An Initial, High-Level Analysis of the 2023 NSF POSE Awards and Recommendations for Enhancing the Solicitation.

Christopher Baker, Associate Director, Open@RIT
Stephen Jacobs, Director, Open@RIT
Kripa Kundaliya, Researcher, Open@RIT
Michael Nolan, Open@RIT

Introduction:

Open@RIT, RIT's Open Programs Office, has analyzed the Phase I and Phase II abstracts from the 2023 awards from the NSF Pathways to Open Source Ecosystems (POSE) as a service to our campus and the greater Open Work community at large. It is our hope that this will help potential applicants for the coming year's solicitation have a better understanding of the program and its alignment with their own efforts overall. We also provide some recommendations to the program itself for increasing clarity of the solicitation.

The following breakdown has three general points of discussion:

1. The research domains in which the funded project exists, such as physical sciences, computing, etc.
2. The specific Open Source Product at the heart of the proposed ecosystem.
3. The types of work/outcomes the funds will be used to achieve.

Limitations of the analysis:

The NSF provides only proposal abstracts of awarded projects for public consumption. These provide fairly limited information for analysis. Abstracts don't follow a specific format, so their structure and the level of detail they provide vary widely.
The terminology used from one abstract to another also varies. For example, the word "infrastructure" could refer to human, physical, or digital infrastructure, and how the term is being used in a given abstract may not be clear. Another example is "platform," which could reference both a software-development platform or a community-building

platform. Both uses of the term could appear either as an existing Open Source Product (OSP) or as a desired outcome of an Open Source Ecosystem (OSE) which NSF funding would be used to accomplish. Other terms follow a similar pattern throughout the abstracts.

Throughout this document, we will include excerpts from some of the proposal abstracts that illustrate points in our analysis. This shouldn't be interpreted as a ranking or weighting of one proposal over the others, but rather the authors finding particular ones that best illustrated a specific piece of the overall analysis. We recommend viewing the full abstracts excerpted here for greater clarity of the points discussed.

Despite the high level of this analysis and the limitations described above, the authors believe this overview should provide some utility to the community of potential future applicants to the program.

The Funded POSE Proposals

This analysis looks at the 19 Phase I and 19 Phase II POSE awardee abstracts found in the NSF awards database made in 2023. Several other awards for conferences or training were also funded under the umbrella of the program, but these outliers are not included in this analysis.

- The NSF limited Phase I proposals to a total budget of $300,000 for durations of up to one year. Awardees were not obligated to submit Phase II proposals in the future.
- The NSF capped Phase II proposals at a total budget of $1,500,000 for durations of up to two years. A successful Phase I proposal was not required for submitting a Phase II proposal.

Out of the 38, only four were from independent, not-for-profit providers of research tools, services, or platforms. Three more were from private corporations doing similar work. The remaining 31 proposals came from academic institutions

Department Statistics for the PI departments from the 31 academic institutions are as follows:

Computer Science	5
Computer Electrical engineering	7
Information technology	2
Human-centered computing	2
Civil, Architectural, and Environmental Engineering	4
Science (Physics, Math, Statistics)	3
Psychology	1
Education and Human Development	2
Biology	4
Agriculture	1

Methodology:
As mentioned above, we analyzed abstracts by Domain, Open Source Product, and OSE Goals. Each of those groups contains subcategories, and most proposals belong to several different, and sometimes overlapping, ones.

Domain refers to the project's functional disciplines — such as science, education, computer science, high-performance computing, etc.

