Title: "A Model for Bioinformatics Training: The Marine Biological Laboratory"

Grant Yamashita, ASU Center for Biology and Society
Holly Miller, MBLWHOI Library, MBL
Anthony Goddard, MBLWHOI Library, MBL
Cathy Norton, MBLWHOI Library, MBL

Abstract:

Many areas of science such as biology, medicine, and oceanography are becoming increasingly data-rich and most programs that train scientists do not address informatics techniques or technologies that are necessary for managing and analyzing large amounts of data. Educational resources for scientists in informatics are scarce, yet scientists need the skills and knowledge to work with informaticians and manage graduate students and post-docs in informatics projects. The Marine Biological Laboratory houses a world-renowned library and is involved in a number of informatics projects in the sciences. The MBL has been home to the National Library of Medicine's BioMedical Informatics Course for nearly two decades and is committed to educating scientists and other scholars in informatics. In an innovative, immersive learning experience, Grant Yamashita, a biologist and post-doc at Arizona State University, visited the Science Informatics Group at MBL to learn first hand how informatics is done and how informatics teams work. Hands-on work with developers, systems administrators, librarians, and other scientists provided an invaluable education in informatics and is a model for future science informatics training.

Keywords: MBL, digital HPS, Embryo Project, MBLWHOI Library, BioMedical Informatics course, informatics boot camp, informatics education, Science Informatics Group

Introduction:

The amount of data produced in the sciences has increased exponentially in the last decade and scientists have not been trained to deal with this information deluge. In biology this includes the ever-growing database of gene and protein sequences as well as the never-ending accumulation of papers, books, and other forms of data. Tools to harvest, parse, and display this information in easily digested forms will not only make work easier to manage, but also lead to new discoveries that hold the promise of solving important social problems [1]. Informatics, which includes recording, analyzing, retrieving, and disseminating knowledge and information, plays a crucial role in today's practice of science. New tools to mine data, interpret it, and make it available to

others have accumulated at a rapid pace, but many of these tools are highly specialized and require integration with other sets of tools. The resulting load of complexity makes it increasingly difficult for scientists and other scholars to keep up with advancements in the field. In short, the rate of growth of informatics tools and applications has outpaced the availability of educational resources in informatics for scientists [2]. Laboratories are increasingly bringing in graduate students and post-doctoral fellows with computer science backgrounds to help manage this flood of data. Indeed, the boundary between science and informatics has blurred, which has led to the ineluctable reality that scientists need to also be trained in informatics to know how to lead a team of students and fellows in these data-driven projects.

Here we introduce a novel way of training biologists and other scholars in informatics that simultaneously addresses the problem of gaining expertise in cutting-edge techniques while learning about how data-centric projects, teams, and collaborations are built. We summarize a recent research stint at the Marine Biological Laboratory (MBL) by a visiting biologist to show the breadth of informatics training and education offered and how this can serve as a model for other visiting scientists and scholars. We specify the goals of the visit and how topics like database structure, repositories, web application development, IT support, agile development, machine-learning and text mining, leading teams, collaborations, and thinking about longevity/ sustainability/archiving were addressed. Finally, we outline the construction of a "boot camp" course in informatics modeled after the National Library of Medicine's (NLM) Medical Informatics Course and end with an invitation for interested scientists and scholars to visit the MBL.

The Model:

The Marine Biological Laboratory (MBL) is located along the southern shores of Cape Cod in Woods Hole, MA. In existence since 1888, it is home to world-renowned research and rigorous advanced courses that span the summer months. The MBL has also hosted more than fifty scientists who eventually won the Nobel Prize. In 1956 alone there were four faculty members in the Physiology course that later went on to win this prize. The MBL also houses a top research library with more than 250,000 volumes, both physical and electronic, that have been collected at the "research level" in science with concentrations in marine and oceanographic disciplines (http://www.mblwhoilibrary.org/index.php). In the last decade the library has become a leader in science informatics, exemplified by innovative taxonomically intelligent applications and services like uBio (Universal Biological Indexer and Organizer; http://ubio.org), and is a founding member of the Biodiversity Heritage Library (BHL, http://

