Datasphere at the Biosphere II: Computation and data in the wild

Bryan P. Heidorn, University of Arizona
Gretchen Renee Stahlman, University of Arizona
Steven Chong, University of Arizona

Abstract
Biological Field Stations provide a unique set of opportunities and challenges for digital curation. The stations serve as the center of short-term and long-term biological research, from biomolecular-scale to ecosystems-scale research. They represent some of the last remaining “natural” areas in certain regions. Stations provide unique information about local biotic and abiotic conditions. Data shared among the stations support continental scale and global research initiatives. The stations themselves support a large number of researchers who often come from multiple universities and other research and teaching institutions around the world. Because of this decentralized user base, it is particularly difficult for stations to capture data and other research products generated by research at the stations. The authors, part of a larger NSF-funded project, conducted a survey of field station researchers and then held a two-day workshop to identify challenges and opportunities for “grand challenge” research questions that could be enabled through development of cyberinfrastructure. We were particularly interested in “long-tail” data (Heidorn, 2008), which refers to large numbers of smaller datasets rather than only the large collections of homogeneous data frequently associated with “big data”. The information gathered through this study will inform future proposals for cyberinfrastructure development.

Keywords: data, cyberinfrastructure, biology, ecology, Software-as-a-Service, SaaS, long-tail
Citation: Heidorn, B.P., Stahlman, G.R., Chong, S. (2015). Datasphere at the Biosphere II: Computation and data in the wild. In *iConference 2015 Proceedings*.

Research Data: In case you want to publish research data please contact the editor.
Contact: heidorn@email.arizona.edu, gstahlman@email.arizona.edu, stevenchong@email.arizona.edu

1 Background

The objective of this study is to identify the cyberinfrastructure needs of small and medium laboratory (SML) teams that could be met with cloud-based Software-as-a-Service (SaaS).The value of research stations is determined in large part by the research that is conducted at the stations. Unfortunately, researchers have had little incentive to provide the research sites with copies of their data. Such data could be valuable to later researchers, since the data can give improved context to future work. Federal agencies such as NSF are now requiring that grant proposals include data management plans, but researchers in ecology frequently do not have satisfactory options for storing their data. An analysis of recent NSF ecology grants indicated that less than 8% of non-genomic data is shared (Hampton et al., 2013). Much of this data becomes “dark data” (Heidorn, 2008). A premise of the Empowering Long Tail Research project is that if we provide incentives for moving data to research station-managed cloud computing around the time of original data collection, researchers will leave their data with the station. A survey and 2-day workshop were conducted to generate grand-challenge research scenarios that would be enabled by SaaS. Results of the workshop are summarized here.

Our population of interest is researchers who used field stations in the Organization of Biological Field Stations (OBFS) (http://www.obfs.org/). OBFS has 208 members and affiliate members. The member stations exist mostly in the western hemisphere, with a preponderance in North America but with members in Latin and South America, Pacific islands, Australia and Europe. Each year thousands of researchers and students visit OBFS field stations. The stations cover 31,205 hectares (120.5 square miles) representing woodlands, wetlands, oceans, prairies, deserts and every other natural area type (Wyman, et al., 2009). Some stations are associated with small colleges or large universities while others are associated with private reserves and foundations or semi-public organizations such as the Smithsonian and the American Museums of Natural History. As a rule, station administrators have little control of the data and research conducted at the sites, so beyond simply a head count of researchers, there is no global repository of publications and data associated with the stations. Part of the goal of the research being reported here has been to identify cyberinfrastructure resources that could be shared among the stations, which would facilitate research and indirectly encourage researchers to deposit data and results.

2 Method

Two groups of participants were selected to provide insights into the operations of biological field stations that could be facilitated by improved cyberinfrastructure. To create a representative sample of ecological regions and ecological questions, the main group of participants was identified through a stratified random sample of researchers who had published papers based on research conducted at field stations selected across the 20 National Ecological Research Network (NEON) regions. Participants in the workshop were chosen by searching the BioOne database of academic journals for first-authors of papers referencing biological research stations in each NEON region. Original authors could nominate substitutes who had been involved in the research. If authors of the initially selected papers could not attend or did not reply, they were replaced by authors of papers about stations that did not yet have representation in each region. To fill remaining slots after the initial pool of invitees had been exhausted, additional workshop participants were selected from the attendance list of an Organization of Biological Field Stations meeting held fall, 2013 at the Southwest Research Station near Portal, AZ. The list of 109 attendees (who had also been included in our initial wave of survey recipients) was randomized, and invitations were issued until all available workshop slots were occupied. Thirteen field station researchers were included in the workshop (one participant from the initial group of 14 confirmed attendees in this category canceled just prior to the workshop). In addition, 6 researcher-administrators were invited to and attended the workshop.

The workshop was held at the University of Arizona's Biosphere II facility. Following an introduction and description of SaaS, participants were encouraged to imagine that they were being sent to inhabit a biosphere for interplanetary colonization, focusing on the data and computational resources they would want to take on the journey to enable study of the biosphere and improved biological management. A set of brainstorming breakout sessions was directed towards research that was being limited by lack of computational tools and organized datasets. Note-takers were assigned to each breakout group, and each group subsequently summarized the discussion to the larger group. This process led to a long list of research questions. Very similar questions were collapsed into single questions and the duplicates were removed. The workshop participants then voted to identify the five most interesting research questions. Breakout groups were then formed around each of these questions. These groups identified the key data and computational resources that would be required to answer the questions. All groups worked together to identify the key shared computational and data resources that were held in common between the grand research questions, representing the most important resources to advance research at the field stations.

3 Results

Results can be categorized into two sets: the selected grand challenge research questions, and the computational and data resources that would be required to answer the questions.

The five research grand challenges were:

- What is the tipping point for habitat destruction that would cause a species collapse?
- How does environmental change impact spatial dynamics among species?
- Are the changes in the population of a species correlated with factors A, B and C, and can these be used to predict future population changes?
- How is sea level rise impacting coastal ecosystems?
- How do climate change, habitat modification and invasive species interact to impact biological community function?

A breakout group was established for each grand-challenge research question. Each group produced a list of resources that would be needed to answer the research questions. While there were many critical resource challenges identified, the following shared themes emerged.

These included a set of data products, data set characteristics and computational tools. The most commonly referenced data sets consisted of remote sensing data of the sites, including MODIS, LANDSAT, LIDAR and others, as well as hydrology data sets, weather data and species distributions over time. Other data include species observations, species characteristics such as phenology, and biotic measures such a plot data. SaaS services would need to provide historical data for the sites in one easy-to-use repository or gateway so that researchers can easily identify what is available. Derived data included species distribution models, vegetation layers, and hydrology models. All of these are generated by a set of existing software packages such as MaxEnt, which should be made available as SaaS. The workshop group identified a chief challenge of data interoperability. Data gathering methods often vary, producing a diversity of data formats. There are currently few accepted standards for data representation, so SaaS services would need to adopt or develop such standards. All of these types of resources require an investment in software, hardware and human effort. If they are developed under the SaaS model, the costs could be amortized over many stations and many research projects, reducing the time spent by researchers in data wrangling to support their projects.

References

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., & Porter, J.H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment, 11* (3), 156-162.

Heidorn, P. Bryan (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends, 57* (2), 280-299. doi: 10.1353/lib.0.0036

Showstack, R. (2014). Field stations and marine labs urged to adapt to changing economies and technologies. *Eos, 95* (32), 286-287.

Wyman, R.L., Wallensky, E., & Baine, M. (2009). The activities and importance of international field stations. *BioScience, 59* (7), 584-592.