

The Open Source SCE

Caroline Phares, Domino Data Lab, San Francisco, USA

ABSTRACT

With the prevalence of collaborative efforts and use of open source on the rise, what new opportunities does this open? For many years, organizations have tried to solve the same problem: implementing the ideal SCE. A few years ago, the PHUSE SCE white paper standardized what the ideal SCE should look like. Today, inspired by that white paper, and following the tide of the open source movement, the Open-SCE community has been launched. This presentation will describe the vision of Open-SCE, review the activity to date, describe the current contributions, and extend an invitation to come along the journey.

INTRODUCTION

Hundreds of thousands of years ago, over the course of human history, homo sapiens coexisted with at least 4 other human species. Today, homo sapiens is the only one that has prevailed and is still in existence. Evolutionary anthropologists have concluded that the reason for this has to do with what is referred to as “cooperative communication”, homo sapiens’ unique ability to work together with others toward accomplishing a common goal. “[Cooperative communication] allows us to plug our minds into the minds of others and inherit the knowledge of generations. It is the foundation for all forms of culture and learning, including sophisticated language.”¹

“Cooperative communication” is not an everyday term. Instead words like “collaboration” or “cooperation” are used to describe something similar. These words are often used interchangeably, or sometimes each may have a slightly different meaning from the other. For the sake of clarity, the term “collaboration” will be used to describe the action of working jointly with others to create a product.

Open source is a collaboration where the product, in this case software, is made available for anyone to use, build on, and/or, within reasonable parameters, contribute to. Although open source is not a new concept, over the course of the past several years, there has been increasing interest across pharmaceutical companies, CROs, software vendors, and even freelance individuals on developing, contributing to, and using open source tools to aid in the statistical analysis and reporting of clinical studies. This paper will describe a vision and journey to expand the open source concept to a set of useful building blocks for an SCE solution.

USING OPEN SOURCE TO DELIVER CLINICAL STUDY ANALYSES

DEFINING OPEN SOURCE

Open source is not a new concept. It has become increasingly popular over the past 30 years. In 1998, the term “open source” was coined, and the Open Source Initiative (OSI) organization established a set of defining criteria². In order for software to be considered open source, according to the OSI, the following conditions must apply:

1. The distribution terms must not restrict the software from being used as a component of proprietary software.
2. Along with the compiled code, the source code must also be available.
3. Changing or building upon the code and distributing the modified or extended software must be permitted.
4. Distribution of the modified or extended software may require it to have a different name or version number.
5. There can be no discrimination against any group or person in the distribution of the software.
6. The software can be used for any purpose, without discrimination.
7. Rights to the software extend to the redistribution of the software.
8. Rights to the software are independent of any other software.
9. No restrictions must be imposed on any other software with which it is distributed.
10. Provisioning of the software should be technology-agnostic.

The OSI has a licensing program, where entities can apply to have their software officially labeled as “open source” after a thorough review of the license terms of the software is conducted by the OSI board. It is worth noting that not all open source projects go through this official process. Today, there are open source licenses with similar criteria that have been evaluated through OSI and can be applied to open source projects. Popular licenses include GNU

General Public License version 3 (GPLv3), Apache License 2.0, Berkeley Software Distribution (BSD), and MIT License.

THE BENEFITS OF OPEN SOURCE

The obvious benefits of open source are (1) that it is free to access and use, and (2) it provides freedom to the programming community to modify and extend the code to accelerate progress and meet specific needs. Since purpose is not restricted, developers can fork projects and build upon what has already been done in a completely new direction. In this sense, open source is a readily-available accelerator. In fact, even proprietary software vendors have embraced open source as a business accelerator and centered their businesses around proprietary enhancements to open source software.

Anecdotally, open source progresses and evolves much more quickly than proprietary software. This may be due to the pool of motivated developers who dedicate their personal time to maintaining and enhancing the software. The motivations of these individuals indicate that they are talented and passionate about improving and applying their skills through useful open source contributions in particular. A recent study showed that developers are motivated to contribute to open source mostly by fun, altruism, reputation, and kinship, and that learning and career development are far greater motivators than pay for these individuals³. Having talented and intrinsically motivated individuals working on an open source product positively impacts the quality and speed at which it is delivered.