www.biodiversitylibrary.org/) and the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC, http://www.iamslic.org/). The Biodiversity Informatics Group (BIG) of the Encyclopedia of Life (EOL, http://eol.org) is also located at the MBL. The goal of EOL is to provide information about all the world's organisms. In order to move towards this goal, the multi-institutional EOL project has developed tools to manage taxonomic names (e.g., TaxonFinder - http://code.google.com/p/taxon-finder/ and taxon name processing - http://code.google.com/p/taxon-name-processing/) and is beginning to explore semantic architectures.

The BioMedical Informatics Course (http://courses.mbl.edu/mi/) is held twice a year at the MBL and is sponsored by the National Library of Medicine. It is a prestigious course that has been running continuously since 1992. NLM Fellows who are accepted into this course are uniquely situated in the medical community to become "agents of change" in their respective institutions [3-4]. The course runs for one week, Sunday night through Saturday afternoon. The group of medical librarians, physicians, and other healthcare professionals experience lectures and classroom exercises during the day and then work in groups on an informatics project in the evenings. On Saturday morning each group presents the results of their project to the rest of the students and the faculty. Faculty in this course are leaders in the field and come from NIH, NLM, Vanderbilt, Columbia, Stanford, Harvard and all manage large scale informatics enterprises. Students interact with faculty not only during the give and take atmosphere of the lectures, but also in the informal sessions and in the dining hall where collaboration often begins. Topics in the 2010 spring course include medical ontology systems, electronic health records, database design, decision analysis methods, disaster informatics, and a peek at the latest technologies and tools being developed in medical informatics. Typically the student project involves the use of the Drupal content management system (http://drupal.org/) to generate a web portal to create a data model, map health-related data, and create consumer health information [5-8].

Recently, the Science Informatics Group (SIG) of the MBLWHOI Library hosted Grant Yamashita, a post-doctoral fellow from the Center for Biology and Society at Arizona State University, which develops the NSF-funded Embryo Project (EP, http://embryo.asu.edu). The MBL is a partner in the EP and the close ties between the institutions has resulted in a number of collaborations. The EP's aims are ambitious - "to document everything related to embryos through all of time, from Aristotle to tomorrow." [9-10]. As the EP progressed it quickly became clear that to further the goals of the project, informatics approaches and techniques would be needed. Specifically, there were questions about how to analyze and manage large amounts of textual data, how to represent and store this data in ways that would make it easily accessible and useable by others, and how to mine embryo-related data in other repositories. In short, the

project required text mining and semantic web tools as well as expertise in library archival methods. Could, for example, the data in the EP be represented via RDF (http://www.w3.org/RDF/) and then shared across similar projects in the history and philosophy of science? And, could this be done programatically via the analysis of largely unstructured texts, many of which have OCR errors or are in other languages.

Additionally, the EP is part of a growing coalition of digital projects in the history and philosophy of science ("digital HPS", http://www.digitalhps.org) that is intent on mapping out a digital infrastructure to bring about new methods and practices in informatics and social science computing to the various fields. Science studies scholars have always emphasized complex explanations of historical events, which are mostly presented as historical and richly contextual narratives and are thus always the end result of years of individual scholarship. What the science studies community has not yet embraced are the enormous benefits of the informatics revolution that has transformed the life sciences with respect to the organization of multiple forms of complex data, shared access to these data, searches in distributed relational databases that are organized around standardized practices of database management, and the possibilities of digital workbenches for collaborative and distributed research. All these developments have also contributed to a robust cyber-infrastructure that has changed the ways biologists go about their research [11].