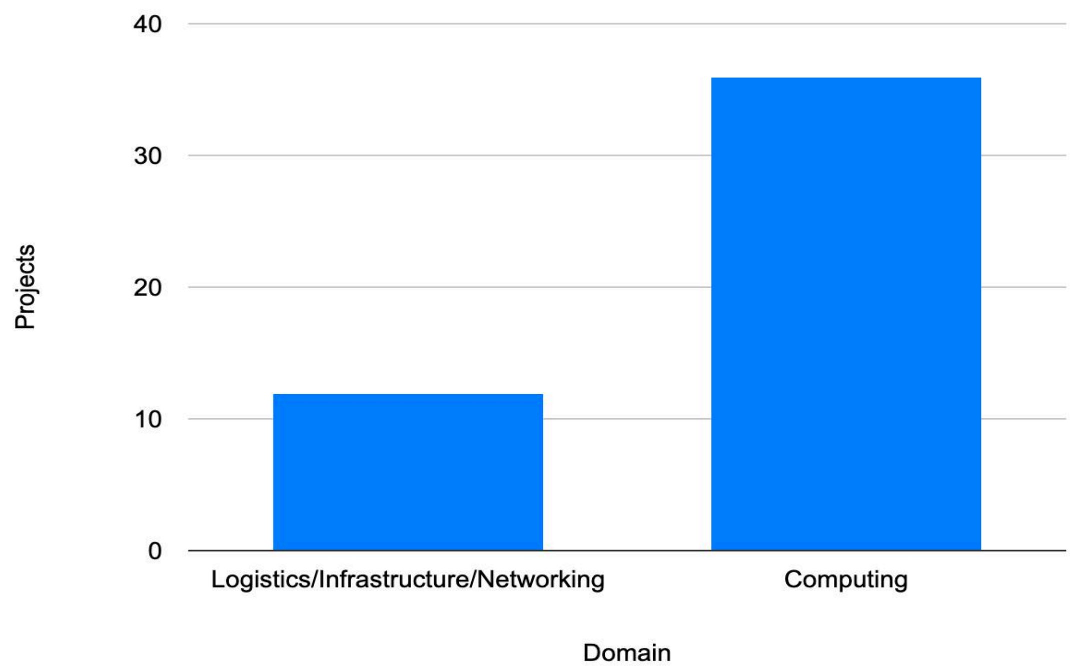
Open Source Product: This is the term the NSF uses in the POSE solicitation to identify the project(s) or research artifact(s) around which the community will be built. These include, but are not limited to, platforms, repositories, software/research tools, and hardware.

OSE Goals: These are based on the desired outcome(s) that funding will be used to achieve. These categories include community plans, ecosystem infrastructure development, governance, community standards, and other types of documentation. Proposals were assigned to as many categories as deemed applicable, and for our analysis, we calculated aggregate percentages based on the total number of projects.

NOTE: For the sake of this survey, percentages will typically represent a category in relation to the whole number of projects (38 projects) **unless expressly stated in the findings.**

We concentrated then on correlations between categories against the whole number of projects, examining what categories were most identified across projects.

For example:
12 of 38 projects are relevant to the **logistics, infrastructure, and networking** domain, and 26 of 38 projects are relevant to the **computing** domain**.**



Domain Analysis
The NSF solicitation supports “Ecosystems” around work in STEM domains. While we see collaboration and sharing, especially data sharing, pursued across a much wider range of domains, the term Open Source (stewarded by the Open Source Initiative) was created for software. By choosing to use that term for the solicitation, rather than a more inclusive one like Open Work, Open Scholarship, or Open Science, the NSF may have inadvertently narrowed the field of proposers. This, and other issues of clarity around the solicitation process, will be addressed in the recommendations section below.

NOTE: when examining the metrics, keep in mind that projects were identified with ALL applicable domains:

- **Computing - 68%**
- **Engineering - 42%**
- **Physical sciences - 39%**
- **HPC - 39%**
- **Education - 32%**

In this analysis, it is possible to calculate correlations between domains. For instance, 43% of all computing-related projects were also relevant to the HPC domain.
Based on our analysis, all projects relate to several domains.

For example:

- **Phase I: Scoping An Open-Source Ecosystem Around Proactive Software Supply Chain Monitoring** - "*…This project tackles the challenge of developing and sustaining a community to provide usable **security**. The project's novelties are in recognizing and building a broader solution that can secure not only cloud systems but emerging applications such as **Artificial Intelligence** and Internet of Things (IoT) as well as mission-critical applications such as the **power grid**."*
    - **Domains**: security, AI/machine learning, and network/infrastructure.

