Sustainability in Research-Driven Open Source Software

Danielle Robinson, PhD and Joe Hand

Code for Science & Society

May 2019

Executive Summary

Research-driven open source software (OSS) is a critical component of the scientific research process and has a broad usage in many data-driven industries. Generally, this software is open source software developed by, maintained by, having contributors from, and/or used by the research community. This includes projects developed for one paper or research lab and projects where researchers are one of many stakeholders. The research OSS community has unique constraints and stakeholders, placing it in a challenging position relative to the broader OSS ecosystem. To build capacity and sustain OSS communities serving the research, we recommend the following:

Growth strategy program for research-driven OSS projects

- **Rationale:** Transitions between project phase (such as those identified in this post) provide key points of leverage to impact project sustainability.
- **Approach:** Provide focused support for projects in transition phases or to support transition through an incubator-style program, combining work sprints with supported implementation and iteration of new practices.
- **Outcomes:** Increased project transparency, successful implementation or maturation of governance models, increased diversity in revenue streams, reduced burnout among team members and leaders, increase in the size and diversity of contributor communities

OSS leadership development program

- **Rationale:** Many small OSS projects underlie our collective scientific knowledge. The longevity of these projects is threatened by project leaders lacking time and resources for mentorship and lack of training and attention to OSS best practices.
- **Approach:** Develop support systems for projects with fewer contributors, where many early-career researchers build skills.

- **Outcomes:** Stronger project leadership, reduced burnout, new leaders who are supported by the OSS community.

Advocacy skill development program

- **Rationale:** The support of institutions (academic, industry, and government) is critical for the future of research-driven OSS.
- **Approach:** Advocacy skill development programs - short term or cohort-based - targeted at developing specific skill sets needed to advocate for sustaining research-driven OSS projects in complex research-focused environments.
- **Outcomes:** Researcher advocates for OSS investment within institutions, increased partnerships with and investment in OSS by institutions.

This report documents our survey of people involved in research-driven OSS, describes key challenges encountered by people working in the space, and identifies where investment can meet immediate needs in the OSS for research community. The increasing importance of software in reproducible research burdens researchers trained in scientific disciplines to produce lasting software products. Supporting and developing the communities building OSS will create capacity to support the future of data-driven research.

1 Background

1.1 Open Source Software for Research: An ecosystem in transition

Over the last decade, research has become increasingly data intensive. In this context, the data science skills that scientists need have grown dramatically (Leeming 2016). At the same time, the field of data science and the development of open source software (OSS) for research has expanded significantly. The growth of OSS in scientific communities has been supported by programs from a variety of organizations. These organizations such as The Carpentries, ROpenSci, Mozilla, and Moore-Sloan Data Science Environments have focused on outreach, OSS and data science education, and upskilling communities to use research-driven OSS and contribute back to OSS. This has brought more people, more diversity, and accelerated scientific progress across fields.

As the use of OSS in research has grown, the research-driven OSS ecosystem has also matured across scientific disciplines. We define the research-driven OSS ecosystem as the projects and organizations that build and maintain software, or support that work through advocacy, administration, training, and community engagement. While free and open source software was popularized with the launch of GNU in 1983, the 2000s saw an expansion of tools designed to handle data which form the foundation of today's data science toolbox. Older projects in this ecosystem are between 10-20 years old, including SciPy (2001), Matplotlib (2003), scikit-learn (2007), ggplot2 (2007), pandas (2008). Newer projects like Julia (2012) RStudio (2011), and Jupyter (2014) are all less than ten years old.

Each project has a unique culture, roadmap, and means of supporting its operations. Organizations that support scientific OSS through training, administration of OSS projects, and scientific software advocacy have developed over the last 10 years as the ecosystem has matured; these include: UK Software Sustainability Institute (2010), ROpenSci (2011), The Carpentries (2012), NumFocus (2012), Mozilla Science (2013), Berkeley Institute for Data Science (2013), Code for Science & Society (2016), and US Research Software Sustainability Institute (2017). (Dates reflect the dates of first release for software or organizational incorporation as listed on Wikipedia). The existence of multiple organizations that support open source data science software development, training, administration, and advocacy reflects concerns about the sustainability of free and open source models of software development across domains (Oberhaus 2019; Eghbal 2016).

Organizations supporting OSS for research have developed varied structures and focus areas, reflecting the vast set of skills required to support the ecosystem. Supporting organizations include nonprofit foundations (NumFOCUS, Code for Science & Society, Mozilla Science Lab), educational and training nonprofits (The Carpentries, ROpenSci), government-funded institutes (SSI, URSSI), and academically affiliated institutes (Berkeley Institute for Data Science (BIDS), and other Moore-Sloan Data Science Environments). These groups have built and maintained software (ROpenSci, BIDS), developed leaders through individual fellowships (SSI, Mozilla Science, NumFOCUS), and operated administrative infrastructure to

allow projects to receive grants and donations (NumFOCUS, Code for Science & Society, and institutionally affiliated organizations like BIDS).

Today, this ecosystem is in transition. The way research-driven OSS is used, shared, and relied on in new contexts has evolved. As the use of OSS expands across industries (Eghbal 2016; Oberhaus 2019), sustainability is an issue across all areas of OSS. Small research-focused OSS projects face industry's growing appetite for data-intensive work, friction in the academic labor market, and a growing backlog of OSS maintenance. This has created a landscape where research-driven OSS, often developed in academic contexts, must manage large communities with small teams whose primary aim is research. Tools developed for scientific use cases are valuable for business contexts, but small teams need support to navigate if, how, or when to develop sustainability models. Philanthropy, corporate support, government grants, and the labor of scientists are all needed to see the research-driven OSS ecosystem through the next five years. But project leaders have little stability, and less time, to focus on long term development.

In this changing landscape — what will prove sustainable in the next 20 years? How can the projects and organizations emerge from today's transition with the resources needed to prioritize innovation in data science for the pursuit of research? How can scientific and public interest communities grow on constrained resources? Does this come down to stable funding and staffing, a more developed community, and/or leadership development? In this report, we approach these questions through surveying and interviewing the research OSS community to better understand community attitudes towards sustainblity in the space.

1.1.1 Goals of this work

This report contains perspectives from interviews with people working in a variety of roles in research-driven OSS, including graduate students, researchers, full time OSS developers, and OSS contributors based at corporations. We pursued this project to give the people building, maintaining, and supporting research-driven OSS the opportunity to voice their concerns about sustainability. As a nonprofit organization supporting a variety of OSS projects, CS& S has had the opportunity to work intensively with OSS through various stages of growth. This work will inform our path forward as we continue to develop programs that support OSS for science and in the public interest, and we hope it will add context and new voices to the conversations on OSS sustainability in research.

1.1.2 Defining Sustainability

As the research-driven OSS ecosystem has grown, a definition of sustainability has been difficult to establish. In fact, the term has been described as "*a complex and multifaceted concept that can be viewed from a variety of perspectives and can include a range of different dimensions and factors*" (Venters et al. 2014). The following works informed our understanding of OSS sustainability:

- "*Sustainability means the capacity of software to endure... [or] that the software will*

continue to be available in the future, on new platforms, meeting new needs" (Katz 2016)

- "Software that considers social, environmental, technical, and economic sustainability" (Lago et al. 2015)
- "*Software you use today will be available - and continue to be improved and supported - in the future*" (About | Software Sustainability Institute n.d.)
- *S*ustainability refers to "qualities of availability (available), extensibility (improved), and the maintainability (supported) of the software" (Venters et al. 2014)

Sustainability in research-driven OSS requires attention to constraints unique to research environments, that are often different than broader OSS ecosystem. Some of the unique conditions of the scientific software landscape include university affiliations supporting in-kind project contributors, grant funding, and the ability for a project with a very small community of users and contributors to become sustainable. By exploring the challenges of sustainability in research-driven OSS through a community-centered lens, we hope to add context to the conversation on the future of OSS generally, as well as the opportunities and challenges of building and maintaining OSS for research.