The science studies community is thus missing out on new ways to conduct and organize research and to store, distribute, and analyze data. One of the main consequences of the incorporation of informatics techniques in biology has been the possibility of large-scale and comparative analysis of data and the integration of detailed experimental research with readily available points of comparison. This strategy has facilitated a bottom-up approach that allows biologists to find patterns of increasing generality. Insofar as one goal of the science studies community is to better understand both individual sciences as well as science at large in its various contexts (technological, theoretical, historical, social or political), it too will have to move beyond the particular and focus on general patterns wherever these do exist, a goal greatly facilitated by various informatics tools that allow large-scale analyses of data.

To better deal with specific problems in the EP and also to facilitate the growth of informatics in the history and philosophy of science communities, Yamashita spent three months under an NSF Professional Development Fellowship working with the Science Informatics Group (SIG). By training with an informatics team in a research environment that promotes informatics collaborations between different projects, the goal was to learn not only technical skills such as ontology development [12], database management, and semantic web application design, but

also how to effectively interact with other projects.

Results:

While the approaches and areas of emphasis would be customized for individuals who come through this program we identified core areas of informatics that Yamashita would learn during his time with the SIG. This included

- database structure
- repositories
- web application development
- IT support
- agile development
- machine-learning and text mining
- leading teams
- collaborations
- thinking about longevity/sustainability/archiving

In addition to reading relevant books and articles on the above topics, much of the training occurred via organic interactions with team members while working on real projects, attending talks, building servers, and participating in group functions like meetings and colloquia. The value of a simple act like walking over to an adjacent office to discuss the merits of Ruby on Rails as a development platform with one of the team members cannot be overestimated. There were many Embryo Project developments that provided ample space in which to experiment, but Yamashita was also asked to participate in and contribute to other SIG projects. By spending time embedded within an informatics team, Yamashita was exposed to a wide variety of relevant experiences. From a systems administration standpoint, this experience included virtualization, bare-metal hardware and server room design, resource and capacity planning, and networking principles. The value of these types of skills increases as informatics becomes a larger part of a typical research grant, where knowing the correct informatics resources to plan for can lead to both time and cost savings.

In a software development capacity, Yamashita learned software design and architecture principals, programming languages, database design, and project management techniques. Pair programming, where a researcher and a developer both work on the same code simultaneously has been shown to enable developers to learn much faster than when using traditional teaching

methods and is a unique aspect of working physically located within a development team [13].

Working directly within an established informatics team, then, exposed Yamashita to areas and challenges that would otherwise risk being overlooked in traditional professional development. This affords a clearly defined view of the roles that comprise an informatics team. The mix of technical and logistical knowledge gained provided Yamashita with a comprehensive picture of how informatics tools can support research and assist in planning informatics aspects in the EP and other digital HPS endeavors.

Because of the diverse expertise at the MBL and the wealth of resources available, specific problems usually had experts that could provide solutions. For example, when we had issues with the handles server for the EP (which maintains and serves a database of persistent identifiers for publications) and wanted to investigate other persistent ID solutions, we were able to discuss potential alternatives with the staff at the MBLWHOI Library. When Yamashita wanted to know more about how agile development is implemented within an informatics group, he was able to discuss it with members of the Encyclopedia of Life's Biodiversity Informatics Group (BIG). Machine learning was discussed and presented as a way to mine textual data for specific information, a useful tool for heavily text-based projects like the EP. And when he needed to figure out how to access datastreams in a Fedora commons repository (http://www.fedora-commons.org/) and export them to another content management system, Omeka (http://omeka.org/), a solution was found after a few brainstorming and code exploration sessions with developers. Equally important were the experiences of reporting on the SIG's projects to the MBL Director and participating in informal biodiversity informatics sessions with the EOL's BIG.