- **POSE: Phase II: An Open Source Ecosystem for Collaborative Rapid Design of Edge AI Hardware Accelerators for Integrated Data Analysis and Discovery** - *"The ecosystem resulting from this project will manage access to a catalog of pre-designed and validated software packages and Intellectual Property hardware blocks, which can be used for both **educational purposes** and to build custom **machine learning computational systems**. The ecosystem will enable application and domain experts from various disciplines and affiliations … to utilize automated design successfully **flows** to create customized machine-learning hardware.*
    - **Domains**: Education, engineering, HPC, AI/machine learning

The education domain was relevant to 32% of projects. However, we identified education resources both as OSPs and as project outcomes. Many projects referenced educational data repositories, Open educational resources (OERs), training events, and software tools for educators and students. However, the project's primary thrust was not to specifically enhance science education as a discipline.

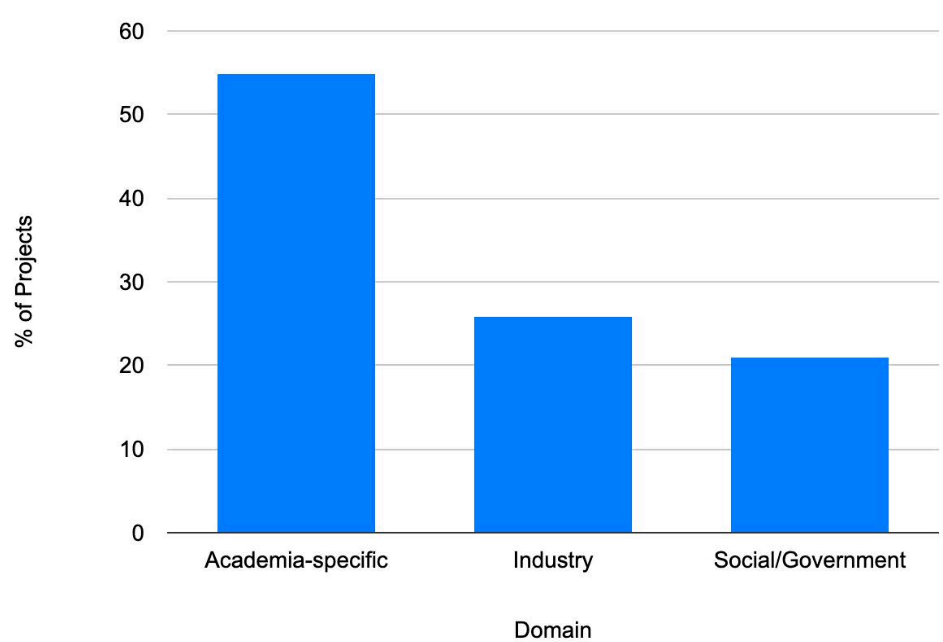
- **POSE: Phase I: An Open-Source Ecosystem for the Creation and Use of Accessible Science, Technology, Engineering, and Mathematics (STEM) Open Education Resources** - *“This project scopes the creation of an Open-Source Ecosystem (OSE) consisting of two existing open-source products: … for publishing OER textbooks that support both instruction and education research in K-12 and higher education.”*

While not citing education as their primary focus, many others indicated education as a secondary or tertiary beneficiary of the work.

- **POSE: Phase II: A Sustainable Open Source Consortium for the Tock Secure Embedded Operating System** - *"…This project lays the foundation to sustain the Tock open source project through documentation, developer tools, security audits, and educational resources as well as establishing stewardship over these.”*

We identified three domains that could be considered their own taxonomy tier: **Academia**, **Industry**, and **Government/Socia**l. We used these to identify the direct impact on these important social sectors that can be considered as domains as well:

- Academia-specific 55%
- Industry 26%
- Social/Government 21%



NOTE: For the sake of this analysis, we have not extrapolated this data to the extent of all possible implications, but this is well worth further examination.

