Journal of
open research software

Zentner, L et al 2014 nanoHUB.org: Experiences and Challenges in Software Sustainability for a Large Scientific Community. *Journal of Open Research Software,* 2(1): e19, pp. 1-5, DOI: http://dx.doi.org/10.5334/jors.bd

ISSUES IN RESEARCH SOFTWARE

nanoHUB.org: Experiences and Challenges in Software Sustainability for a Large Scientific Community

Lynn Zentner*, Michael Zentner*, Victoria Farnsworth*, Michael McLennan*, Krishna Madhavan* and Gerhard Klimeck*

The science gateway nanoHUB.org, funded by the National Science Foundation (NSF), serves a large scientific community dedicated to research and education in nanotechnology with community-contributed simulation codes as well as a vast repository of other materials such as recorded presentations, teaching materials, and workshops and courses. Nearly 330,000 users annually access over 4400 items of content on nanoHUB, including 343 simulation tools. Arguably the largest nanotechnology facility in the world, nanoHUB has led the way not only in providing open access to scientific code in the nanotechnology community, but also in lowering barriers to the use of that code, by providing a platform where developers are able to easily and quickly deploy code written in a variety of languages with user-friendly graphical user interfaces and where users can run the latest versions of codes transparently on the grid or other powerful resources without ever having to download or update code. Being a leader in open access code deployment provides nanoHUB with opportunities and challenges as it meets the current and future needs of its community. This paper discusses the experiences of nanoHUB in addressing and adapting to the changing landscape of scientific software in ways that best serve its community and meet the needs of the largest portion of its user base.

Keywords: science gateway; infrastructure; software; licensing

Background

The Network for Computational Nanotechnology Cyber Platform (NCN CP) is responsible for the operation and support of the science gateway nanoHUB.org, serving a large scientific community centered around the nanoscience and nanotechnology fields. There are many other science gateways, portals, and scientific social networks that also serve unique communities such as ours and that are also exploring the landscape of scientific software and how the related issues impact their communities. Some gateways have been discussed in the context of their use of grid resources, such as the variety of platforms associated with TeraGrid/XSEDE [1]. Gateways such as GridChem, LEAD, nanoHUB, and CaBIG were observed to share common goals despite serving communities as disparate as meteorology and cancer research. Serving their communities required having an infrastructure in place that allowed the gateways to not only lower barriers in providing access to HPC resources to researchers who may not have had that access on their own, but also do it in a way that addressed usability for those who may be unfamiliar with complex computational infrastructure and simultaneously serve the needs of a community that may include students beginning to learn about computation and modeling as well researchers studying complex problems. Other infrastructures serve scientific communities by functioning under a social network style paradigm, such as MyExperiment [2]. The MyExperiment platform allows a variety of scientific domains to share scientific workflows and other digital objects and drive collaboration through that sharing. MyExperiment utilized workshops and feedback to allow the needs of the end-users to help focus design decisions. Likewise, they developed processes for contributing content that provided appropriate attribution and protected the rights of the content developers, while still promoting sharing and allowing ease of contribution. MyExperiment leverages the participation of the community of users who are not active contributors through usage data and reviews provided back to the community.

The concept of community may be central or peripheral to the organization and operation of various science gateways. Merriam-Webster defines community as a "body of persons of common and especially professional interests scattered through a larger society." In the case of nanoHUB, we see our community as central to the success of our gateway. We have a challenge of supporting a variety of community members, with varying viewpoints

* Network for Computational Nanotechnology, Purdue University, West Lafayette IN, USA
lzentner@purdue.edu

Corresponding author: Lynn Zentner

and needs; as in the case of MyExperiment, generally many wish to participate and utilize nanoHUB while a significantly smaller subset provide the content critical to maintaining a dynamic online facility for that community. nanoHUB has a significant ten year history of data and experience. nanoHUB users exceed 300,000 annually and access a portfolio of over 4400 resources contributed by over 1,600 authors, including over 340 simulation tools contributed by 441 software authors and developers. Such broad use and a large, vibrant community provide continued opportunities for growth, but careful management of policies and processes is necessary to anticipate and meet associated challenges.