2 Our Approach

2.1 Survey

We surveyed people who contribute to and use OSS for research. This survey was open to the community for 30 days. Survey results are summarized in section 3. The survey began with text informing participants that data would be used by Code for Science & Society and NumFOCUS to better support OSS in research.

Interviewees were given the option of:

- going on the record (their name would be associated with the notes on the interview and quoted in this report
- anonymously on the record (no name would be associated with the interview notes)
- off the record (statements would inform the work, but no notes or quotes would be taken).

Participants were informed that this work was funded by a grant from the Gordon and Betty Moore Foundation grant to Code for Science & Society, as described in this post.

2.2 Interviews

Survey participants were invited to follow up with a 30 minute interview conducted by the authors. Findings from these interviews are summarized in section 3.

3 Findings

3.1 Respondent Involvement in OSS for Research

The survey was released openly and shared with communities including NumFOCUS, URSSI, and general open science community via Twitter, LinkedIn, and mailing lists. 63 respondents answered the survey, which was up for a month. This group gave detailed and uniques perspectives, and while this small number does not represent the entire community, we hope that these perspectives can add context to the larger conversation of research driven OSS sustainability.

What types of research-driven OSS involvement are represented? Survey respondents classified themselves as users of research-driven OSS (94%), contributors to and engaged in discussion on these projects (both 79%), maintainers of projects (63%), and leaders of projects (50%). Survey respondents were able to select multiple choices to best classify their involvement and follow up with free text to further describe their association. In the full text responses the following languages, packages, and projects were referenced: Python (10), Jupyter (8), R and RStudio (8), Yellowbrick (7), scikit-learn (4), PyMC3 (4), Pandas (3) Scipy(3) among others.

What professional roles are represented? Students, faculty, researchers, and other academic professionals made 57% of survey responses. Freelancers (13%), non-research industry and corporate roles (14%), and research-focused industry and corporate roles (10%) made the remainder of the responses.

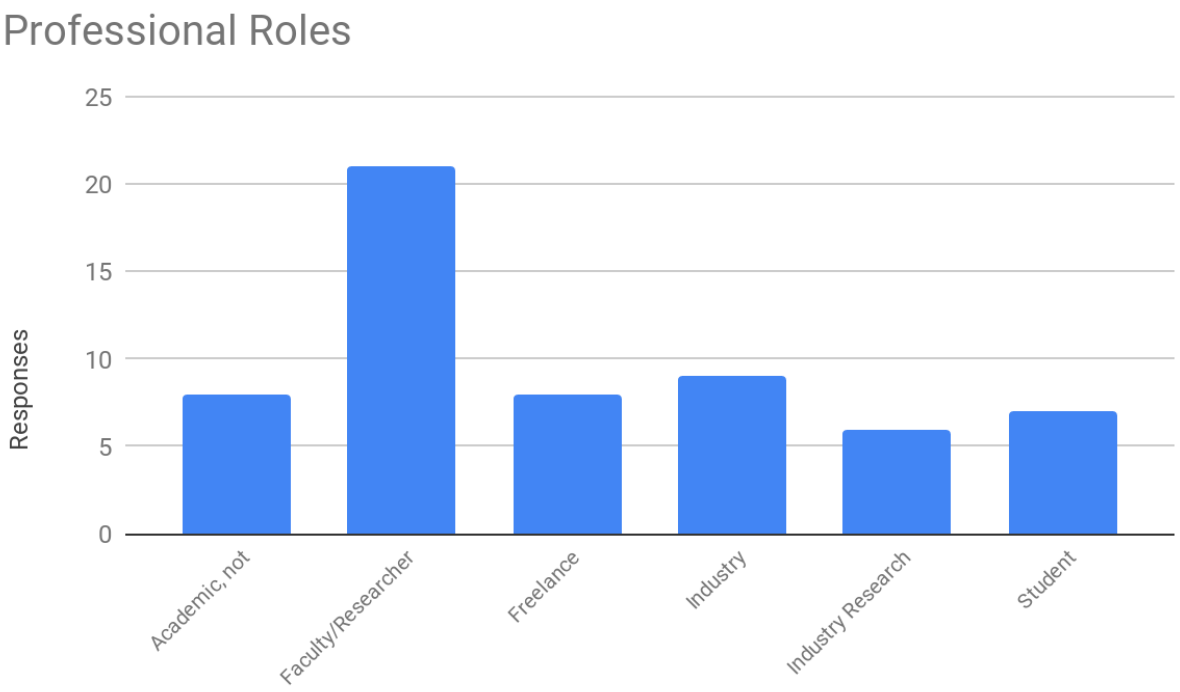

Figure 1. Professional roles of survey respondents. *From left to right: Faculty or other academic professional, not in research focused role; Faculty or researcher at an academic institution; Freelance or contract worker; Industry or corporate professional, not in research focused role; Researcher in industry or at a corporation; Student at an academic institution. $n = 63$.*

Paid or unpaid labor? There are many ways to make OSS part of a person's paid work. For example, researchers who develop software as part of pursuing a larger research question are paid by grants and fellowships. OSS software developers can support their work on research-focused OSS with grants and donations. And OSS contributors based in industry or corporations may contribute as a part of their full time jobs. Amongst our survey respondents, most were, in some way, paid for their work. Most are based at academic and research-focused jobs. For 59% of respondents, their involvement is part of their paid employment. For 30% or respondents, their involvement in OSS for research is not part of their paid work. The remaining 10% of respondents indicated that work on OSS is "sort of", "sometimes" part of their paid work, or offered further clarification on the relationship of open source work to their job. A small fraction, 8%, work on research-driven OSS over 30 hours per week. 27% spend between 11 and 30 hours a week on this work. 32% spend between 3 and 11 hours per week and 32% spend less than 2 hours.

Reflecting these ratios, of the people who are paid to work on research driven OSS, about 25% of them spend at least half of their time on this work, with 6% spending at least three-

quarters of their time on this work. Approximately 42% spend between more than 10% time but less than 50% on this work. And a third of respondents spend less than 10% of their time on open source work.

