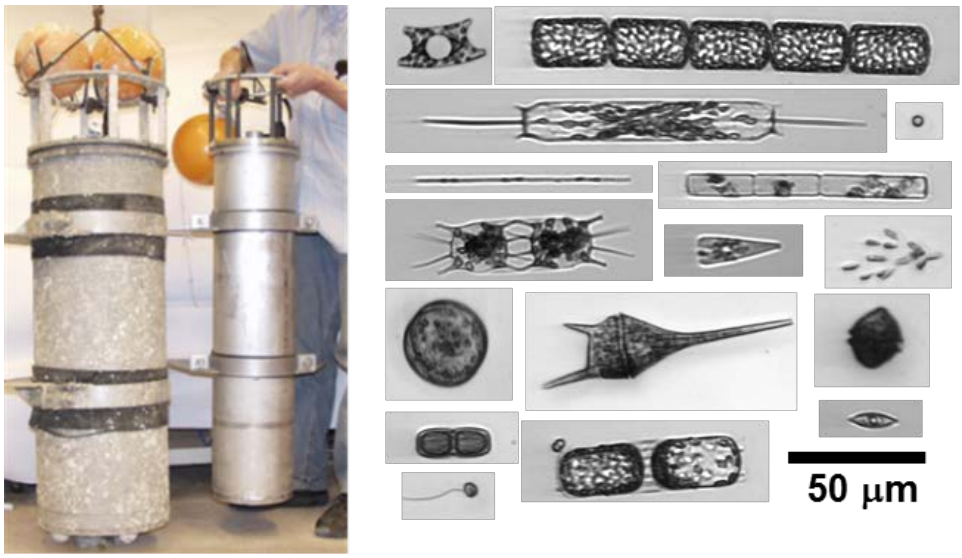

Figure. 1. Imaging FlowCytobot ready for deployment in the ocean (left), along with plankton images collected by Imaging FlowCytobot at the Martha's Vineyard Coastal Observatory[1]. The research prototype (instrument on left) collected the time series observations used in this study and the commercial design (right instrument) is now available from McLane Research Laboratories, Inc[2]. The instrument system automatically acquires 1000s of high resolution images of microscopic plankton every hour and can operate for >6 months unattended.

APPROACH

We have adopted an interdisciplinary strategy that has enabled rapid development of solutions with high scientific impact. In sharp contrast with the traditional focus on attempting to adopt the latest innovations designed by computer scientists, we have instead emphasized building partnerships amongst oceanographers generating and using data and computer scientists motivated to improve specific science outcomes. The role of these interdisciplinary teams has been to identify, design, implement, and evaluate replicable methodologies and technical information systems, without predisposition towards any particular technology or novelty of components. In fact, we have found that in many cases, rapid advances have not required technological innovations, but rather effective communication, focus on science outcomes, and an iterative design and evaluation process. In this work we have modeled methodology developed in the Tetherless World Constellation[3] at Rensselaer Polytechnic Institute, and which has been previously applied to other data-intensive earth science applications (Rozell et al. 2010; Zheng et al. 2011).

In the design process, initial steps involved developing a set of use cases describing current and proposed applications of IFCB and IFCB data, and analyzing those use cases in terms of the

[1] http://www.whoi.edu/mvco
[2] http://www.mclanelabs.com/
[3] http://tw.rpi.edu/