The nanoHUB Community

The nanoHUB community can be thought of as being composed of several overlapping groups of stakeholders. At its core, nanoHUB serves its user community with cutting edge tools and learning materials that may be so new as to not yet have been presented in textbooks. This user community consists of users in both the research and educational arenas and benefits from and helps speed the transition of research code into use by other researchers as well as into the classroom. We have documented the rapid transition of research codes into the classroom, on average, in less than 6 months from the initial publication of the code on nanoHUB [3].

A second group served by nanoHUB is the group of researchers who themselves are developing code related to their scientific research areas. nanoHUB, through its easy to use Rappture Development Toolkit [4], allows the scientists, with minimal training, to deploy and maintain their code in a way that is easily accessible. This approach eliminates the "middle man," a computer scientist previously required to rebuild and maintain code for use on the web, effectively disenfranchising the original author. nanoHUB presents the originating scientists with an opportunity to share their research products easily and continue to stay involved in the ongoing support, maintenance, and enhancement of their code, with significant and measureable impact.

Lastly, nanoHUB plays an active role in the cyberinfrastructure community. nanoHUB found such success within its own scientific community that the infrastructure powering it was extracted in order to bring similar HUB technology to other scientific areas [4]. The result is that nanoHUB now contributes to and benefits from development efforts to expand and improve the functionality of the core infrastructure, known as HUBzero. The nanoHUB cyberinfrastructure does not operate in a vacuum, but rather takes the opportunity to leverage and incorporate appropriate technologies, such as Pegasus workflow management tools [5, 6], to the benefit of the nanoHUB users and tool developers. The nanoHUB and HUBzero teams approach incorporation of other existing technologies as driven by use cases and leveragability. When there is a compelling use case and the opportunity for extensibility, development work for implementing the existing technology is prioritized accordingly.

Opportunities and Challenges

Managing and growing a successful cyberinfrastructure such as nanoHUB presents a variety of opportunities and challenges, particularly in regard to software. nanoHUB is in the somewhat unique position of dealing with issues related to two types of software: the open source HUBzero software that powers the infrastructure as well as the many scientific codes contributed and deployed by the nanotechnology community. Over the years, nanoHUB has explored several issues related to software deployment and publishing, including licensing, intellectual property, export control, incentives, and quality.

Maximizing Participation Relies on Tolerant Licensing

nanoHUB has repeatedly considered the implications of licensing, both with regard to its core HUBzero platform software and with respect to scientific code contributed by its community members. The HUBzero code is available through regular open source release under the LGPLv3 license. The LGPLv3 license allows academic institutions and even commercial entities to use, modify, and redistribute this code—even to sell the code or services related to using the code—provided that they make any changes available to entire community. This has the potential to build up an ecosystem around the code, and encourages all parties to share their enhancements with one another. Because we are using the "lesser" General Public License, however, researchers can treat our toolkit as a library and keep the analysis tools that they create and deploy within the HUBzero framework as proprietary code, or license their tools under any other license they choose. Developers are encouraged to feed changes back to HUBzero, such that they can be considered for the next open source release. Other open-source license choices were considered, ranging from relatively permissive in the case of the Berkeley Software Distribution (BSD) licenses to the full General Public License (GPL), which must be carried forward in all subsequent derivatives. The LGPLv3 license and approach lies near the middle of the spectrum and was selected in order to encourage open collaboration without being overly restrictive. Changes to the platform carry the same license, while the developer can license new components as they choose.

Keeping the code closed source did not fit with the intended model for the HUBzero platform, because it was always intended to allow people to modify and contribute to the code. To make matters as simple as possible, the HUBzero Foundation owns the code, so that there is one copyright holder. The Foundation also acts as the reviewer of new code contributions, with an established review process in place. To date, about 12 contributions to the code have come from outside the HUBzero team, ranging from patches to full components.

A similar context drives how we handle licensing of scientific codes on nanoHUB. We believe that a steady stream of high quality, open access content is necessary to continue to grow and maintain a vibrant community and strive to lower barriers to dissemination of content on nanoHUB.