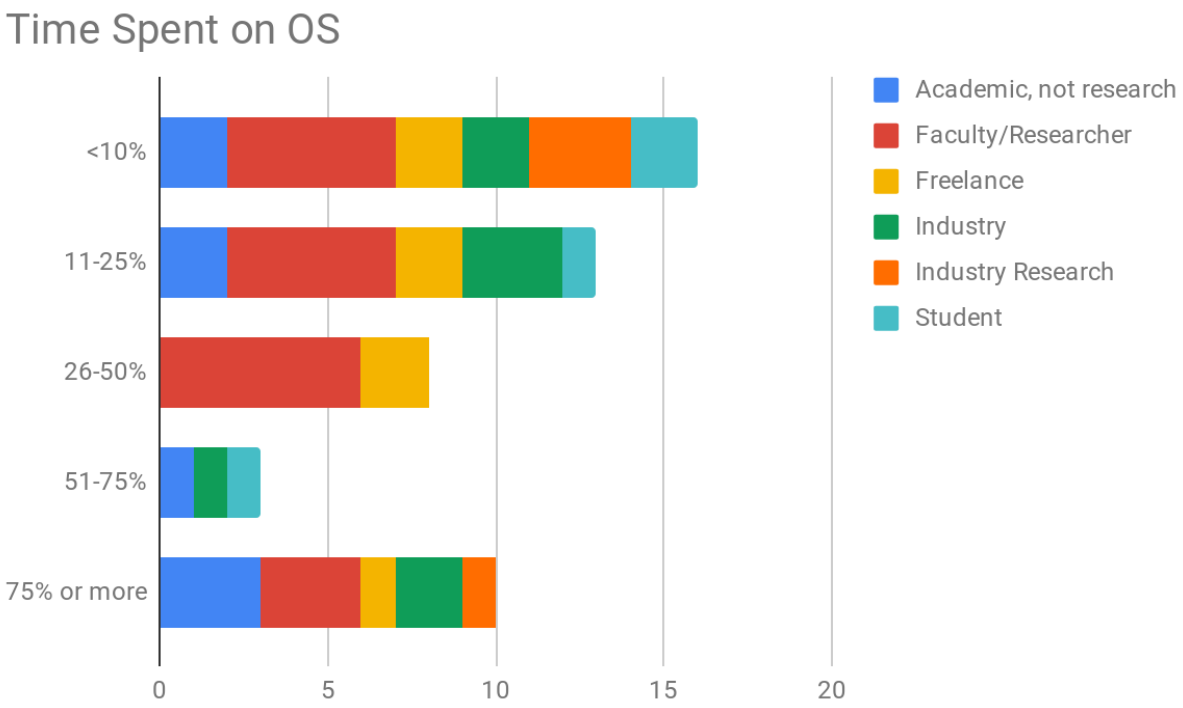

Figure 2. Time spent on Open Source Software by Job. *Breakdown of time spent on OSS, color coded by professional role.*

3.2 Sustainability Challenges

Survey participants were asked "Which of the following themes related to sustainability are current challenges to the projects you're involved with?" and offered ten themes, outlined below, as well as the option to write in a response.

- **Leadership** & **Team Management**: Project management, setting and sticking to clear goals and priorities, providing mentorship and direction to paid and volunteer contributors
- **Community Development** & **Management**: Bringing on new contributors, creating pathways to leadership, moderating community conversations, enforcing code of conduct
- **Communication**: Internal Communication includes the clarity and consistency, who is tasked with internal comms. External Communication includes how a project presents itself, who is tasked with external communications

- **Recognition**: Distribution of workload, avoiding burnout and maintainer fatigue, opportunities for professional advancement for both paid and volunteer contributors
- **Funding and Fundraising**: Business development, fundraising from grants and donations, securing stable funding and/or in-kind developer support
- **Stability and Maturity of Tool**: Maintenance of documentation and software, ability to quickly address issues
- **Governance Issues**: Decision making processes, steering of project priorities, transparency or opacity of project governance
- **Administrative Issues**: processing reimbursements, getting people paid, handling project finances, who is tasked with handling administrative tasks

Results provide a glimpse into the areas critical to project sustainability in the experience of our respondents. The top five themes that are current challenges to projects are: funding (64%), leadership (61%), community engagement and development (61%), burnout and recognition (59%), and stability of the tool (56%).

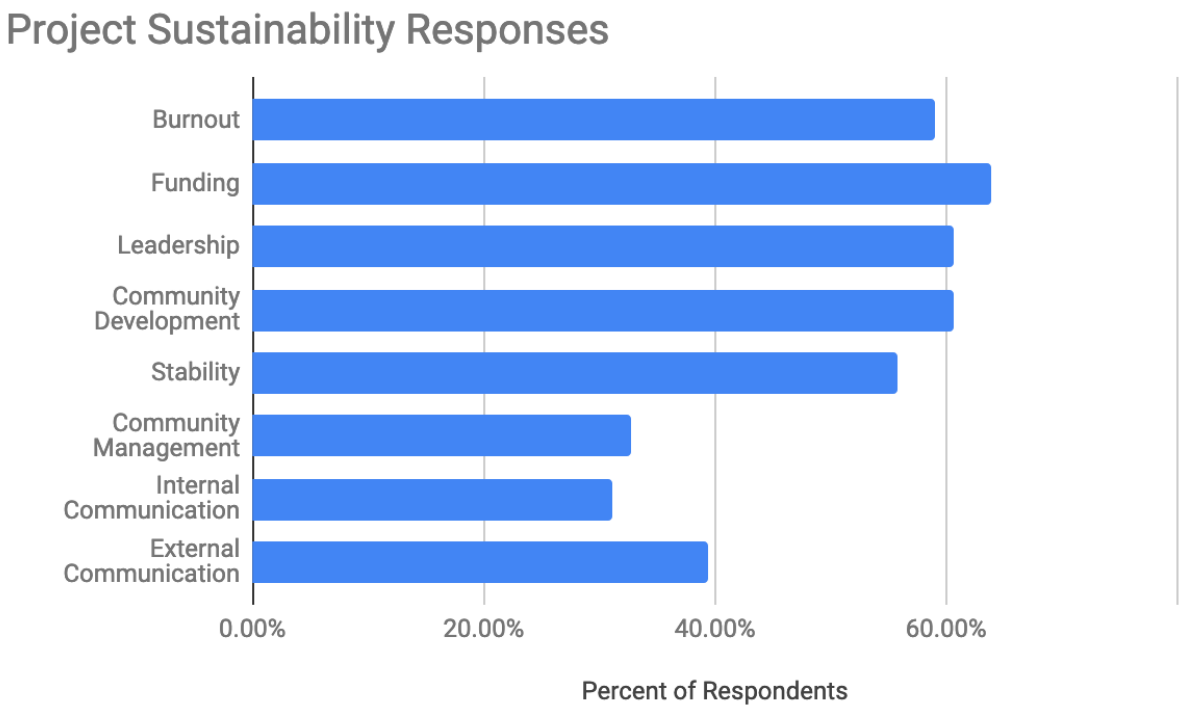

Figure 3. Challenges to Project Sustainability. *Current challenges to projects are: funding (64%), leadership (61%), community engagement and development (61%), burnout*

and recognition (59%), and stability of the tool (56%), external communication (39.3%), community management (32.8%), and internal communication (31.1%).

