Conceptualization of a US Research Software Sustainability Institute (URSSI)

Jeffrey C. Carver
University of Alabama

Sandra Gesing
University of Notre Dame

Daniel S. Katz
National Center for Supercomputing Applications; University of Illinois Urbana–Champaign

Karthik Ram
University of California, Berkeley

Nicholas Weber
University of Washington

The US National Science Foundation (NSF) funded conceptualization of a US Research Software Sustainability Institute (URSSI; http://urssi.us/) intends to make the case for and plan a possible institute to improve science and engineering research by supporting the development and sustainability of research software in the US. In this column, we—the URSSI's principal investigators (PIs)—introduce the broader *CiSE* community to our project.

WHY ARE WE DOING THIS?

Research software is essential to progress in the sciences, engineering, humanities, and all other fields. From 1995 through 2016, the NSF made 18,592 awards to projects referencing "software" in their abstracts, for a total of $9.6 billion. Indeed, modern research is inescapably digital because the processes that create, store, and analyze data and publications use tools and methods expressed in software. An examination of 40 papers in *Nature* from January to March 2016 showed that 32 explicitly mentioned software, with each paper mentioning an average of 6.5 software tools. Of the software tools mentioned in these articles, almost all were research software.[1]

In many fields, research software is produced within academia, by academics who range in experience and status from students and postdocs to staff members and faculty. The academic environment in which this software is developed, maintained, and used is quite chaotic with regard to the software development lifecycle. This is partially because the academic environment and culture have developed over hundreds of years, while software has only recently become important—in some fields over the last 60+ years, and in others, just in the last 20 or fewer years.[2] Further, only recently have frameworks such as science gateways, software repositories, and virtualization been widely available to significantly lower the barriers to sharing such software.

Copublished by the IEEE CS and the AIP
1521-9615/18/$33 ©2018 IEEE

Although much research software is developed in academia, important components are also developed in national laboratories and industry. Wherever research software is created and maintained, it can be open source (most likely in academia and national laboratories) or commercial/closed source (most likely in industry, although industry also produces and contributes to open source.)

The open source movement has created a tremendous variety of software, including software used for research and software produced in academia. This plethora of solutions is not easy for researchers to find and use out-of-the-box.[3] Standards and a platform for categorizing software for communities are lacking, which often leads to novel developments rather than reuse of solutions.[4] Three primary classes of concern are pervasive across research software and have stymied it from achieving maximum impact:

- *Functioning of the individual and team:* issues such as training and education, ensuring appropriate credit for software development, enabling publication pathways for research software, fostering satisfactory and rewarding career paths for people who develop and maintain software, and increasing the participation of underrepresented groups in software engineering.
- *Functioning of the research software:* supporting sustainability of the software; growing community, evolving governance, and developing relationships between organizations, both academic and industrial; fostering both testing and reproducibility, supporting new models and developments (for example, agile web frameworks, software as a service), supporting contributions of transient contributors (for example, students), and creating and sustaining pipelines of diverse developers.
- *Functioning of the research field itself:* growing communities around research software and disparate user requirements, cataloging extant and necessary software, disseminating new developments, and training researchers in the usage of software.

The goal of this conceptualization project is to create a roadmap for a URSSI to minimize or at least decrease these types of concerns. To do this, the two aims of the URSSI conceptualization are to

1. Bring the research software community together to determine how to address the issues about which we have already learned (described below). In some cases, there are already subcommunities working together on a specific problem, including those that we are part of, but those subcommunities might not be working with the larger community.[5] This leads to a risk of developing solutions that solve one issue but don't reduce (or might even deepen) other concerns.
2. Identify additional issues URSSI should address, identify communities for whom these issues are relevant, determine how we should address the issues in coordination with the communities, and determine how to prioritize all the issues in URSSI.

Although we believe this project is urgently needed by the US research community, we are not working in a vacuum. For example, there are the NSF CSSI program (https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505505), which supports our award; two ongoing institutes in science gateways (SGCI; https://sciencegateways.org/) and molecular sciences (MolSSI; http://molssi.org/); a recently completed conceptualization in high energy physics (S2I2-HEP; http://s2i2-hep.org/); two other conceptualization projects now underway in geospatial software (http://www.ncsa.illinois.edu/enabling/software/gsi) and computational fluid dynamics; and a large number of software development and maintenance projects. In the UK, the Software Sustainability Institute (SSI; https://www.software.ac.uk/), which has been in operation since 2010, is an inspiration and a potential model for our work. And the US Department of Energy supports the Better Scientific Software (BSSw; https://bssw.io/) community, which curates, creates, and disseminates information that leads to improved software for the advancement of computational science and engineering (CSE) and related technical computing areas, with a particular interest in CSE on high-performance (parallel) computers.

