EOL-BHL-NESCent Research Sprint Report

Cynthia Sims Parr[1] and Craig McClain[2]

[1] Encyclopedia of Life, National Museum of Natural History, Smithsonian Institution, Washington DC 20013-7012
[2] National Evolutionary Synthesis Center, 2024 W. Main St., Suite A200, Durham, NC 27705

ABSTRACT

There are exciting biological questions which require programming and large online data resources to address, yet many traditionally-trained biologists lack the informatics skills needed for successful analysis. In the last decade, new aggregated, open data resources have become available, but these are not often being leveraged effectively for big data research. Using a new hackathon-inspired format, the EOL-BHL-NESCent Research Sprint facilitated scientific discovery by supporting teams to use online data resources such as the Encyclopedia of Life and Biodiversity Heritage Library to answer their biological questions. We describe the methods of the research sprint and present results indicating its success in producing publications, in introducing scientists to large-scale informatics resources and approaches, and in encouraging new collaborations.

Keywords: informatics, biodiversity, collaboration

INTRODUCTION

The EOL-BHL-NESCent Research Sprint was a four-day meeting (4-7 February 2014) addressing pressing biodiversity questions held at the National Evolutionary Synthesis Center (NESCent) in Durham, NC USA. The event was motivated by our belief that there are exciting questions which require programming and big data resources to address, yet many traditionally-trained biologists lack the informatics skills needed for successful analysis. At the same time, we wanted to help informaticians apply their techniques more directly to answer urgent ecological and evolutionary questions being asked by biologists. Thus, our research sprint would enable biologists and informaticians to make discoveries through joint activities. Finally, the event would allow us to test the suitability for informatics research of major resources such as the Encyclopedia of Life (EOL), particularly its recently launched TraitBank (Parr et al., view), and the Biodiversity Heritage Library (BHL) (Gwinn and Rinaldo, 2009).

This event represents a new kind of intense face-to-face event called a *research sprint*, which was designed to accelerate discoveries in this domain using these specific online resoruces. However, the research sprint approach should be useful to many other domains, addressing general problems in data sharing and big data analysis, e.g. Borgman (2012). It was modeled after hackathons, which have been highly successful at producing new tools and promoting collaborative, open source software and communities. The term *research sprint* has even been used by Google Ventures (Margolis and Zeratsky, nd) to describe a fast-paced process of market and user research. However, the primary aim of our research sprint was to produce scientific research results rather than informatics tools, algorithms or products. Of course, new tools and techniques may be needed to achieve research goals, but the focus on biological research (in this case) insures that these are only developed where existing tools don't exist. As with hackathons, education and networking are also important.

METHODS

Five months prior to the event we invited biologists to submit research questions they wished to address using EOL or BHL or other large-scale biodiversity online resources. Organizers reviewed submissions and chose those most likely to succeed. We then matched each biologist with a programmer that we or NESCent staff knew to have relevant skills or who had responded with interest to broadcasted tweets or other announcements about the event. Although biologists could suggest informaticians who might

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.503v1 | CC-BY 4.0 Open Access | rec: 17 Sep 2014, publ: 17 Sep 2014

be appropriate for their project, few did so. This suggests that the Research Sprint event was reaching biologists not previously experienced in informatics approaches. While teams of two were initially planned (one biologist plus one programmer), several groups acquired additional remote or local participants or participated in more than one team. Overall, 13 biologists were teamed up with 11 informaticists on 9 projects. Teams were supported by 6 support staff from EOL, BHL or NESCent. The group is pictured in Figure 1.

In the weeks prior to the four-day event we encouraged participants to exchange background reading and ask questions by email. During the sprint, teams began data mining and querying and started exploratory analyses. Day 1 included introductions and flash talks so that teams became aware of each others' projects before diving into their own. After day 1, each day started with a brief tutorial on a topic that arose in the previous day's discussions. For example, a demonstration on using the EOL API. Teams shared experiences at the end of each day, prompted by the following requests:

- Day 1: Share any barriers to work
- Day 2: Share progress made towards goals
- Day 3: Share a visual representation of progress or results
- Day 4: Share a plan for finishing analyses and publications or grant proposals

Figure 1. Participants in the EOL-BHL-NESCent Research Sprint.

RESULTS AND DISCUSSION

The four-day research sprint duration was long enough to start producing results (these began emerging in only 2.5 days) but not so long that momentum flagged.

Sprint projects topics included: vulnerability to climate change, color in butterflies, character displacement in vertebrates, conservation risk of amphibians, defining habitat, text mining to track anatomical terminology over time, mutualisms and global change, gaps in knowledge about the tree of life, niche width and biodiversity. An additional topic arose during the meeting: rate of species description in gastropods. Several other projects emerged that may have potential, including new tools and features. While it is possible that not all of these projects will result in publications, a collection of manuscripts appears in PeerJ (The EOL-BHL-NESCent Research Sprint collection).

How successful was the EOL-NESCent-BHL research sprint in achieving its goals? In addition to the growing collection of research papers, EOL and BHL gained significant feedback about how scientists would like to use their resources, for example, most researchers used R for their analyses and combined data from EOL with data from other sources. Moreover, a post-event survey suggests that participants thought the event was a success. The survey was administered by Survey Monkey and completed by 12 participants, representing a 50% response rate. A quarter of the respondents indicated no prior knowledge of informatics techniques; after the event all respondents feel they were now at least somewhat knowledgable. All were likely (6/12) or very likely (6/12) to submit a manuscript for publication. Some (4/12) were likely or very likely to use the results of the sprint in grant proposals. All but one of the twelve gained at least one new collaborator (one collaborator: 5, two collaborators: 3, three or more collaborators: 3). Nearly all were satistified (5/12) or very satisfied (5/12) with the length of the sprint and all were satisfied (5/12) or very satisfied (7/12) with the small teams format of the sprint. Most important, free responses indicated that participants most valued the opportunity to connect with other researchers and the new ideas and approaches that were shared.

In conclusion, the research sprint format was effective in facilitating research utilizing major online resources such as EOL and BHL. It was also successful in introducing scientists to larger-scale informatics approaches and in encouraging new collaborations.

ACKNOWLEDGMENTS

Support was provided by the Richard Lounsbery Foundation, Robert T. and Catherine T. MacArthur Foundation, and the National Evolutionary Synthesis Center (NESCent) NSF EF-0905606. Thanks to Danielle Wiggens, Lori Houjak, Patrick Leary, Jeremy Rice, William Ulate, John Mignault, and Katja Schulz.

REFERENCES

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078.

Gwinn, N. E. and Rinaldo, C. (2009). The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA Journal*, 35(1):25–34.

Margolis, M. and Zeratsky, J. (n.d.). The GV research sprint: a 4-day process for answering important startup questions — Google Ventures.

Parr, C. S., Wilson, N., Schulz, K. S., Leary, P., Hammock, J., Rice, J., and Corrigan, Jr., R. J. (in review). TraitBank: Practical semantics for organism attribute data in Special Issue on Semantics for Biodiversity. *Semantic Web Journal*.