underlying abstract information model and the set of human and software actors engaged in the activity. The information models, activity diagrams, and other formal descriptions produced in this phase served as common reference points helping establish where in the system we could make high-impact improvements (Figure 2). Some of these "pain points", such as difficulty distributing software and tracking changes to it, were addressed by adopting standard software engineering practices such as shared issue tracking and revision control systems. Others, such as the inadequate performance of retrospective processing, required development of new software and the use of new hardware. In early prototyping, web services were developed that provided remote access to IFCB data, including standard variants of the non-standard data formats produced by IFCB. For example, IFCB images are not stored as image files but rather as uncompressed pixel data that is separate from the headers and other meta-information required to reconstruct the images. The prototype web services produced standard image variants on demand and so provided stable URLs from which all IFCB images could be retrieved. This proved to be such a useful way of accessing IFCB data that it led to the development of a comprehensive set of web services providing access to associated metadata and even the raw data, in a variety of standard formats such as CSV, XML, RDF, RSS, and Zip. The web services enabled improvements in near-real-time processing as incoming data could be syndicated to remote processing machines using RSS. They also enabled the development of a web-based dashboard providing a convenient user interface for browsing and linking to IFCB data (http://ifcb-data.whoi.edu/). From being accessible only using custom software on an internal network, IFCB data has now become a globally-accessible data resource.

Our technical approach is motivated by the observation from many previous science informatics efforts that reformatting, curating, and migrating data to centralized repositories is prohibitively expensive for many small research projects, especially research projects with substantial and constantly-growing data holdings. As a result, solutions are needed that place a minimal technical burden on scientists while providing access to data in as ubiquitous and timely a fashion as possible. World Wide Web standards such as HTTP, XML, RSS, and JSON provide a level of standardization and interoperability that is unmatched by any other informatics technologies; they interoperate with desktop operating systems, high performance computing systems, mobile devices, and social networks. They also decentralize access to data, which solves the problem that centralized repositories are not large or responsive enough to provide near-real-time access to instrument data from the growing global environmental observatory infrastructure. And finally the use of URLs to identify data items "at birth" provides global identification and the ability to cite data in literature and other communications without having to undertake any additional curation work or deposit data in a central repository.

DISCUSSION

The prototype replicable information system we have produced for Imaging FlowCytobot data provides simple and ubiquitous access to observational data and products via web services in standard formats; accelerates image processing by enabling algorithms developed with desktop applications to be rapidly deployed and evaluated on shared, high-performance servers; and improves data integrity by replacing error-prone manual data management processes with generalized, automated services. The system includes a web-based dashboard (http://ifcb-data.whoi.edu/) that provides near-real-time browsing and access to images (Figure 3) and image products (Figure 4) from anywhere and on a wide variety of devices. It also provides remote access to IFCB images, data, and metadata so that they can be processed on large, multi-CPU clusters, which has enabled reprocessing existing data with improved algorithms in a fraction of the time (weeks instead of years) that would have been required using previous approaches. A consistent URL scheme and public web service means that it is no longer

necessary to know details about local storage or have access to internal networks or custom software. Links to image data can now be cited in publications and other documentation, emailed, posted to social networks, etc., and used to provide both browsing and download of individual images or entire datasets. A key collaboration strategy has been to enable these new capabilities without disrupting current practices (e.g., no requirements to change native data format or rewrite analysis software); new solutions are provided as an augmentation rather than a replacement of existing, working systems. The new data system is currently in use by multiple researchers deploying Imaging FlowCytobots and providing input that is informing continued development, and elements also being tested with other types of ocean imagery.

ACKNOWLEDGMENTS

This research was supported by grants from the Gordon and Betty Moore Foundation, NSF, NASA, and ONR (NOPP). We are indebted to Rob Olson, Alexi Shalapynok, Taylor Crockford and Emily Peacock for expert assistance in the lab and field and to the Martha's Vineyard Coastal Observatory operations team (J. Fredericks, H. Popenoe, and J. Sisson) for logistical support that has made time series observations possible. We also thank other members of the Ocean Imaging Informatics team at WHOI for contributions to the creative process around conceiving and adopting novel informatics solutions.