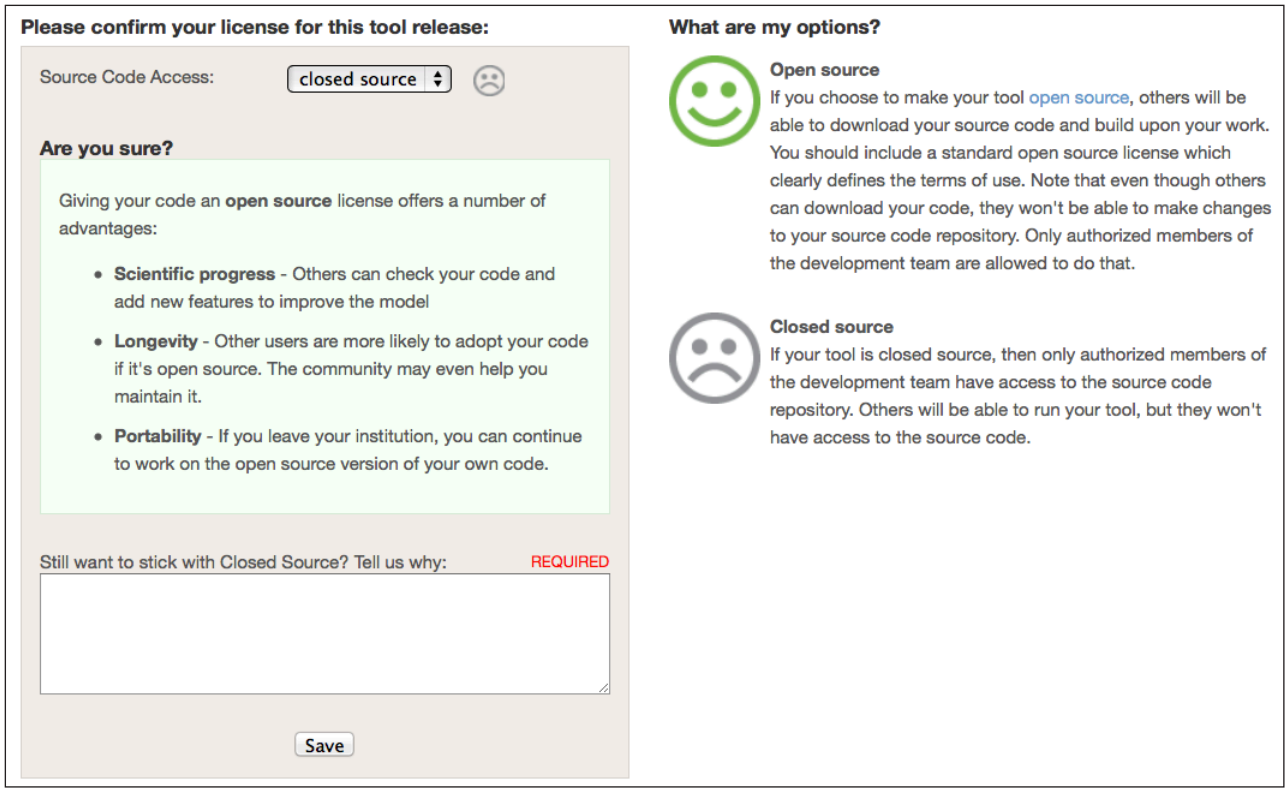

Figure 1: Before a developer can publish a tool, they must confirm the license for their tool. This text encourages selecting open source, provides several reasons to do so, and requires the developer to provide a reason for choosing closed source.

When an author submits code on nanoHUB, we provide as much flexibility as possible to them so that they can contribute code to the community and still meet the requirements of their funding agency, institution, as well as their personal intellectual property concerns. In general all the codes need to be in an ***open access*** form, where any nanoHUB user can run and execute the scientific codes via a graphical user interface, much like an app on an iPhone, except that the science codes on nanoHUB run on remote venues, either at Purdue or on the grid, on behalf of the user. We provide an opportunity for contributors to license their code as ***open source***, and have found that out of 341 currently published tools, 50 tools have utilized an open source license, with their authors choosing a variety of flavors of that license, ranging across GPL, NCSA, BSD, and LGPL.