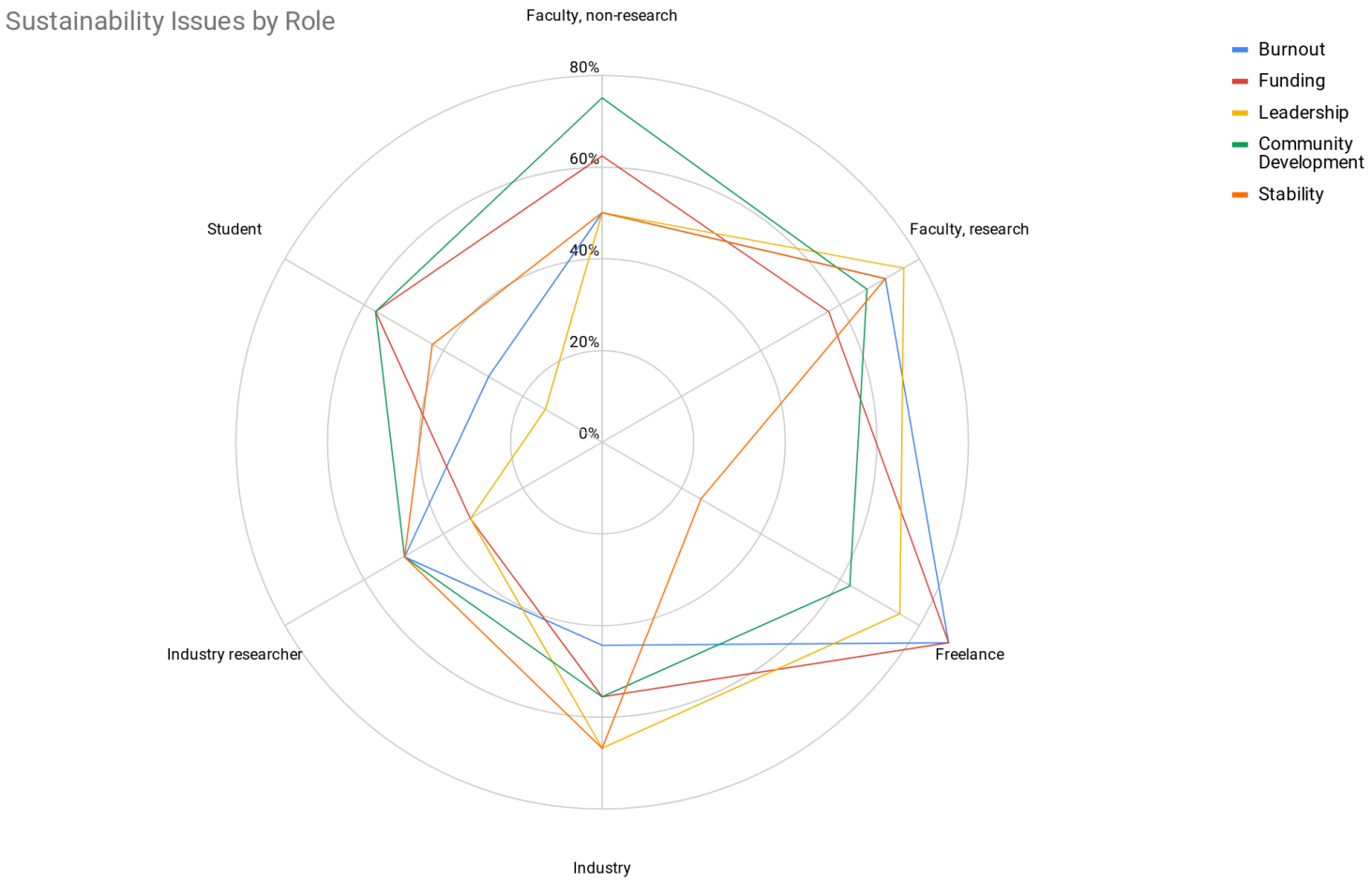

Figure 4. Sustainability Issues by Role. *Faculty and academics in research roles were most concerned about leadership, burnout and recognition, and stability of the tool. Faculty in non research roles were most concerned about community development and funding. The top concerns of freelancers were funding and burnout. People in industry flagged leadership and stability of tool (non-research industry roles), and leadership and community development (research roles in industry). Students' top concerns were funding and community development.*

To allow respondents to go further into this topic, this question was followed by two questions with free text answers. The first question, "*What issues related to open source data tool sustainability (defined as a project's capacity to operate stably) do you face that make it harder for you to do your research/work?*" targeted sustainability in their project or community. The second question, "*Please share your perspective on the biggest challenges facing open*

source data tool sustainability today" allowed for comments on the field more generally. In the free text responses, funding, leadership, and maintenance emerged as key themes.

Funding: Respondents focused on funding as a major sustainability challenge. Challenges cited include:

- Finding funding for maintenance: "*we find it hard to get money for maintenance*" , lack of "*access to capital to cover infrastructure cost and maintainer time*" - more on maintenance below.
- Paying market rates for developers: "*Constraints from University policies on pay*"
- Business development: "*conflicted about charging fees/subscriptions but are unsure how else to raise funds*"
- The tension between novelty and infrastructure: "*Funders want shiny and exciting which may be good for scientific research but is absolutely not appropriate for scientific software*" and "*funding is all around new features that we promised to funding agencies*"
- And the frank:"*How do we get the $ 1,000,000/year we could use to sustain our project?*"

Leadership: Mentoring, professionalization, strategy, and project management were all cited as issues within respondent's communities or issues in the field as a whole. Leadership related themes included:

- "*Making our project more professional*"
- Critiques of personal practices: "*(a) There are only 24 hours in the day; (b) I'm extremely disorganized*"
- Operational issues: "*At the moment administrative issues are the main bottleneck in our org*"
- Comments on project strategy: "*Vague long-term vision that leads to unclear ownership and imprecise scope, which in turn hurts fundraising and promotion efforts*"
- Leadership development and training: "*The biggest challenge is that people aren't trained to actually lead a project through different stages and scales*" and "*people who ∗start∗ open projects should not be the people who lead them, but so many others are perceived as grabbing an opportunity when they try to lead the project to the next stage.*"
- One comment, "*Bringing on core contributors that are true software developers,*" underscores the desire to recruit people who can contribute at a high level quickly. However the realities of the job market may make this difficult or impossible for projects without stable funding. Mentoring junior developers into these roles can fill this gap but may be perceived as another time-sink by project leadership.

Maintenance: Many people commented on maintenance. Maintenance is the process of an Open Source project ensuring existing software works as expected, community questions and bug reports are addressed, and documentation, websites, and other communication and software publishing sites continue to be available. Maintenance requires time, money, and commitment to the project. Finding the time or funding to perform maintenance is a clear issue for project leaders, as was mentoring others to help in this task.

- The challenge for some is finding the right people to bring into maintainer roles: "*Finding people to work on the fundamental tools*" and "*Finding and mentoring junior maintainers to replace senior ones.*"
- The issue for others is paying for maintenance: "*As developer of open source data tools the main issue is getting funding for software maintenance. Very few are ready to pay for that.*"
- For some, "*Maintaining [the project] is a near full time job, but it is not my actual full time job. I feel a lot of pressure to maintain the project or I will let people down.*"
- Practical outcomes of having a few maintainers, "*proportion of users to maintainers makes maintaining projects difficult*" and "*developers are overburdened with technical support work.*"
- One respondent encompassed the fears of many by articulating "*I just don't have enough time to do all the things that I want to do, and because they all require a level of expertise that is difficult to teach, I haven't been able to easily recruit people to continue my work*"

In summary, the responses show a community of committed OSS for research users and developers who are struggling with funding, leadership, and finding the resources to do critical project maintenance. While this survey is a small sample, we believe that the perspectives of different types of research-driven OSS community members perspectives are critical as we seek realistic sustainability strategies that will fit this diverse community.

3.3 Interviews

3.3.1 Professional roles and scientific disciplines of interviewees

35 people provided a contact email for a follow up interview, and of those, 21 have been completed. People interviewed include scientists at all levels, software developers, and more. Research areas of interviewees included astrophysics and astronomy, neuroscience, computer science, environmental science, and social science. I also spoke to people at large tech companies, nonprofits, and startups who use and/or develop research-driven OSS. Some people were early career, while others were at more advanced career stages and in leadership roles.

3.3.2 Perspectives of interviewees

Community members are in general agreement that there are transition points where critical issues including funding, leadership, burnout, community engagement, and tool maturity can conspire to hinder a project's overall stability. While there is no universal agreement on a unified solution, digging into these critical issues lead us to identify key areas for support to build an ecosystem of OSS for research that will meet the challenges of the next decade.

In this section, we discuss the themes raised by interviewees. These themes are:

- What is Sustainability?
- Leadership, Mentoring, and Project Management in Research Driven OSS
- The Threat of Burnout
- Long Tail Research: Small projects are a big part of the ecosystem
- Funding, Fundraising, and Business Development
- Resourcing Maintenance for the Long Term
- Relationship of Research-Driven OSS with Universities and Research Institutions
- Relationship of Research-Driven OSS with Corporations and Big Tech

What is Sustainability?

While there is much discussion of the importance of sustainability, there is little agreement on what it means. Is a sustainable OSS project a business? Funded by donors? Staffed by folks through in-kind support agreements? Are Jupyter, Dat, Yellowbrick, or Pandas sustainable - if so, why? There is no clear measure of a project to show that it is sustainable, and many projects will move in and out of phases where they are more or less sustainable.