In software engineering, the intersection of open source and open standards has propelled innovation and progress, and has indeed reduced the time and cost of developing software. HTTP, XML, and JSON are just three of many examples. The hope is that this will translate directly to statistical programming for clinical study submissions. It is expected that readily available, free, open source tools and packages will reduce the effort, time, and cost of creating statistical analysis deliverables for clinical study submissions. The success of projects like OpenCDISC and Admiral will be strong indicators.

CHALLENGES TO APPLYING OPEN SOURCE TO REGULATED STUDIES

Embracing open source in statistical programming has not been without its challenges. This particular movement has been active for many years. However, it was only within the past several months that the first real regulatory submission was achieved where the statistical analysis deliverables were produced using R. Some of the challenges needing to be overcome to reach this success continue to be debated.

One of the biggest challenges to embracing open source, specifically R, for regulatory submissions, is a concern regarding regulatory compliance, particularly validation. Advocates for R and open source established the R Validation Hub (<https://www.pharmar.org/>), a collaborative effort focussed specifically on addressing this concern. Additionally, many companies have developed validation frameworks for R packages. The concern around validation has largely been addressed, however a remaining challenge to validating open source is that it progresses quickly. A validation effort is required for each package and version that is intended to be used in producing submission deliverables. Validation does require time and effort, and therefore, companies do incur a cost to use open source.

Another challenge to fully reaping the benefits of open source is that, unless someone is contributing as a free agent, sponsor companies must agree to the release of intellectual property (IP) created by their employees. Even if someone wants to contribute, unless their contribution is allowed or approved by the company they work for, they could be putting themselves at legal risk. Sponsor companies must support their employees in contributing to open source, and the process of obtaining approval must not be difficult.

Even if sponsor companies are supportive, the remaining challenge to consider as well is motivating talented programmers to contribute. How can open source project owners and champions attract the most talented programmers to contribute to their projects? A recent study found that the most highly skilled programmers have the greatest motivation to contribute when the repository owners (1) show no commercial interest, (2) are generous in accepting external contributions, and (3) are quick to respond to merge requests and provide feedback⁴.

AN OPEN SOURCE SCE

Statistical Computing Environment (SCE) is a term that has been commonly and frequently used in the pharmaceutical industry to describe an analytics platform supporting the analysis and reporting of clinical studies. In 2019, a collaboration among several prominent pharmaceutical companies was launched with the purpose of recommending a standard set of requirements for a modern SCE. A resulting "SCE White Paper" was published in 2021. According to the white paper: "there are many off the shelf solutions available to support multiple aspects of an analysis and reporting platform, we are yet to see one that comprehensively meets the full set of Pharma end-user

requirements and future-ready ambitions, leading to many large organizations developing their own independent customized analytics solution.”⁵

Over the course of the last 3-4 years many pharmaceutical companies have invested in technology modernization programs that include implementation of a new SCE. The challenge still remains in meeting the entire set of recommended standard requirements. Some companies have chosen to build rather than buy in order to meet as many of the requirements as possible. Some companies have compromised on future-readiness in order to implement a more familiar user experience. Some companies have chosen the flexibility of modern data science platforms to meet future-ready needs, while building components to facilitate the more familiar ways of working. The idea of an open source solution to the SCE problem was born out of the latter.

THE OPEN-SCE

BIRTH OF A VISION: THE AUTHOR’S JOURNEY

I am currently employed by Domino Data Lab. I am also a former customer, having been involved in the selection and implementation of the Domino platform at GSK as part of the SPACE program. I co-authored and presented a paper on the experience at the 2022 PHUSE US Connect conference in Atlanta. This is where my journey begins.

Domino is an analytics and data science platform that fundamentally meets an important majority of the recommended standard SCE requirements. Domino is also API enabled which makes it a good starting point and accelerator for developing immediately usable components of an optimal SCE solution. The companies who have chosen this path to implementing a modern SCE have been eager to work together to drive Domino’s product direction as well as share solutions. Not only have they been eager to work together, but consistent with the collaborators on the PHUSE 2021 SCE White Paper, there are significantly greater similarities in their SCE needs than there are differences.

Meeting the needs of the industry in a timely manner was and is my primary goal. Open source by nature can progress and be enhanced more quickly than copyright licensed software. In my experience, I witnessed skilled customers of Domino quickly develop UI and CLI tools to enhance their SCE experience. The idea occurred to me that perhaps the needs of Domino’s SCE customers could be met more quickly through the development of open source components of a Domino-based SCE. Perhaps this could even be applied more broadly.

If industry experts can collaborate on developing a standard set of requirements for a modern SCE, can a similar concept apply to developing software to meet those requirements? Can a community of knowledgeable, skilled, and creative professionals be inspired to contribute their efforts to innovating and developing fit-for-purpose SCE components? Would companies be willing to release IP developed by their employees for such an initiative? These were the questions that led me to form the Open-SCE community.

GETTING OFF THE GROUND

The idea for an open source community may begin with one person, however it takes a village to make it real. For the Open-SCE, Domino executive support and sponsorship provided the necessary foundation. Domino sponsored the setup of a GitHub organization and domain name. In addition to sponsoring the framework, Domino executives developed an approval process for releasing IP to the Open-SCE community. Without Domino’s executive support, the Open-SCE would likely have not progressed beyond just an idea. In addition, a community of interested volunteers was crucial. Initially the community began with a handful of Domino employees who began by establishing the foundation with the implementation of the Open-SCE GitHub organization and web site. The Open-SCE community was officially announced with a call for volunteers in March 2023 in Orlando at a Domino-sponsored breakfast event.

The first to take action were James McDermott and Guy Garrett from Achieve Intelligence, who volunteered their time in helping to develop the charter and governance framework for the community. Later, Achieve Intelligence also made the first open source contribution to the organization: the Study Migration Tool project. Study Migration Tool is a Plotly-Dash Python application initially developed by Achieve Intelligence that generates visualizations of study milestone activity for analysis and migration scheduling.

Domino Data Lab followed by contributing two projects to the organization: Domino Audit Solution and Code Translation Toolset. Domino Audit Solution provides three types of audit trail reports for activities on the Domino platform. Each of the audit trail reports can be accessed in three different ways: 1) A CLI, which is designed to run either inside a Domino workspace or externally, 2) an API endpoint, authorized by a valid Domino API key, and 3) a

web application, deployed as a standalone service. Code Translation Toolset includes a number of assets for automatic code translation between R, Python, and SAS using generative AI.

KSM Technology Partners contributed a fourth project: Domino Fluent CLI (Command Line Interface). Fluent CLI is an enhanced CLI for interacting with the Domino Platform. It provides the ability to output results as plain text, JSON, or XML; it is a zero dependency native executable; it provides detailed context sensitive help for each command; and it allows users to easily script and chain commands together in their favorite shell (Bash, PowerShell, etc).

THE COMMUNITY

The Open-SCE community aims to foster collaboration, innovation, and the development of high-quality open-source projects to deliver a collection of tools and applications which enable more efficient workflows for statistical programming organizations to deliver quality statistical analyses for regulatory submissions, interfacing with the Domino Data Lab platform in addition to other technologies. By promoting transparency, inclusivity, and shared ownership, we strive to create a vibrant and welcoming environment for individuals passionate about open source software and driving positive change through data science and statistical programming in pharmaceutical research and development. The Open-SCE community is centered upon the principles of openness, collaboration, and transparency.

The organizational structure of the community contains a board of trustees, a content management team, and a framework management team. Decisions regarding community-wide matters, such as project direction, major changes, or the adoption of new initiatives, should be made collaboratively through open discussions and consensus building. In matters of conflict, the board of trustees will vote on resolutions. The content management team will be responsible for managing and approving the quality of new content and standards applied to contributions to the community. Updates to content will be managed on a project basis by project leadership which will be subject to the standards as defined by the content management team. The framework management team is responsible for ownership and management of the Open-SCE GitHub organization. This team will approve new projects and create new repositories as required. All of the open source projects under the Open-SCE organization are required to be licensed under GPLv3. The rationale for this is that the GPLv3 license makes all of the projects free for use but also mandates that any enhancements made to the code are contributed back to the community.