REFERENCES

Campbell, L., R. J. Olson, H. M. Sosik, A. Abraham, D. W. Henrichs, C. J. Hyatt, and E. J. Buskey. 2010. First harmful *Dinophysis* (Dinophyceae, Dinophysiales) bloom in the U.S. is revealed by automated imaging flow cytometry. J. Phycol. **46:** 66-75

Olson, R. J., and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. Limnol. Oceanogr. Methods **5:** 195-203

Rozell, E., A. Maffei, S. Beaulieu, and P. Fox. 2010. A framework for integrating oceanographic data repositories. 2010 Fall Meeting, AGU, San Francisco, Calif., 13-17 Dec http://tw.rpi.edu/media/2011/07/13/10a4e/S2S-AGU-Poster-Small.pdf

Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr. Methods **5:** 204-216

Sosik, H. M., R. J. Olson, and E. V. Armbrust. 2011. Flow cytometry in phytoplankton research, p. 171-185. *In* D. J. Suggett, O. Prasil and M. A. Borowitzka [eds.], Chlorophyll a fluorescence in aquatic sciences: Methods and applications. Developments in Applied Phycology 4. Springer

Zheng, J., P. Wang, E. W. Patton, T. Lebo, J. S. Luciano, and D. L. McGuinness. 2011. A semantically-enabled provenance-aware water quality portal. *In* M. B. Jones and C. Gries [eds.], Proceedings of the Environmental Information Management Conference 2011 (EIM 2011). University of California.doi:10.5060/D2NC5Z4X. http://tw.rpi.edu/media/2011/07/14/41a/EIM2011-SWQP_V1.6.pdf ; https://eim.ecoinformatics.org/eim2011/eim-proceedings-2011/

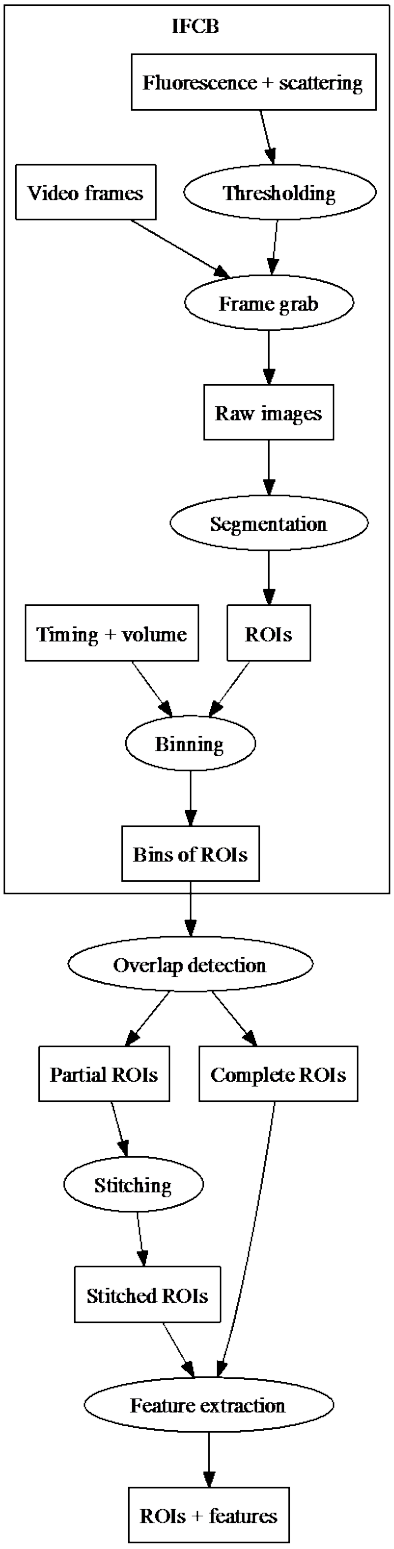

Figure 2: IFCB image processing diagram, produced during initial prototyping of web services. Regions of Interest (ROIs) correspond to the cell or other target present in each image.