"*What resources are required to run a project? If you know what you need - you can think about how to get it sustainability.*" - Abby Cabunoc Mayes

"*Sustainability means can a project do what it needs to do over a period of years.*" - Christie Bahlai

Acknowledging, and planning for, the transition points that a research-driven OSS project will move through over a period of years is the best any leader can do to improve project sustainability. Stages, phases, and transition points have been discussed as a framework for thinking about sustainability at meetings convened by the US Research Software Sustainability Institute (link to blog).

Addressing the risks of unsustainable projects, Konrad Hinsen's discussion of software collapse was summarized by Daniel S. Katz in our interview (link):

"*How do projects deal with all the changes that can happen underneath them? 4 models in how to build the building when the earth is shaking. Different options that the project can choose.*" - Daniel S. Katz

Leadership, Mentoring, and Project Management in OSS

"*Open source is unique and the best place to develop as a leader because of the openness and transparency.*" - Christopher Aedo

The discussion of resourcing maintenance and the research software engineer (RSE) career path lead into ideas from interviewees on how to best develop leaders in this field. Abby Cabunoc Mayes who runs Mozilla's Open Leadership Training (OLT) program encouraged people to take a realistic approach to the time it takes to develop leaders in a project:

"*It takes longer than a year to develop people to move into leadership... the hardest part is finding the right person and having the time to invest in mentorship.*" - Abby Cabunoc Mayes

This comment echoes back some of the survey answers where people highlighted the lack of experienced software developers who could take on maintenance and core project work. Mentorship takes time, and the work of Abby's Open Leadership Training (OLT) program has illustrated the impact of cohort based learning followed by a train-the-trainer approach where mentees move on to become mentors. While the work of running an OSS project is immediate, Abby suggests making an up-front ask for project leaders to invest in mentorship, perhaps two hours a week. Even a small shift in resource allocation like this can change the way a project engages with its community and create pathways to leadership that help avoid burnout and talent replacement problems.

Other interviewees focused on the management part of leadership. Again, echoing the free text answers, most project leaders are not trained managers nor even trained software developers. They are scientists, researchers, or data scientists first. However, in a small project everyone must wear multiple hats. Nathan Goldbaum pointed to issues around resourcing management, and called out the taxing nature of this work:

"*Lack of dedicated people to do project management stuff... it takes a lot of social effort. Making sure meetings happen, making sure GitHub issues are in a good state. Someone needs to make sure we all actually use Trello, or it becomes stale quickly.*" - Nathan Goldbaum

The work of running a project adds a layer of work that includes short term planning, day to day management of tasks, and long term project strategy. When leadership roles are volunteer, underpaid (or lacking the potential for promotion), or part-time it can add

instability to a project.

"*People are hungry for advancement.*" - Titus Brown

"*Money and leadership are related. You can't afford not to be paid.*" - Travis Oliphant

"*Room for leadership support in a dispersed community of people working at a fraction of an FTE (full time effort of an employee)?*" - Steve van Tuyl

Project leaders take on the hard work of project strategy and governance. On the topic of governance, interviewees highlighted the wide range of governance models out there and identified concerns about prioritizing governance work. Interviewees took the view that project governance models need to grow with a project to best support a growing community. Interviewees identified transition points - rapid growth of users, growth outside the domain of origin, and even leadership crisis - as events that can push a project to update their governance. When large projects operate with small-project governance models, interviewees noted that decision-making bottlenecks increase maintainer burnout and can slow technical progress.

"*It's important that people have the mental and financial time and space to think about governance. It's really hard to get people to focus on this.*" - Tania Allard

"*[some projects have] incredibly immature governance models... you need to think about governance - but can put it off when [the project is] well funded*" - Titus Brown

"*[I see an] older style of leadership in open source. Usually what this means is 1-3 people, usually men, usually white, who have worked their butts off to build a community or tool. They've grown the thing. But the fact that they were able to build the project means that they are not well suited to see it scale.*" - Anonymous

Interviewees identified the advancement and stability as an issue for career development within existing research organizations. As many interviewees were based in academia, there was discussion of the role people see OSS skills playing in their career. Some see OSS best practices as part of their road to running a research group:

"*Having good software is the craftsmanship of research. The bottom up approach, good software = good science. I want to run a research group. Developing this skill to improve my research and set up to run a research group.*" - Anonymous

But others want careers that focuses on software development for science:

"*First job as an RSE [research software engineer] I had a reckoning about the traditional academic path. I want to do the job of an RSE. But not to run a research group.*" - Tania Allard

Of the people who participated in the survey, most were affiliated with an academic institution. Building structures that support OSS leadership development in research environments is a key area for program development. These topics are explored in depth in the sections Relationship of Research-Driven OSS with Universities and Research Institutions and Relationship of Research Driven OSS with Corporations and Big Tech.

The Threat of Burnout

"*Another ambient threat is that the success of the project kills it.*" - Dietrich Ayala

"*[if your project is] sustained on money you have a burn rate - you know how long your money will last. But for projects that depend on people, it's the same but we call it burnout. How long will the people last? One way or another you spend them and need to replenish them (money or people).*" - Anonymous

When a small project becomes successful, the team will feel the increased workload immediately. For small teams where each person plays multiple roles, this can have a devastating impact, interviewees discussed the realities of working on small teams:

"*Too many jobs for one person.*" - Nathan Goldbaum

"*Reality is 1/2 time research science, 1/2 open source developer and community and infrastructure... no 100 % FTEs [full time equivalent staff] on the project*" - Joe Hamman

Benjamin Bengfort is co-founder and mainter of Yellowbrick, an affiliated project with NumFocus. It is an open source, Python project that extends the scikit-learn API with visual analysis and diagnostic tools. He summed up issues with burnout in the following way:

"*When I think about what's happening with open source now - research tools critical to academia like scikit-learn and numpy depend on the mental health of a handful of people who are doing it for love and not money or recognition. That's not sustainable.*" - Benjamin Bengfort

Long Tail Research: Small projects are a big part of the ecosystem

On the topic of work that support a single publication but are unlikely to develop a community of users, interviewees pointed out that many scientists learn data skills and open source best practices in these small projects. This is particularly true in today's labor market, where many early career researchers do not have exposure to data science education before their work demands it.

"*I am recruiting students with no software development experience and I train them to be data scientists. The best thing that a funder could do would be to support me to upskill my students - Carpentries, ROpenSci are options but my students don't see themselves there yet.*

If I had time I would run a local cohort based thing, like Mozilla study groups or a study group for a class." - Christie Bahlai

Others expressed the need for core facility or permanent faculty research software engineer (RSE) support for long tail research. This could reduce the need for everyone to become an expert, allowing domain experts to focus on their scientific questions and collaborate with data science and software experts where needed. This concept came up as a solution in many discussions:

"*Starting to bat around the idea of a coordinated open source data science institute - an org within NCAR that serves the general population of researchers*" - Joe Hamman

"*The long tail researchers [are] not going to be able to develop or refactor software. RSEs that are underwritten by someone and can work to develop software with the long tail of research... create a pool of support.*" - Tania Allard

This model is being tested at institutions like the Turing Institute. Predictably, some interviewees wanted to see the data on how these research software cores function:

"*How well do existing RSE positions/centers work? Are there metrics that demonstrate value and how do they compare against other types of institutional hiring?*" - Hao Ye

Students and early career researchers are the key group to be trained, particularly in long tail research contexts. Early career exposure to OSS best practices had a large positive impact on many interviewees. However, interviewees believe that institutions are not set up to do this work in the near-term. Interviewees cited The Carpentries, OLT, locally operating interest groups as making a positive impact in the space. Despite the work of these groups, they called for more support for long tail researchers looking to do quality data science with OSS. Interviewees discussed their experiences, both those who don't currently contribute to big projects and those who see it as their duty to mentor others to contribute:

"*I do use git to maintain my own work, but not to enter a large community with people I don't know. It is intimidating.*" - Anonymous

"*A lot of people consume open source software products and don't think that they could become a contributor. Once you get people over that barrier, it's a huge shift. It's exciting. We don't expect people to be open source savvy because it's easily trainable.*" - Christopher Aedo

Fundraising, and Business Development

"*A lot of money gets spent on software - it is not always flowing down to support projects.*" - Travis Oliphant

Most interviewees devoted some part of their time to funding, fundraising, and how to

appropriately obtain resources for a project.