While scientists can and will share their tools with the community through the nanoHUB open access policy, the above numbers indicate that a *requirement* by a cyberinfrastructure such as nanoHUB for contributors to share their source code through an open source license would drastically and negatively affect the sharing of tools that we have seen historically. *We believe that the cyberinfracture can play a role in providing education on the benefits and best practices of open source release, but ultimate choice of the actual license belongs to the tool authors, funding agencies, and supporting institutions.* As a first step in that education process, nanoHUB has updated the wording on the tool submission page to encourage developers to consider an open source license (**Figure 1**). We also began requiring developers who did not choose to publish their tool as open source to provide the reason for their choice. Though the developer provides their reasoning in free text, reviewing the submissions provides a few main reasons for not selecting an open source license including: the tool is still under development/review or is still being refined, the tool contains code from another developer who has not authorized open source, and, in one case, the tool was developed to provide an example to students who will be writing their own code. Additionally, one developer pointed out that although their tool has many users, no one had requested the source code, so they felt that open access to use the tool was more important to their user community than having access to the source code. We have also become aware recently that developers can be confused as to which open source license to select and are considering suggesting a recommended license that would suit the needs of the largest group of our developers.

World Complexities Demand Flexible Software Access

Another issue nanoHUB has needed to consider is related to restrictions regarding export control. The content authors carry the responsibility of knowing whether there are any export restrictions on the code they deploy on nanoHUB. Our contribution process allows them to restrict access to their code accordingly, allowing the choice of full access, restriction to US users only, restriction to non-D1 nations, or in the case of commercial software that may be licensed only to a particular set of users,

Figure 2: Snapshot overview of the nanoHUB tool "Bandstructure Lab". 38 citations refer to citations in the scientific literature. Usage in classes has been determined by an automated correlated behavior analysis.

licensing to particular groups. Of the 343 tools currently published on nanoHUB, 8 are restricted to US users, 6 are restricted to non-D1 nations, 2 are restricted to Purdue users, and 8 are restricted to a particular group of users. The remaining tools are open access. While the majority of tools are open access, we maintain that the above choices allow us to support the greatest number of users with the greatest number of tools, which allowing contributors to control access in a way that meets any institutional, funding agency, or commercial requirements.

Incentives and Low Barriers to Participation Keep the Content Pipeline Flowing

A last set of considerations revolves around incentivizing contributions while maintaining quality. As mentioned above, we strive to lower barriers to contribution in order to maintain a steady flow of content to our community. However, we must balance the ease of contribution with maintaining a level of quality in our contributed software. Software contributors are strongly encouraged to provide at least minimal documentation, such as a first time user guide as well as scholarly publications that support the scientific approach and content of the code. nanoHUB also provides a mechanism for developers to create regression testing suites for their code, such that code revisions can be vetted against these tests.

With over ten years of experience in hosting scientific tools, the nanoHUB team has concluded that the user community can be a strong partner in crowd-sourcing quality control. Through open, transparent mechanisms such as reviews, question and answer forums, wishlists, citation counts, and usage statistics, it is easy for users to see which tools are actively used and maintained, and the highest quality tools are allowed to bubble to the top. These same mechanisms provide an incentive for authors to contribute and maintain their codes, providing both a heartbeat of the quality and usefulness of a particular tool as well as quantifiable measure of the tool and author's impact on the scientific community. See for example **Figure 2**, which shows the snapshot view of the very popular tool *"Bandstructure Lab"* on nanoHUB.org. This tool [7, 8], like any other published tool, has a digital object identifier that can be cited in scientific publications. Lastly, we take the view, as do most tool authors, of the tools as a formal publication, which carries the author's name publicly. It is therefore in the responsible author's interest to deliver high quality material. Additionally, nanoHUB has an established procedure for identifying and vetting citations to nanoHUB tools and resources in the scientific literature [9] and provides that data with each tool or resource. Currently, 118 nanoHUB tools have been cited at least once, but only 14 tools have been cited 10 or more times. We intend to further study the relationship between usage, citation, and licensing choices to determine what correlations exist and whether any of these factors significantly influence the others.