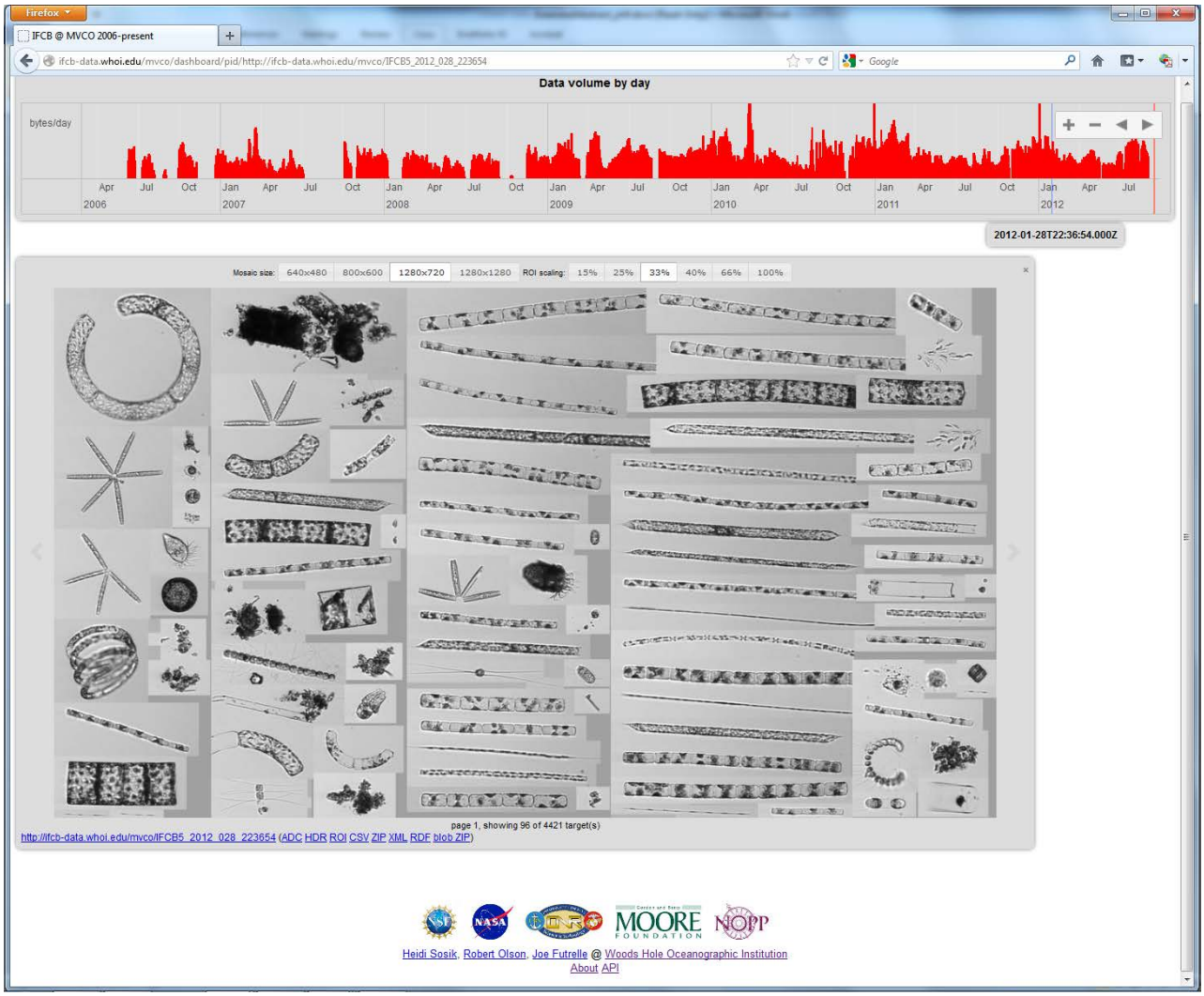

Figure 3. Snap shot of the web-based IFCB data dashboard (http://ifcb-data.whoi.edu/) showing the time series navigation tool (top) and a mosaic of images (phytoplankton, microzooplankton, and detritus) from a single time series sample selected by the user. The dashboard provides links to URLs accessing results ranging from each individual full resolution image (including metadata), to all images in a sample bin (zip compressed), or metadata for an entire sample bin (in various formats, e.g., HTML, XML, RDF). This specific sample can be viewed in the dashboard at http://ifcb-data.whoi.edu/mvco/dashboard/pid/http://ifcb-data.whoi.edu/mvco/IFCB5_2012_028_223654