In addition to the resources at the MBL, a gift from the George F. Jewett Foundation (235 Montgomery Street, Suite 612, San Francisco, CA 94014) provided funds to invite a number of visiting lectures based on topics of interest were held. All the informaticians were able to interact with the speakers in an informal setting: P. Bryan Heidorn from the University of Arizona ("Biodiversity Informatics: Mining Untapped Resources", http://www.slideshare.net/pbheidorn/mblwhoil2010-heidorn); Peter Fox from Rensselaer Polytechnic Institute ("XInformatics: bridging the gap between science and discipline-neutral cyber infrastructure with semantics", http://hickory.eol.org:8081/display/public/XInformatics); and Tom Mitchell from Carnegie Melon University ("Read the Web", http://hickory.eol.org:8081/display/public/Read+the+Web). These lectures were open to the Woods Hole community and very well-attended. Indeed, one of the MBL's strengths is its ability to attract leaders of the field even in the winter.

Too often informatics projects are "accidentally" constructed (Peter Fox, personal communication). This training provided an understanding of how to construct informatics projects from the ground up, purposefully. What kind of expertise is needed? What hardware decisions need to be made? What metadata considerations ought one to think about and what are the repercussions of choosing a specific repository over another? The answers to these questions partially depend on the specific needs of the project. Yet, it is evident that almost all informatics projects require a rich mixture of developers, designers, systems administrators, and content experts. Knowing how to choose the right personnel and the proper hardware and understanding how to develop the right tools for the project requires learning and surveying the general landscape of informatics. These practical concerns are just as important as learning to code. But even if the end goal for a researcher is not to lead an informatics project, this kind of training still holds immense value because it brings the researcher into the common lexicon of the informatician. Communication is enhanced and the language of informatics is better understood. The researcher is better able to articulate details of his/her data to the informatician, thus enabling the informatician to make informed and educated decisions about the data on behalf of the researcher. Armed with this knowledge scientists and scholars can, like the BioMedical Informatics Fellows, become agents of change within their fields.

Future Plans/Further Development:

We anticipate that this kind of hands-on approach to informatics training will become more and more prevalent as the lines between science and informatics become further blurred in the near future. Therefore, Yamashita's visit to the MBL can be viewed as a model for informatics training that could certainly be mimicked by other institutions and programs. As a follow-up to Yamashita's time at the MBL we are in the process of developing an informatics "boot camp" for scholars in the humanities and biodiversity fields, a condensed version of what Yamashita experienced in his three-month visit. Modeled after the highly successful BioMedical Informatics Course, the one-week boot camp would bring together interested scholars in the humanities (or biodiversity fields) to work on specific aspects of their individual projects. Lectures, exercises, and working groups would bring participants up to speed on contemporary topics in informatics as decided by the organizers and faculty. Unlike the Medical Informatics course, however, the boot camp would require that participants already have extant projects. The course would be a time to not only learn about the basics of science informatics and survey the diversity of tools and techniques currently available, but also delve into project-specific issues. In so doing, participants not only learn about how to create solutions for their particular sets of problems, but also learn techniques and utilize tools that they otherwise might not have

encountered.

The SIG and the MBLWHOI Library also encourage longer visits by scientists and scholars who wish to learn more about informatics and its applicability to different kinds of science. Specific programs of study can be constructed and tailored to individual interests and problems. The MBL is a unique place and an even rarer institution, one which supports a vibrant community of informatics researchers working on assorted projects ranging from biodiversity to cell imaging to marine ecosystems. It has promoted and cultivated informatics training and education for almost two decades and continues to grow in the diversity of fields in which they are experts. We can think of no better place to be.

For more information about informatics opportunities at the MBL, please contact Cathy Norton (cnorton@mbl.edu) or Holly Miller (hmiller@mbl.edu).

Key Points:

- More so than ever, the boundaries between science and informatics are blurred, yet few scientists are trained in informatics.
- Scientists need to understand informatics approaches to advance their research and also to lead graduate students and post-docs in their labs.
- The MBL, which is home to a world-renown library and research in BioMedical informatics, biodiversity informatics, science studies informatics, and science informatics in general, is an ideal place for scientists and scholars looking for training opportunities in science informatics.