Conclusions

At the vanguard of the science gateway community, nanoHUB has established itself as a leader both in hosting of scientific code and development of a production-level, open source cyberinfrastructure platform. The nanoHUB team continuously considers and adapts to the rapidly evolving challenges facing the scientific and software community. We have found that a flexible approach safeguarding the rights and concerns of software authors while striving to quickly bring quality codes to a larger audience leads to the best potential for accelerating the transition of research from the labs and the scientists to the classroom and the greater scientific community.

Acknowledgements

Mark S. Lundstrom founded nanoHUB.org in 1998. In 2005, Michael McLennan created the Rappture Toolkit and Rick Kennell wrote the scalable middleware of HUBzero that, respectively, enable and power interactive nanoHUB simulations. The Network for Computational Nanotechnology (NCN) manages nanoHUB.org and is funded by NSF Award # EEC-1227110.

References

1. **Wilkins-Diehr, N, Gannon, D, Klimeck, G, Oster, S** and **Pamidighantam, S** 2008 TeraGrid Science Gateways and Their Impact on Science. *Computer*, 41(11): 32–41. DOI: http://dx.doi.org/10.1109/MC.2008.470
2. **De Roure, D, Goble, C** and **Stevens, R** 2009 The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, 25: 561–567. DOI: http://dx.doi.org/10.1016/j.future.2008.06.010
3. **Madhavan, K, Zentner, M** and **Klimeck, G** 2013 Learning and research in the cloud. *Nature Nanotechnology*, 8: 786–789. DOI: http://dx.doi.org/10.1038/nnano.2013.231

4. **McLennan, M** and **Kennell, R** 2010 HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Computing in Science & Engineering*, 12(2): 48–53. DOI: http://dx.doi.org/10.1109/MCSE.2010.41
5. **Deelman, E, Singh, G, Su, M-H, Blythe, J, Gil, Y, Kesselman, C, Mehta, G, Vahi, K, Berriman, G B, Good, J, Laity, A, Jacob, J C** and **Katz, D S** 2005 Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming Journal*, 13(3): 219–237
6. **McLennan, M, Clark, S, Deelman, E, Rynge, M, Vahi, K, McKenna, F, Kearney, D** and **Song, C** 2014 HUBzero and Pegasus: integrating scientific workflows into science gateways. *Concurrency and Computation: Practice and Experience*. DOI: http://dx.doi.org/10.1002/cpe.3257
7. **Mehrotra, S, Zentner, L, Klimeck, G** and **Vasileska, D** 2011 nanoHUB.org - the ABACUS tool suite as a framework for semiconductor education courses. In: Proceedings of the 11th IEEE Conference on Nanotechnology, Portland, Oregon on 15–18 August 2011, pp. 932–936. DOI: http://dx.doi.org/10.1109/NANO.2011.6144581
8. **Mukherjee, S, Paul, A, Neophytou, N, Kim, R, Geng, J, Povolotskyi, M, Kubis, T, Ajoy, A, Novakovic, B, Steiger, S, McLennan, M, Lundstrom, M** and **Klimeck, G** 2014 Band Structure Lab. Available at: https://nanohub.org/resources/bandstrlab. DOI: http://dx.doi.org/10.4231/D3SF2MC1C
9. **Klimeck, G, Adams, G, Madhavan, K, Denny, N, Zentner, M, Shivarajapura, S, Zentner, L** and **Beaudoin, D** 2011 Social Networks of Researchers and Educators on nanoHUB.org. In: Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Newport Beach, California on 23–26 May 2011. Washington, DC, USA: IEEE Computer Society, pp. 560–565. DOI: http://dx.doi.org/10.1109/CCGrid.2011.33

How to cite this article: Zentner, L, Zentner, M, Farnsworth, V, McLennan, M, Madhavan, K and Klimeck, G 2014 nanoHUB.org: Experiences and Challenges in Software Sustainability for a Large Scientific Community. *Journal of Open Research Software,* 2(1): e19, pp. 1-5, DOI: http://dx.doi.org/10.5334/jors.bd

Published: 9 July 2014

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS