Understanding the Sustainability Challenges for Building Open-Source Scientific Software

Shurui Zhou
Assistant Professor
@shuishuiblue
Sustainability Challenges in Open-Source

• Sustaining the project
• Sustaining the community
Open-Source Scientific Software

Interdisciplinary collaboration

https://github.com/scijava/
Takeaway: Sustainability problem can get worse

- Two-fronted risk
- Requires both domain-specific knowledge & SE knowledge
Interdisciplinary Collaboration when building AI-based Software

- Different experts tend to focus on different stages in the machine learning lifecycle

Interdisciplinary Collaboration when building AI-based Software

MLOps
- Model Requirements
- Data Collection
- Data Cleaning
- Data Labeling
- Feature Engineer
- Model Training
- Model Evaluation
- Model Deployment
- Model Monitoring

User Interface
- Payment
- User Account
- Results & Editor
- Audio Upload
- Speech Recognition
- Monitoring
- ML Pipeline
- Database, Hadoop, Kafka
Our Focus: Sustainability Challenges when Building Scientific Software in Open-Source

• Majority of development work is done by scientists
• Professional SDE may be employed to create and maintain the software

Q1

What are the major obstacles when an interdisciplinary team builds and maintains a scientific OSS?
What are the main factors for sustaining the scientific OSS community?
A Case Study on a scientific software in Physics domain

Q1: Science-related challenges in OSS context
Q2: OSS-related challenges in scientific comm.
A brief intro of Moonpie

- A software ecosystem
- 1 core package & 50 interoperable packages
- Core package has 1.6K forks
- Over 10 years
- The project has been cited over 5.5K times
- 41 core contributors
- > 400 contributors in total
A Case Study on a scientific software in Physics domain

Q1: Science-related challenges in OSS context

Q2: OSS-related challenges in scientific comm.

Mining Repositories

Interview

Survey
Q1: Science-related challenges in OSS context

• Focusing on 41 core contributors

• Understanding the type of contributions

  • The code file can be divided into 2 categories:

    • Infrastructure

    • Domain-specific
Q1: Science-related challenges in OSS context

number of merged commits

Infrastructure & Management-related

Domain-specific
Q1: Science-related challenges in OSS context

number of merged commits

800 1600 2400 3200 4000

Infrastructure & Management-related

Domain-specific
Conflicts between different mindsets

• Incentives of making contribution
• Prioritizing the tasks

“rigid coding standards... they need to accept the flexible nature of scientific software collaboration”
Conflicts between different mindsets

• Perception of seniority

“senior researchers in the decision-making position sometime ignore certain PRs because they do not see the value of the research.”
A Case Study on a scientific software in Physics domain

Mines Repositories

Q1
Science-related challenges in OSS context

Q2
OSS-related challenges in scientific comm.

Interview
Survey
Q2: OSS-related challenges in scientific community

Survey questions to disengaged contributors:
1. incentive
2. reason of disengagement
3. suggestion of improving sustainability
Incentives

- GSoC
- Invitation
- Learning
- Pay
- Altruism
- Own-use

Reasons for disengagement

- Lacking SE background
- One-time opportunity
- Lacking science background
- Project is stable
- Focus shifted
NOT MUCH
WHAT'S NEXT? We need a different strategy
We need a different strategy

Lowering the entry barrier in science

• Make both SE and science more accessible
  • Documentation for source code and scientific theory
• More guidance on Good First Issues
• Tooling support?
We need a different strategy

Recognizing the participation and contribution

• Citing the tool you are using

• Acknowledging the impact of contribution
  • Quantifying the impact?
  • Identify the usage in a large scale
Acknowledgement

Jiayi Sun
PhD student
U of Toronto

Aarya Patil
PhD student
U of Toronto

Youhai Li
Undergrad
U of Toronto

Jin Guo
Professor
McGill University
Understanding the Sustainability Challenges for Building Open-Source Scientific Software

Recognize the unavoidable tension

Improving accessibility for both code and science

Giving credits to the contribution

Shurui Zhou
@shuishuibleu