Analysis of Open Source Products (OSPs):
Multiple OSPs were identified in each proposal. Similar to the percentage of proposals relevant to the **computing** domain (68%), 74% of all projects contain a software OSP. We attribute this large number to the same familiarity software developers across all domains have with Open Source code and the practices of Open ecosystems. While software may only be a component rather than the focus of some projects, the volume of software-related proposals is indicative of this prior familiarity with Open Source and its community practices.

In the same vein, 71% of all OSPs were research tool OSPs. While we have not extrapolated further on most categories, we noted that 71% of all research tool OSPs, were, at least in part, software-based OSPs.

For Example:

- **POSE: Phase I: Evolving Exosphere with Community-Driven Software Stewardship** - *"Exosphere is an innovative web-based interface for non-proprietary cloud computing infrastructure. Exosphere empowers researchers to wield advanced* ***cloud-based research tools*** *without needing advanced systems administrator skills. The project's novelties include: providing a* ***user-friendly dashboard*** *to manage cloud computing, networking, and d****ata storage resources****, and providing interactive access to these resources via web browser.*
  …Investigators and community contributors build Exosphere with a fully public development process. They deliver the result as ***free and open-source software****. This approach has made Exosphere the most widely-used interface for Jetstream2, a national-scale research cloud. It has also resulted in advanced features such as push-button elastic virtual clusters and reproducible data science workbenches."*
  - **OSPs**: software tool, repository, data collection, research tool

Outside of Software and research tools, 68% of all proposals listed a repository OSP, and 34% contained hardware OSPs. Both heavily correlated with a software tool, and this is consistent with the current state of hardware development and particularly Open Hardware, which by certification, must run Open Source Software.

NOTE on Education OSPs: as highlighted in domains, 26% of all projects contained education and training OSPs, and those projects were likely to exist within the science education domain.

For Example:

- **POSE: Phase II: Cultivating Modeling Literacy and Practice through a NetLogo Open Source Ecosystem:** *Over the past two decades, the NetLogo team at Northwestern University has developed a large **open-source codebase** consisting of the NetLogo agent-based modeling (ABM) environment and many associated products. NetLogo has become the leading ABM **platform**, with hundreds of thousands of users including both researchers and **K-16 educators**. **Researchers** in the natural sciences, social sciences, and policy have published thousands of **scientific articles** using NetLogo. Educators use NetLogo to **engage students** in learning about complex systems through inquiry-based modeling activities.*
    - **OSPs**: platform, repository, software tool, research tool, OERs
    - **Domains**: science education, physical sciences, computing