Given these existing activities, part of our challenge is to define how we will work with these other groups. For example, we might decide that they perform an activity so well that we should

point to it, such as the SSI's software guides. Or we might decide to either duplicate or enhance an activity they do to expand its impact, such as working with the SGCI to offer incubator services to a wider community than just gateway developers. Or we might decide to collaborate with one or more groups, such as on policy campaigns aimed at providing better career paths for research software developers in universities.

WHAT ARE WE DOING?

The conceptualization project has two primary aims that drive the activities. First, we plan to bring the research software community together to determine how to address the known issues related to research software. Second, we plan to identify additional issues that URSSI should address, how we should address those issues, and how to prioritize those issues. To accomplish these aims, we are engaging with a broad and diverse set of stakeholders via three types of activities: community workshops, a community survey, and community engagement activities.

Community Workshops

The community workshops will engage the community through discussions and brainstorming to focus in on the key priorities for the anticipated institute. The first large workshop, held at the University of California, Berkeley 10–12 April 2018, focused on building community support, identifying those areas in which an institute could impact the community, and those where more discussion is needed. This workshop consisted of a series of breakout discussions based on topics raised by the workshop participants, who had diverse backgrounds and roles in academia and industry. These topics included models for sustainability, funding for sustainability, receipt of credit for contributions to open source software, career paths for scientific software developers, and usability. We identified a number of approaches to addressing these topics by discussing how a potential institute could be organized around various themes such as training, policy advocacy, community building, new project incubation, and provision of support and services to scientific software developers. We anticipate future workshops will provide more input and help refine the specific topics that a potential institute would address and how best to organize this institute to effectively accomplish its goals.

While the first workshop was broad and useful for identifying a large number of issues that an institute could address, the remaining workshops will be smaller and focused on the more specific topics that emerged from the first. On the basis of analysis of the discussions from the initial workshop, we will identify topics and relevant people for the remaining workshops. The results of these workshops will again be publicized on our project website.

Community Survey

To obtain a broader set of inputs than can be collected from in-person workshops, we are developing a survey. The goal of this survey is to gather input regarding the key challenges related to sustainable scientific software. The results of the community survey will be instrumental in ensuring that the vision and goals of an institute align with the key needs of the scientific software community.

Ethnographic Studies

To better understand the diversity of practices, governance models, and community development activities necessary to successfully sustain software projects, the URSSI conceptualization project will conduct a set of ethnographic studies. These studies will focus on generating rich, descriptive explanations about both the successes and failures of particular scientific communities that have been maintaining shared software over time. Data and findings from these studies will inform the URSSI's design and feed back into planned services and coordination with other NSF Sustainable Investments Institutes.

Community Engagement Activities

We plan to use six complementary and interconnected activities to engage the research software user and developer community: the community workshops and the community survey described above plus the following:

- *Web presence.* A web presence is crucial for a successful community-based project. The URSSI website contains information on the goals of the institute conceptualization project, along with a copy of the conceptualization proposal. We will continue to add information to the website over the project's course.
- *Blog.* The website will also host a blog where we will publish content from the conceptualization community as well as from others interested in this topic.
- *Reports and data.* Each workshop and the survey will result in a large amount of data and a curated report. We will make this information available on the project website.
- *GitHub repository.* To allow members of the community to interact with the project, both during and after the face-to-face workshops, we have set up a public GitHub repository (https://github.com/si2-urssi/).

The PIs' and senior personnel's networks provide us with a good starting point for engaging with the relevant communities. To ensure we address underrepresented groups in STEM (science, technology, engineering, and mathematics), we are reaching out to organizations already involved with XSEDE and the SGCI, which both focus on women and minorities; organizations with minority membership such as the National Medical Association; and chairpersons of relevant departments at the more than 100 US historically black colleges and universities. The community engagement activities will extend and formalize the community interested in the goals and vision of a URSSI, and will strengthen the final URSSI concept.

WHAT IMPACT DO WE EXPECT?

If this planning process is successful, and an institute (hypothetically also called URSSI) is created and operates for some years, we expect that the research software world will change as a consequence, with the impact felt in at least three ways. The first will be through direct participation with research software teams, resulting in

- a set of software across a variety of fields that has been improved by URSSI consultants and URSSI-trained developers; such packages will acknowledge these efforts and the URSSI contributions will be seen as a "quality stamp" for these packages; and
- software projects advised through URSSI reaching higher levels of sustainability, as measured by increased engagement with research communities, increased measured usage and citation, and expanded development communities (development and maintenance supported by more individuals, at a larger number of institutions and organizations).