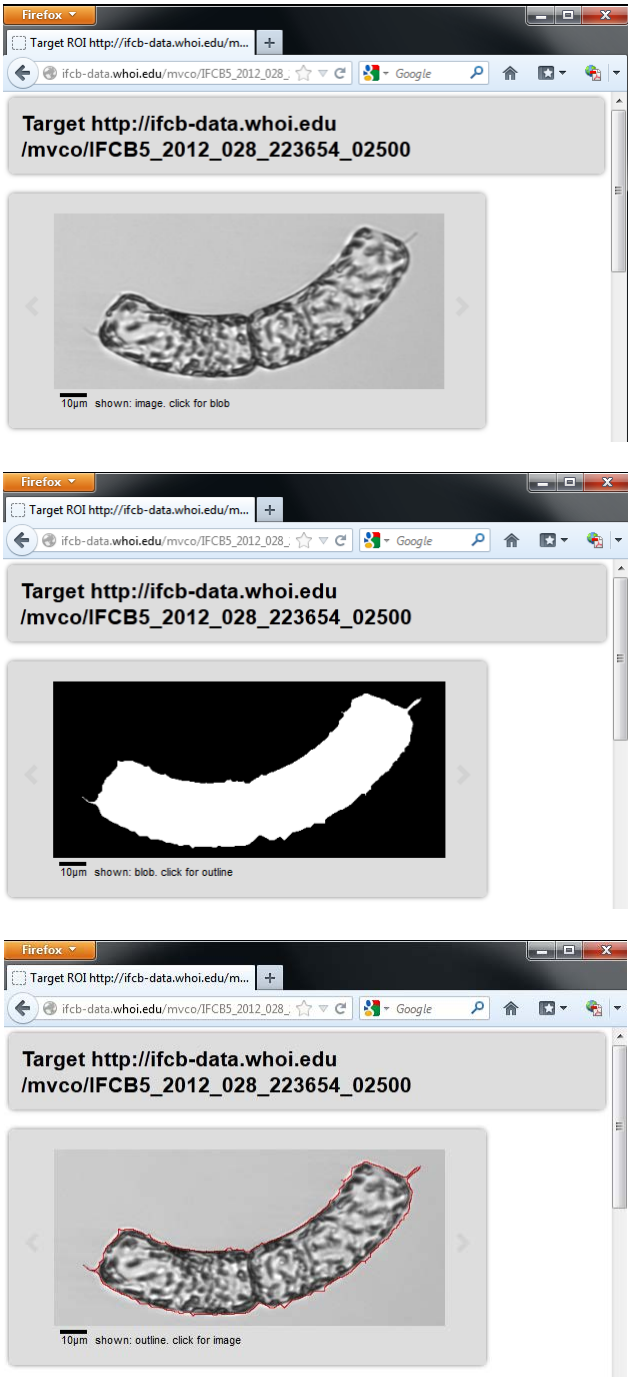

Figure 4. Alternate views from the data dashboard showing a selected individual image (top) and some products after early stages of image processing: "blob" image or bitmap mask separating the target from its background (middle), and blob perimeter superimposed on the original image (bottom). Any image in the data set can be viewed in this way, and web services provide open access to both the images and the products. This specific image series can be found at http://ifcb-data.whoi.edu/mvco/IFCB5_2012_028_223654_02500.html (where clicking on the image allows toggling between the original image and product views).