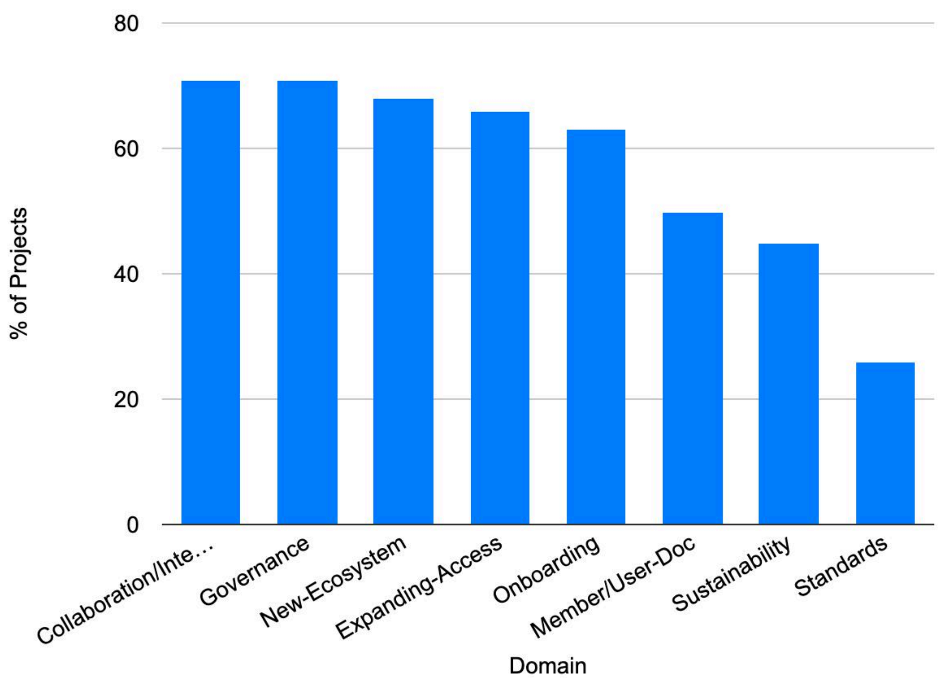
OSE Outcomes

NOTE: a distinction between communities and ecosystems: informal collaborative communities may exist around an OSP but lack sustainability elements. An ecosystem would be the sustainable structure inside which the OSP community can proliferate.

POSE intends to invest in OSE health focusing on sustainable infrastructure for viable communities through roadmaps, policies, governance structure, and documented procedures to ease entry for new contributors. These serve to distribute responsibility across the community, and promote interdisciplinary collaboration.

Open Source communities of practice often evolve informally as a way to pursue common interests. As a result, the use of open-source products can grow beyond the capacity of the original software design and/or the original software team. These cases can place the entire OSE in a sustainability crisis. An overview of a case study on a situation like this can be found in *Conceptual Mismatches,* Chua, M. and Jacobs, S. While no two communities might operate exactly the same, productivity should be considered the primary focus of builders and maintainers. Processes for collaboration become essential to the core templates of these plans and will help inform future Open community plans.

In all, 71% of projects stated collaboration and interdisciplinary work, as well as governance as a core outcome. The other most common outcomes new ecosystem 68%, expanding access 66%, and onboarding 63%, 32% Member/user Documentation (combined), 45% Sustainability, 26% standards



Outcomes by Phases: According to the solicitation language, Phase I focuses on analysis and planning for an OSE, and Phase II proposals concentrate on implementing or expanding the OSE. With this in mind, we have extrapolated the outcomes by phases to identify correlations between various outcomes and phases.

For Example:

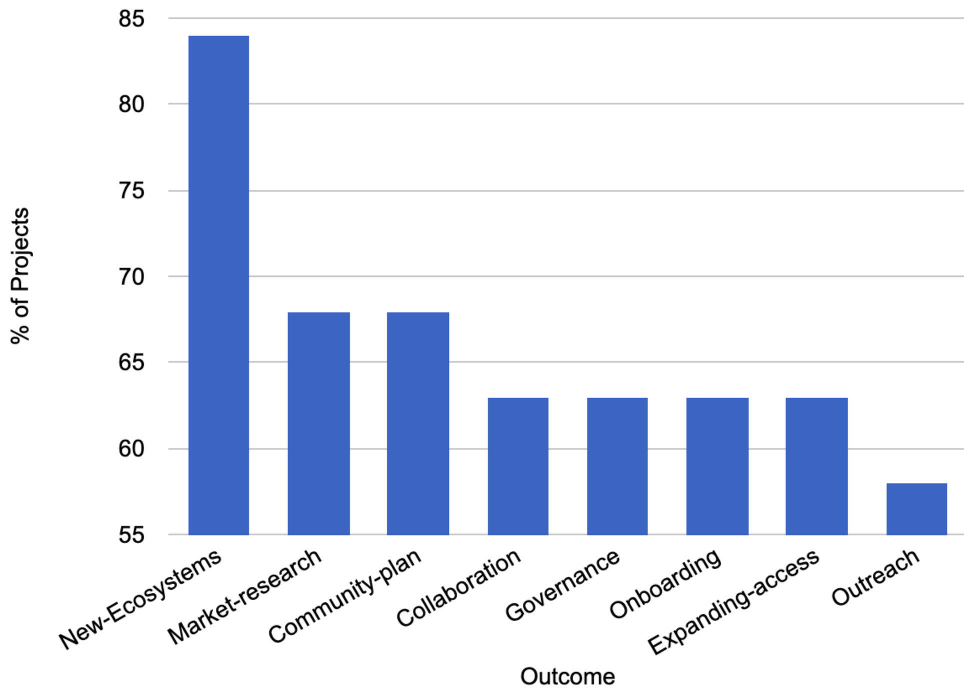
- **POSE: Phase I: Advancement of an open-source hardware and software ecosystem for the Open Source Bionic Leg —** ***Discover*** *the appropriate ecosystem and deployment mechanisms through* ***interaction*** *with* ***existing and future users****. Build and refine the open-source* ***infrastructure and governance tools*** *through discussions with other open-source ecosystems, especially organizations that manage open-source hardware… disseminate information about the Open-Source Leg project and its use through conferences and web-based media.*