Some interviewees focused on business development, academic institutions, corporations, and other potential client bases. Many people identified the difficulty of working in a small research community where a tool is developed to meet a need, but not a market. People also discussed wanting to devote time to business development, but having so much other work that it didn't seem possible. Others expressed little to no interest in developing paid services, some citing concerns about how that might move the project away from its core goals. These interviewees preferred to maintain the work through volunteer and in-kind labor to keep it centered on its mission, even if it was a lot of work to manage that labor.

While some people focused on building their teams with in-kind support from academia, corporate, and occasional contract labor, others saw the need to build up organizational infrastructure around their projects by developing products and services for sale. For a set of projects in our space, moving along a product development route is the way to handle a scaling community and limited research funding. In keeping with the open philosophy, the business models interviewees discussed focused on services, analytics, and expertise that could be layered on top of a core OSS product. One interviewee neatly summed up the challenges of both moving from a tool to a product and from tool builder to market builder:

"*You can build a tool and it'll be a helpful technical solution to the problem. Let's think of a hammer as an example tool. A product is not a tool, it's quite different. If the hammer is a product, it needs to distinguish itself among a market of other hammers. This helped me see the difference between a tool and a product. Tools are not products. We had been building tools, but we didn't know anything about products.*" - Anonymous

"*You have to be able to sell something to someone with money... We don't have a market for what we are doing. We have had to engineer the cr∗p out of a market. We needed to come to where people are, understand them. And it took years.*" - Anonymous

Others focused on building corporate support through in-kind support, donations, and contracts for research-driven OSS. As discussed in Relationship of Research-Driven OSS with Corporations and Big Tech, much of the corporate support that flows in to OSS for research is staff time. Interviewees identified OpenCollective, TideLift, and other platforms that facilitate donations as valuable services to help showcase which corporations are funding OSS. Travis Oliphant expressed a common theme - leveraging the optics of sustainable use of OSS:

"*Certification to companies that are using open source sustainably... Are you just a user? Are you a sustainability champion?*" - Travis Oliphant

Resourcing Maintenance for the Long Term

"*Research software engineer contracts are 1-3 years. Software needs to take a 5-10 year view.*" - Tania Allard

Interviewees took a long-term view when thinking about the sustainability of research-driven OSS, however many cited frustration about the challenges of working on a five to ten year plan with one to three years of funding. Nearly all cited maintenance as a concern. Joe Hamman summed up the issue from a scientist's perspective:

"*How do we fund short term dev and sustainable maintenance? [If we] fund some open source software development for 2-3 years. There's no plan beyond that ... What happens at the end of a grant? Will community step in? Things that are popular and useful tend to find sustainability via in-kind work from scientists.*" - Joe Hamman

Hamman mentions in-kind support here, which is a critical resourcing mechanism for many projects. In-kind support is support given to a research-driven OSS project in the form of staff time from a university, corporation, or other organization. This support was raised as both a solution and a problem by interviewees. In-kind support often means that a person will be balancing other responsibilities. But many interviewees involved in scientific OSS expressed their need for full time developer work (or even short sprints of full time work) on their projects.

Stable, supported access to research software engineer (RSE) expertise is a model that institutions including the Turing Institute, Allen Institute for Brain Science, Chan-Zuckerberg's Biohub, Moore-Sloan Data Science Environments and others have piloted with success. However, embedded or otherwise accessible data science teams are not present at most institutions. Stability was cited as the number one concern of one interviewee, Tania Allard, who recently left an academic RSA role for a big tech company. She commented that it is nearly impossible to prioritize a project's long term needs on short contracts with unstable funding, stating:

"*The most successful teams have underwritten research software engineers - no worries about contracts running out. Can prioritize long term, mid term, short term projects.*" - Tania Allard

Relationship of Research-Driven OSS with Universities and Research Institutions

Many of our interviewees are in some way affiliated with academic institutions. They articulated changes they'd like to see in the way that academic institutions support OSS for research. One key issue is incentives and credit for research software and data science work that will be recognized for academic advancement:

"*There's a lot more going on as an academic that's done to direct research efforts - RFPs, conceptual ideas - where does research code tool building and sustainability fall in? It doesn't fit into existing slots in the system.*" - Hao Ye

"*What I do in software is not what I am evaluated on for my job [as a tenure track faculty*

member]... Institutions have trouble classifying data science." - Christie Bahlai

There is an opportunity to continue to advocate for stable, institutionally supported research software engineer positions and/or faculty who develop and contribute to OSS for research. Titus Brown suggests adapting the Cooperative Extension System (CES) model for software professionals based at universities. The CES model has created a home for agricultural specialists, outreach, and training in the university system. Titus envisions that a CES model would create:

"*Community of practice - extension specialists evaluating each other ... This should gel well with the stated mission of a university ... [and provide] sustainability mechanisms and admin infrastructure*" - Titus Brown

An entry point for this idea is Titus' OSS sustainability and Ostrom's CPR design guidelines. Reconceiving how software expertise can be housed in a university is a worthwhile endeavor. While some privately funded entities, such as those listed above, are able to pilot models more quickly, developing long-term plans with the support of government agencies and institutions (such as URSSI is doing) is a more stable way to resource maintenance on research-driven OSS.

Another issue raised is funding. Many interviewees articulated the wish that their institution would stably resource data science, whether by developing an RSE center or by directly supporting the tools.

"*Most of the users are from universities. But [project] never had a donation from a university - why?*" - Francesc Alted

"*If institutions can not support open tools - the tools will die.*" - Steve Van Tuyl

Some interviewees expressed concern about university procurement procedures. In other words, it is still easier for an institution to buy an institutional software licenses than it is for the university to making a donation to a software foundation. Some expressed in interest in navigating university procurement, while others expressed frustration that their university was paying to license proprietary software rather than paying into the open source, mission-aligned projects.

Relationship of Research-Driven OSS with Corporations and Big Tech

As the business value of data continues to grow, OSS developed in academia is now used more widely. We were able to speak with people in corporations and tech who use OSS data tools and contribute back. They described their motivations for finding ways to contribute and/or fund to OSS data tools developed in science through their work:

"*These [projects] need to be funded now in case we need them in 5-10 years.*" - Ricardo

Mendez

"*I do this because all the really really great career things that people in open source did for me. My career is where it is because of people in OS.*" - Christopher Aedo

"*Companies are seeing a shift where individual devs (data science, engineers, etc) have a lot more influence over purchase decisions - no longer an executive and a sales rep making a purchasing decision. It's important for a company like IBM to talk to developers, both internal advocacy and external.*" - Bradley Holt

They also commented on the constraints that they face working within a company that supports open source but ultimately needs to prioritize its business:

"*We do have support from the company to work upstream and commit back to open source but it's narrowly focused... internal resource constraints. We'd like to commit more time. But we don't have time*" - Christopher Aedo

"*We fix this bug because we use this thing.*" - Dietrich Ayala

"*Ultimately, isn't this the issue we all face (in academia or in business)? If we're donating resources in a way that is undervalued and invisible, we'll always run up against the need to up-prioritize the business (be it profit-making or getting tenure). How can we talk about OSS work in a way that makes it* ***part of*** *the business, rather than some sidecar gig?*" - Steve Van Tuyl

Interviewees based at corporations also discussed the utility of the software foundation, as a way for corporations to indirectly collaborate. Funding via software foundations and coordination in-kind labor support is a key part of sustainability for large projects:

"*Only sustainable open source at any scale is when companies are working together to fill the gaps. This is when needs can be met. You can collaborate via a software foundation.*" - Bradley Holt

All the people we spoke to people at corporations who make contributions themselves, or run teams that make contributions to OSS. While in-kind contributions were viewed as both a blessing and a curse by project leaders we spoke to, Dietrich (formerly of Mozilla Corp, now at Protocol Labs) points out that the monetary value of these in-kind contributions is often significant:

"*If usage of open source software was measured at large corporation and hours were tracked and you were paid for those hours using a very traditional model, [it would be similar to] paying for software as services rendered.*" - Dietrich Ayala

Corporate support of in-kind labor can be quite long-lived, with some interviewees citing open source projects they've been involved in for ten years.