Funding:

This work was supported by the National Science Foundation [0926026 to G.Y., SES-0623176]; Jewett Foundation; Ellison Medical Foundation.

Acknowledgements:

We thank the MBL, ASU, and the Center for Biology and Society for their support. G.Y. thanks the MBLWHOI library staff and especially the SIG - R Schenk, A Shipunova, L Akella, C Ha, and J Hufnagle - for their hospitality in welcoming me as part of the team. G.Y. is grateful for the love and support of Juri, especially being away for those many months.

Grant Yamashita is a post-doctoral fellow in the Center for Biology and Society at Arizona State University. He is an evolutionary biologist who works on the historical and conceptual issues associated with germ-line evolution and the evolution of multicellularity. Currently, he is interested in bringing informatics approaches to the many digital projects in the history and philosophy of science.

Holly Miller is the project leader of the Science Informatics Group at the MBL. A biochemist and enzymologist by training, she became interested in developing more efficient and useful ways to organize and distribute scientific data. Her obsessions at the moment are scientific ontologies and the semantic web.

Anthony Goddard is a developer and systems administrator with the Science Informatics Group of the MBL. He specializes in Linux, Ruby, OS X, virtualization and distributed architecture. He is currently interested in bringing innovative technology solutions to traditional scientific research and data management problems.

Cathy Norton is the Director of the MBLWHOI Library and Vice Chair of the Biodiversity Heritage Library. She has been the PI on the BioMedical Informatics course for more than 20 years and also manages the Information Technology Department at the MBL. Her current interests are in provenance and preservation of digital assets and open access to scientific literature and data sets.

References:

1. McGary KL, Park TJ, Woods JO et al.. Systematic discovery of nonobvious human disease models through orthologous phenotypes. PNAS 2010;107:6544-49.
2. Heidorn PB, Palmer CL, Wright D. Biological information specialists for biological informatics. Journal of Biomedical Discovery and Collaboration 2007;2:1-5.
3. Bennett-McNew C, Ragon B. Inspiring Vanguards: The Woods Hole Experience. Medical Reference Services Quarterly 2008;27:105-110.
4. Brummit A. The Woods Hole Experience: How a weeklong intensive course provides a deeper understanding of the interrelated disciplines involved in medical informatics. MD Computing 2001;January/February:34-6.
5. Baldwin P. NLM/MBL Informatics Fellowship: Is It for Hospital Librarians? Journal of Hospital Librarianship 2004;4:89-94.

6. Wilson L, Gordon MG, Cornelius F et al.. National Library of Medicine and the Marine Biological Laboratory Biomedical Informatics Fellowship - One Team's Experience. In Park HA et al. (eds). Consumer-Centered Computer-Supported Care for Healthy People. Amsterdam: IOS Press, 2006,957.
7. Anderson K, Markland M. NLM's Medical Informatics Course. NLM Technical Bulletin 2002 (5).
8. Bridges J, Miller CJ, Kipnis DG. Librarians in the Woods Hole Biomedical Informatics Course. Medical Reference Services Quarterly 2006;25:71-81.
9. Maienschein J, Laubichler M. The Embryo Project: An integrated approach to history, practices, and social contexts of embryo research. Journal of the History of Biology 2010;43:1-16.
10. Laubichler M, Maienschein J, Yamashita G. The Embryo Project and the Emergence of a Digital Infrastructure for History and Philosophy of Science. Annals of the History and Philosophy of Science 2007;12:79-96.
11. Ouzounis CA, Valencia A. Early bioinformatics: the birth of a discipline - a personal view. Bioinformatics 2003;19:2176-90.
12. Gruber T. A translation approach to portable ontology specifications. Knowledge Acquisition 1993;5:199.
13. Lui KM, Chan KCC. Pair programming productivity: Novice–novice vs. expert–expert. International Journal of Human-Computer Studies 2006;64:915-92.