    - **Outcomes**: community plan, market research/evaluation, governance, interdisciplinary collaboration, new ecosystem, new innovation, conference workshop,

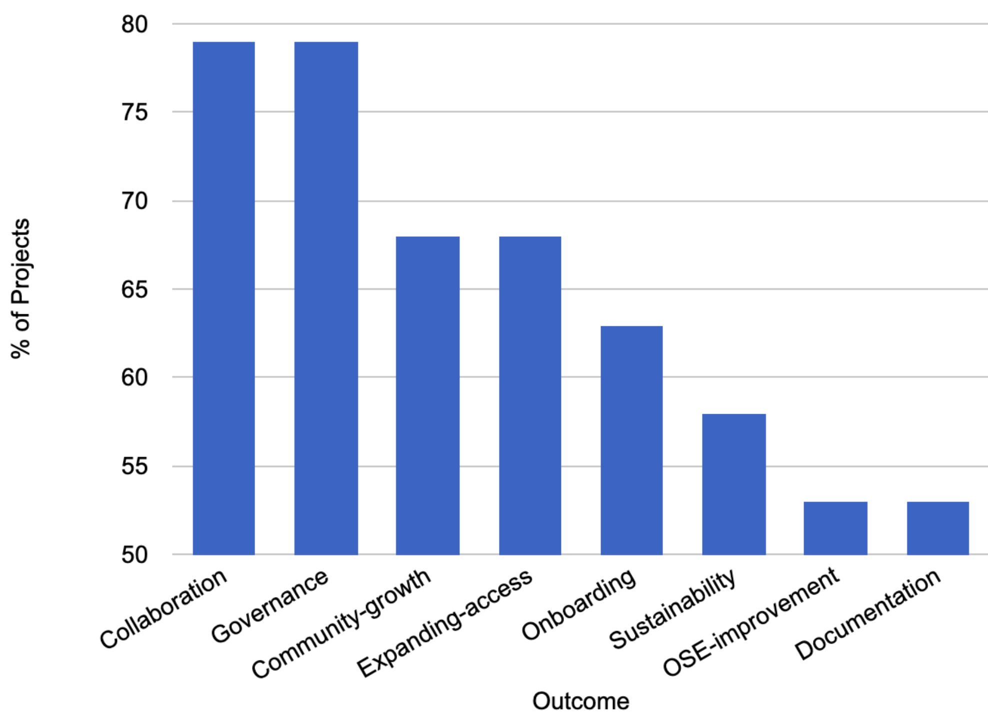
- **Phase II: Growing GRASS OSE for Worldwide Access to Multidisciplinary Geospatial Analytics —** *The demand for GRASS GIS is increasing and therefore a* ***coordinated effort*** *is needed to* ***grow the number of researchers*** *and other software* ***users*** *at* ***universities****,* ***government*** *agencies, and* ***businesses*** *who can* ***help maintain****, and* ***improve access*** *to, GRASS GIS.* ***Expanding the number and diversity of contributors*** *from academia and research, non-profits, and industry will* ***facilitate development of new geoprocessing*** *engine features and ensure long-term maintenance of contributed research models.*

    - **Outcomes**: community growth, new innovation, expanding access, interdisciplinary collaboration, security, outreach, onboarding, sustainability

Following the explanation above, the most relevant outcomes to ***Phase I*** proposals were **new ecosystems** 84%, **market research/evaluation** 68% and **Community plan** 68%, collaboration, governance, onboarding, expanding access at 63% and outreach at 58%.