Ricardo J. Méndez of SamsungNEXT took a unique perspective on OS. He actively advocates for his company to give unrestricted donations to open source projects that he believes are important. He prefers to avoid a contract with strict deliverables, and rather focus on giving creative people the freedom to continue their work, on tools he believes will be important in the next ten years.

"*I am a big fan of open source ... stuff that's not easy to fund. You won't get money back in a year by investing. But these are things we as a society need to invest on. If we push people to come up with a business model too early, it's a big loss.*" - Ricardo Mendez

When asked about why they work open source, people in corporate roles said that it comes down to attracting talent. If people can see the quality of your work and your team's style it raises the profile of the organization, as Christopher Aedo of Teradata put it:

"*We're growing our team... [we get] more impact from being a good OSS citizen to attract talent.... A dev will look at actual code and interact [with the team].*" - Christopher Aedo

3.4 Key Takeaways

3.4.1 Leadership Development

- OSS leadership is a unique skill, but one that can be learned
- At different levels of OSS leadership, needs are different. For example:
    - A developing leader who may be taking ownership of one aspect of a larger OSS project, learning to wrangle contributors, learning to lead in OSS comms channels
    - An existing project leader who may be handling the growth of a quickly scaling software project, managing a team, fundraising

3.4.2 Burnout

- Prioritize time for leadership development and mentoring.
- Make space for project leaders to talk honestly about challenges.
- Support for career development and advancement within OSS communities and governance structures.
- Value OSS participation in a way that supports career development and advancement at institutions (academic, industry, government) where OSS contributors are employed.

3.4.3 Long Tail Research

- Long tail research is a big part of the scientific OSS enterprise.
- It is a key data science training ground.

- Early career exposure to OSS best practices can have large positive impact on a person's career and skills as a data scientist.
- There are unmet needs around basic data science and open source training for early career researchers.
- Developing RSE core facilities is under discussion but information and case studies on how to resource, staff, and develop these facilities will be useful to the field.

3.4.4 Funding and Business Development

- Several interviewees identified *time* to focus on funding and business development as a key issue, specifically:
    - time to step back and assess the best path forward with their community.
    - time to investigate new sources of funding.
    - time to develop and iterate a business model.
- The level of support from academic and corporate users of OSS is a continuing frustration.
- Many in research-driven OSS sector operate on small teams with no internal experience developing, testing, and iterating models.
- Role of philanthropy:
    - Philanthropic grants are seen as a launch-pad, but in practice interviewees had difficulty building business models on short timelines.
    - Philanthropic grants are also seen as a way to develop OSS that will later be supported by a federal agency or institution, but in practice transferring responsibility from a funder to an institution is not straightforward.
    - Grants focused on business development, to free-up staff time to devote to this and other strategic issues, are requested by some interviewees who are struggling with how to grow.

3.4.5 Maintenance

- It is hard to find funding for maintenance.
- It is also hard to find people at the right skill level to do maintenance work.
- Maintenance work is often done by short contracts, this is a source of frustration as it makes long term planning difficult.
- Strategic advocacy at institutions could help support software maintenance through stably funded positions.

- Maintenance is tied in to fundamental issues in the space and may serve as a barometer for the health of a project, or the space more generally. It is linked to:
    - Leadership and burnout: mentoring others to do maintenance to reduce burnout
    - Funding: ability to resource the work
    - Community engagement: the relationship between users (including institutions/corporations) and contributors/maintainers

3.4.6 Relationships with Academic Institutions

- Academic institutions contribute to OSS for research through the in-kind labor of their faculty, staff, and students who choose to contribute to OSS.
- However, OSS labor is rarely valued within academic systems (like promotion processes), and may even be seen as a drain on resources.
- Development of research-driven OSS aligns well with university missions, but more work is needed to see that expand to financial and structural support.
- There is frustration that academic institutions are major users of research-driven OSS but not generally major funders.

3.4.7 Working with Corporations and Big Tech

- Champions at corporations will lobby for in-kind and monetary support to research-driven OSS projects.
- Corporations will work on what they use, large or small projects.
- Long, stable in-kind relationships with corporations can be a part of project sustainability plans.
- Corporations see good OSS citizenship as a way to recruit talent.

4 Recommended Work Areas for Immediate Impact

4.0.1 Growth strategy program for research-driven OSS projects

Rationale: Organic growth of a project can be a boon, bringing new visibility and users to the project. Unfortunately, it can also overburden existing resources and project structures. At these moments, research-driven OSS projects need to step back and articulate what their long-term strategy is for the project. But it is at these moments that this is the hardest to do, because unexpected growth and attention can feel extremely urgent. This growth without strategy leads projects to an unsteady state, quickly leading to burnout.

In our interviews and conversations on sustainability, one theme repeatedly came up - project transitions are a key point of potential failure and thus an opportunity for positive intervention.

Project transitions take many forms. There are positive transitions, such as growth in the user community pushing a project to rethink governance. This type of transition is an opportunity to improve and mature the project. There are also difficult transitions, such as the loss of a person in leadership or the end of a grant. These challenges are also opportunities to do things differently to set up structures to avoid relying too much on one person or funding source.

To understand these transition points, the identification of project archetypes has been indispensable. URSSI and other communities have developed several archetypes to build common understanding and communication around OSS research projects (Mozilla 2018; Benthall 2019).

Ultimately, there are many directions that a research-driven OSS project can take to achieve a stable state. Unlike a typical, non-academic software startup, growth and raising investment funds are not priorities in research-driven OSS. Instead, leaders are looking for space, time, and network connections to support them as they move their project along its roadmap. Some roadmaps follow a traditional product and business development route, while other projects plan to stay small and operate to serve a distinct scientific community. However, project leaders identified time and lack of deep experience in certain areas (governance, business development) as key blockers to successfully moving research-driven OSS projects through these transitions.

Approach: We recommend leveraging transition points, through an incubator-style growth strategy program designed to quickly and confidently take projects to a new stage.

To provide focused support for OSS for research projects, such a program will be:

- Specifically focus on research-driven OSS projects
- Structured in a way that combines focused work sprints with supported implementation and iteration of new practices
    - Multiple in-person work sprints that enable project leaders to step back and focus on project strategy and developing new models (business, fundraising, governance)
    - Supported by remote programming that focuses on implementation (and iteration) of organizational changes
- Focused on projects in transition and/or growth phases, rather than early stage business development (as is common for traditional incubators)

Outcomes: The goals for impact on participating research-driven OSS projects include:

- Increased organizational transparency
- Successful implementation or maturation of governance models
- Increased diversity in revenue streams
- Reduced burnout among team members and leaders
- Increase in the size and diversity of contributor communities

4.0.2 OSS Leadership Development

Rationale: Many small OSS projects underlie our collective scientific knowledge. They are threatened by overburdened project leaders lacking time and resources for mentorship and attention to OSS best practices.