Second, the general services offered by a potential institute will provide resources to the entire research software community:

- Best practices for software as a service for research software that will be established and disseminated to the research community. This might include greater understanding of industry practices as well as how these need to be adapted and modified for research software and the development models commonly used for research software.
- Workforce development activities that will result in an increased number of highly skilled developers who are working on research software systems and increased diversity in research domains, education, gender, and ethnicity.
- Software training becoming common in many departments in universities, in the same way lab training is now common, and more universities offering software training in a non-curricular method. URSSI will be aligned with Software and Data Carpentry and other "carpentry" efforts to promote the training and shared experiences needed for successful projects of all scales, not just successful individual researchers. URSSI will be the

"host of last resort" for synchronous and asynchronous training material, and will work with large disciplinary communities to customize content for them. Almost all such material will be openly shared, via CC-BY or more permissive licenses.

Third, as a formal institute and through its involvement with the research software community, URSSI will be an important voice on issues regarding the elevation of research software as a recognized intellectual contribution. Specifically, we hope that outcomes of our work will be as follows:

- *Software will be discussed more and will receive increased attention.* Specifically, more NSF announcements and solicitations will be aware that software is a key element of research, and will have instructions for proposers to describe software and its disposition, and for reviewers to judge it. Additionally, we expect that experts in all aspects of the software lifecycle will be invited to participate in review of proposals. We expect similar improvements in DOE, NIH, and private foundation opportunities. URSSI will be seen as the central US entity that promotes and improves research software. The URSSI website will be a central place for researchers to discover the latest tools and newest opinions on the state of research software, and URSSI-organized and URSSI cosponsored events will be where research software developers meet. Members of the URSSI staff will be invited keynote speakers at a variety of general and discipline-specific conferences and workshops.
- *Software will be seen as a valid research product.* Software metrics will be an accepted part of faculty hiring and promotion decisions in universities, and more universities will provide opportunities for software professionals to build careers. URSSI-suggested language will be commonly found in recommendation letters. Software will be viewed in parallel with data and other nontraditional scholarly products.

HOW CAN OTHERS PARTICIPATE?

Our goal is to interact with as much of the scientific software community as possible. Therefore, your input and participation are both welcome and desired. We have identified a number of ways that you can participate in URSSI. First, you can take the community survey, which we will make available on our project website once it is ready for distribution. Second, you can join our mailing list to be notified of upcoming events (see details on our website). Third, you can participate in our focused workshops. If you are interested in participating in these workshops or in other activities, please see our website and GitHub repository, or contact us at contact@urssi.us.

REFERENCES

1. U. Nangia and D.S. Katz, "Understanding Software in Research: Initial Results from Examining Nature and a Call for Collaboration," *Proc. Workshop on Sustainable Software for Science: Practice and Experiences* (WSSSPE5.2), 2017; doi.org/10.1109/eScience.2017.78.
2. I. Foster, "2020 Computing: A Two-Way Street to Science's Future," *Nature*, vol. 440, no. 7083, 2006, p. 419.
3. L.N. Joppa et al., "Troubling Trends in Scientific Software Use," *Science*, vol. 340, no. 6134, 2013, pp. 814–815.
4. J. Howison et al., "Understanding the Scientific Software Ecosystem and Its Impact: Current and Future Measures," *Research Evaluation*, vol. 24, no. 4, 2015, p. 454.
5. D.S. Katz et al., "Summary of the First Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE1)," *J. Open Research Software*, vol. 2, no. 1, 2014, p. e6.

ABOUT THE AUTHORS

Jeffrey C. Carver is a professor in the Computer Science Department at the University of Alabama. Contact him at carver@cs.ua.edu.

Sandra Gesing is a research assistant professor at the University of Notre Dame. Contact her at sandra.gesing@nd.edu.

Daniel S. Katz is the assistant director for Scientific Software and Applications, National Center for Supercomputing Applications; and a research associate professor at the University of Illinois Urbana–Champaign. Contact him at d.katz@ieee.org.

Karthik Ram is a Data Science Fellow in the Berkeley Institute for Data Science at the University of California, Berkeley. Contact him at karthik.ram@berkeley.edu.

Nicholas Weber is an assistant professor at the University of Washington. Contact him at nmweber@uw.edu.