The outcomes most relevant to ***Phase II*** were **collaboration** and **governance** at 79% **community growth** and **expanding access** at 68%, onboarding 63% sustainability 58%**,** existing OSE improvement and documentation (both user and member) at 53%.



Outcomes with the greatest difference between phases were: market research 68/26, community plan 68/26, improving existing 10/53, community growth 32/68, CI 0/32, new ecosystem 84/53, user documentation 26/53, sustainability 32/58, outreach 58/37.

General Conclusions

Successful POSE awards landed primarily either in projects that originated in computing disciplines or relied heavily on software to do their research. Many of them had education/educators as a primary domain or “customer” and most of them pointed to educational use as an additional area that would benefit from their work.

The majority of the efforts had an unstructured user community already extant, with a number of proposals looking to merge two or three existing OSP and their OSEs together.

Recommendations:

1. **Change the name:** If the goal of the POSE program is to attract projects beyond computational domains and software tools, a name change might be in order. The term Open Source is a term that specifically addresses licensing of software,. Over the last decade or so, Open Source has been inappropriately applied to a wide variety of products being distributed with flexible licensing or released into the public domain, i.e. without patent or copyright. Creative Commons, the entity that has created variations on traditional copyright to allow for more flexible use of IP, does not recommend their own licenses for software. Instead, they point software creators to Open Source licenses. Entities like the National Academies of Science Education and Medicine, and HELIOS Open, use the terms Open Scholarship and Open Science. Should the NSF elect to move to one of these, it would allow the acronym to stay the same but could communicate a broader interest in more diverse proposals.
2. **Clarify the solicitation:** In Open@RIT’s work with RIT faculty and the broader network of peer organizations and institutions involved in the Open Work space, we’ve seen those interested in **POSE** confused on whether efforts beyond pure computation and hard data collection could be included. Often, there were also questions on whether a percentage of the funds could be used for technology or software development that would broaden the recruitment and /or access to a broader-based ecosystem.
   For example, would improving OSP technology for contribution pipelines and systems be acceptable as part of the effort funded by POSE? Or, is developing

different permissions structures and filtering into existing OSPs within POSE's scope an appropriate use of funds?
These kinds of efforts can broaden participation or cleanly define how different members of the OSE can contribute and engage, so it would benefit the clarity of the solicitation to directly address these types of activities as within or outside of the scope of acceptable use of POSE funds

3. **Provide some examples of accepted proposal narratives:** While there are many clear reasons why full proposals should not be posted for public view, providing access to the purely narrative sections of a few successful proposals would provide some guidance to those considering submission. The National Endowment for the Humanities and the National Endowment for the Arts do this on their web pages. As stated in the limitations section, this analysis is based on the limited information available in the abstracts published by NSF POSE. Examining the extended narrative portion of the proposals would allow for more comprehensive and beneficial support for proposers and findings of future analysis efforts such as this one.
4. **Provide Webinars and Trainings on Open Source Sustainability Models:** In the informal interactions with potential applicants described above, Open@RIT was asked about a requirement for awardees to "become a business." Though Open Science and Open Source Software are not identical per se, there are decades of Open Source sustainability models beyond pure commercialization that can be applicable, either directly or with adaptation, to Open Science/Scholarship.
The NSF is providing post-award funding to support I-Corps training for Phase II awardees who want it, but that training tends to focus on tech transfer business creation processes. It would be fairly easy to assemble a group of experts in the field to develop or point to existing materials around a variety of Open Source Software sustainability models.
5. **What happens after implementation?** While the program supports the creation of OSE plans and governance (Phase I) and the initial implementation of those plans (Phase II) there should be mechanisms in place to support these initial efforts beyond "turning on the engine." It takes time for these ecosystems to demonstrate the return on the investment and truly determine if successful OSEs grow the impact and translation of their OSPs.
As NSF is just getting started supporting OSEs, there would be tremendous benefit to supporting and analyzing outcomes over several years beyond initial implementation. By doing so, NSF would identify the qualities and methods of projects that achieved their goals. Such data could lead to a collection of best practices and guidelines for Open Science and Open Scholarship.