Throughout our research, today's leaders in OSS research projects expressed concern about the software and data science skills gap they encounter when seeking new contributors or working with early career researchers. They cited both the need for skilled people to contribute to research-driven OSS and the need for early career scientists to learn and implement OSS best practices in research. Despite the need, these leaders highlighted the difficulty in finding time for mentorship of early career researchers on OSS.

Long tail research software projects, research-driven OSS projects that are one-off or those with very small communities, are many scientist's first introduction to OSS. This is a key area of intervention, as it is the initial exposure to open source practices. While long tail research software projects do not manage large user communities, they are a valuable stepping stone to understanding what is required to develop and maintain scientific OSS projects. Early career researchers are looking to build OSS skills to advance their scientific careers. Long tail research software projects are training grounds for future OSS leaders, who move on to contribute to larger projects or launch their own initiatives.

While research-driven OSS community members are trained as researchers, leading or contributing to a successful OSS project forces them to "wear multiple hats" , including: project manager, community manager, technical writer, marketer, entrepreneur, fundraiser. Because of the structure of academic institutions, this work is not generally rewarded in faculty or trained in students. Without acknowledgement of these roles and training for them, people working on OSS in research must work with their existing skills and resources, often leading to burnout. Survey respondents and interviewees understood the importance of these skills, but found prioritizing training or mentoring impossible.

Approach: To address this skills gap, develop new leaders, and support long tail research OSS - we recommend a cohort-based program to develop leaders and build communities of practice at research institutions. Such a program will be:

- Focused on research-driven OSS projects

- Centered on leadership development in OSS and practicing those skills in long tail and other real world research OSS contexts, including:
    - Documentation
    - Code review
    - Community engagement
    - Writing for code reuse and reproducibility
    - Project management
    - Developing funding
    - Demonstrating the value of research OSS contributions
        * Designed to grow regional communities of practice
        * Targeted to both new or actively contributing to an existing OSS project

Interviewees in academic roles clearly articulated that they want to build local communities of practice in OSS for research, but lack the resources to do it. The Carpentries, ROpenSci, and Mozilla's Open Leadership Training (OLT) were cited by interviewees as organizations and programs that have made a positive impact in early career scientists' OSS skills. Building a program of cohort-based learning is a tested approach that adds structure to the OSS mentoring process and builds communities of practice. Upskilling early career scientists, and supporting time-strapped mentors, will build stable communities of practice in supporting long tail research projects and beyond. A program that combined the Openscapes and Carpentries' focus on scientists with OLT's cohort engagement would be impactful in this space.

Outcomes: While upskilling early career scientists was identified by researchers as a critical gap, meeting this need alone will not address larger, systemic issues (for example, incentives in academia). Training new contributors may initially add to the burden on already overworked maintainers. By supporting mentors, and coupling this work with investment in project development through an incubator program and sustainability skill building, we hope to push the entire OSS for research ecosystem to a more sustainable place.

This program will provide a leadership pipeline to improve the capacity of long tail projects and provide a place for mentorship of potential leaders for larger projects. The goals for impact include:

- Enhancement of research reproducibility, code reusability, documentation, and overall software quality in smaller OSS for research projects
- Development of OSS leadership capacity in smaller projects

- Support the growth of a skilled force of early career data savvy scientists
- Reduced burnout and enhanced mentorship capacity for local OSS leaders

4.0.3 Advocacy skill development program

Rationale: As many projects in research-driven depend on contributors who are OSS are affiliated with institutions (academic, industry, government) it is critical to develop advocacy skills in the research-driven OSS community. Each type of institution has a unique environment. Academic environments are unique, and pushing for change and resources inside the academy is a skill of its own. With academic advocacy skills, research-driven OSS leaders can advocate for changes at organizations and institutions, a prerequisite to institutional and government agency support for OSS over the next 20 years.

Approach: We recommend advocacy skill development programs targeted at OSS project community members at all levels. To provide focused support to improve coordinated advocacy for OSS, such a program will include:

- Focused, independent or conference-affiliated 1-3 day work sprints
- Or a cohort-based learning program to allow for more support on advocacy and implementation.
- Development of resources based on the experiences of existing OSS projects working to improve alignment between institutions and the OSS research communities, including:
    - Where to look for institutional resources to support OSS
    - Policy change at academic, industry, and government institutions
    - Demonstrating the value of OSS to institutions that engage in research and data science

Outcomes: As our interviewees stated, leadership in OSS is a unique skill, but one that can be learned. Similarly, building advocacy skills to push for more support for research OSS at academic institutions is a skill that can be learned. Measures of impact of this program include:

- Evidence of greater institutional support for research driven OSS through:
    - OSS-friendly university policies
    - Permanent and long-term OSS-focused positions at institutions
    - Increased institutionally-backed funding in the OSS ecosystem
- Documentation that makes the case for training in OSS skills to institutions, such as:

    - Data and case studies of the value of OSS skills for student job placement
    - Data on the cost savings for institutions of OSS use in research
    - Data on the scientific value of OSS in research

- Freely available reports and resources to support academic advocacy, including:

    - Research OSS Champion Stories: Documentation of case studies of institutions that are actively funding and supporting OSS for research.
    - OSS for Research Champions Tool-Kit: Research to document what universities need to support OSS from an administrative standpoint, including answers to the following questions:
        * What base of faculty, staff, student support is necessary to demonstrate a campus need?
        * What mechanisms work to move money from a university to an OSS project housed at a software foundation? Or incorporated as a company?

Acknowledgments

This work was supported by the Gordon and Betty Moore Foundation. Special thanks to Dr. Gina Helfrich who provided feedback on the initial survey and supported its dissemination to the NumFOCUS community. The authors thank all the people who completed the survey, participated in interviews, commented on drafts of the report, and otherwise supported this work.

References

"About | Software Sustainability Institute." n.d. Accessed March 25, 2019. https://www.software.ac.uk/about. Benthall, Sebastian P. 2019.
"URSSI: Software Incubator Workshop: A Synthesis." URSSI Blog. February 25, 2019. http://urssi.us/blog/2019/02/25/software-incubator-workshop-a-synthesis/.

Eghbal, Nadia. 2016. "Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure." https://www.fordfoundation.org/media/2976/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure.pdf.

Katz, Daniel S. 2016. "Defining Software Sustainability." Daniel S. Katz's Blog. September 13, 2016. https://danielskatzblog.wordpress.com/2016/09/13/defining-software-sustainability/.

Lago, Patricia, Sedef Akinli Koçak, Ivica Crnkovic, and Birgit Penzenstadler. 2015. "Framing Sustainability as a Property of Software Quality." Communications of the ACM 58 (10): 70–78.

Leeming, Jack. 2016. "How Is the Rise of Data-Intensive Research Changing What It Means to Be a Scientist? : Naturejobs Blog." September 9, 2016. http://blogs.nature.com/naturejobs/2016/09/30/how-is-the-rise-of-data-intensive-research-changing-what-it-means-to-be-a-scientist/.

Mozilla. 2018. "Open Source Archetypes: A Framework for Purposeful Open Source." https://blog.mozilla.org/wp-content/uploads/2018/05/MZOTS_OS_Archetypes_report_ext_scr.pdf.

Oberhaus, Daniel. 2019. "The Complicated Economy of Open Source Software." Motherboard. VICE. February 14, 2019. https://motherboard.vice.com/en_us/article/43zak3/the-internet-was-built-on-the-free-labor-of-open-source-developers-is-that-sustainable.

Venters, C., C. Jay, L. Lau, and M. K. Griffiths. 2014. "Software Sustainability: The Modern Tower of Babel." For Sustainable .... http://eprints.hud.ac.uk/22201/1/paper2.pdf.