Decisions regarding community-wide matters, such as project direction, major changes, or the adoption of new initiatives, should be made collaboratively through open discussions and consensus building. All members shall have an equal opportunity to voice their opinions and contribute to the decision-making process. In the event of conflicts or disagreements within the community, members are encouraged to engage in constructive dialogue to find resolutions. Mediation, facilitated by team leaders or the board of trustees, may be utilized when necessary. The goal is to promote understanding, find common ground, and preserve the community's well-being. Projects initiated within the community shall be owned collectively, with no individual having exclusive control. The community may establish project maintainers or core teams responsible for overseeing specific projects, but these roles should be based on meritocracy and open to rotation. The community shall ensure that no individual or group unduly influences project direction. All community members are expected to adhere to the community's code of conduct, which outlines acceptable behavior, inclusivity, and respect for others. Harassment, discrimination, and offensive conduct will not be tolerated.

Contributions from all members are encouraged, regardless of their level of expertise. Contributions can include code, documentation, bug reports, feature requests, design suggestions, translations, and more. A well-defined contribution process shall be established by the framework management team, including guidelines for issue tracking, code review, and project submission. Contributions shall be recognized and celebrated within the community. Acknowledgement can take various forms, such as credits in documentation, listing contributors in project repositories, or highlighting significant contributions in release notes. Recognition motivates and encourages community members to continue their valuable contributions.

CONCLUSION

As human history has shown, collaboration is integral to progress and success. The connection between collaboration and progress is aptly captured in the often quoted African proverb. "If you want to go fast, go alone. If you want to go far, go together." Human experience in scientific and technological progress appears to follow the same pattern, and the open source movement has embodied the sentiment. With the 2021 SCE White Paper, the statistical programming community reached a junction between SCE and open source. The Open-SCE community aims to follow through and continue on the path toward the ideal, modern SCE, collaboratively through open source. It has started small, but the vision is large. In order to reach that vision, help is needed. The conclusion of this paper is a call to action for individuals and organizations interested in contributing knowledge, skills, and IP to the

Open-SCE. Anyone interested in getting involved with the Open-SCE community should start by visiting the community website: <http://open-sce.org>.

REFERENCES

- ¹"Survival of the Friendliest", By Brian Hare and Vanessa Woods, Scientific American 323, 2, 58-63 (August 2020)
<https://www.scientificamerican.com/article/humans-evolved-to-be-friendly/>
- ²"The Open Source Definition (Annotated)", Open Source Initiative, created 2006, modified 2023,
<https://opensource.org/definition-annotated/>
- ³"The Shifting Sands of Motivation: Revisiting What Drives Contributors in Open Source", By Marco Gerosa, Igor Wiese, Bianca Trinkenreich, Georg Link, Gregorio Robles, Christoph Treude, Igor Steinmacher, Anita Sarma, 43rd ACM/IEEE International Conference on Software Engineering, Madrid, Spain. June 2021,
<https://arxiv.org/abs/2101.10291>
- ⁴"What makes the right OSS contributor tick? Treatments to motivate high-skilled developers", By Inna Smirnova, Markus Reitzig, Oliver Alexy, Research Policy, Volume 51, Issue 1, 2022,
<https://doi.org/10.1016/j.respol.2021.104368>
- ⁵"SCE White Paper", By Mark Bynens (JnJ), Sheetal Patel (GSK), Olivier Leconte (JnJ), Delyth Jones (GSK), Sam Warden (GSK), Sascha Ahrweiler (Bayer), Oliver Richter (Boehringer Ingelheim), Jorine Putter (GSK), Joseph Rowley (Novartis), MarieClaude Laramée (Novartis), Des Burke (GSK), Holger Dach (Bayer), Mario Lozina (Boehringer Ingelheim), Jon-Paul Mewes (Roche), Paul Fioole (JnJ), Eunice Ndungu (Merck), Mary Kuklinski (BMS), Timothy Kelly (BMS), Timothy Stuart Pearce (Pfizer) and Gary Chen (Pfizer). 2021, PHUSE.
https://phuse.s3.eu-central-1.amazonaws.com/Events/2021/US+Connect+2021/SCE_White_Paper_Final_15Dec2021.pdf.
- ⁶ <https://open-sce.github.io/>

ACKNOWLEDGMENTS

KSM Technology Partners
Achieve Intelligence
Domino Data Lab

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Caroline Phares
Domino Data Lab
Email: cmphares@gmail.com
Web: <https://www.linkedin.com/in/caroline-m-phares/>

Brand and product names are trademarks of their respective companies.