# Understanding Practices and Challenges of Developing Sustainable Research Software: A Pilot Interview

Daniela Feitosa*
Christina von Flach*
dfeitosa@ufba.br,flach@ufba.br
Federal University of Bahia, Institute of Computing
Salvador, Bahia, Brazil

Joenio Costa*
joenio@gmail.com
Université Gustave Eiffel, Laboratoire Interdisciplinaire
Sciences Innovations Sociétés (LISIS)
France

## ABSTRACT

Research software is the software developed in universities and research laboratories as part of or to support their research. Research software sustainability, that is, the software's ability to last and continue to be supported over time, fosters long-term reproducibility. In this context, it is essential to understand and describe the current practice in universities and research laboratories concerning the development of sustainable research software. However, such information is seldom available, even for research projects funded by national agencies.

In this paper, we present the results of a pilot interview study conducted with a research group in Applied Physics, whose members historically developed most of the supporting research software. We interviewed the leading investigator to collect information about his knowledge of open science, research software, and software engineering practices, the challenges and supporting factors to develop software, and possible efforts to make it sustainable. The pilot study allowed us to refine the study design to support a more comprehensive study with other groups at the same University.

## KEYWORDS

Research software, Sustainable software, Qualitative studies.

## 1 INTRODUCTION

The increasing use of *Research Software* (RS), that is, software developed in universities and research laboratories as part or to support their research, has instigated the scientific community's concern with its sustainability and ability to support the reproduction of studies by independent researchers.

Software sustainability has to do with the software's ability to last and continue to be supported over time, implying the longevity and maintainability of the software and its communities [5]. Reproducibility is about ensuring that anyone with *access* to the data and software can feasibly reproduce results, both to check them and to build on them [7].

It is essential to understand and describe the current practice in universities, research laboratories and industry concerning the development of RS and evaluate their sustainability for long-term research [1, 5]. However, such information is seldom available, even for research projects funded by national agencies [1].

In this context, the main goal of our research is to undertake a full-scale investigation with research groups to characterize the current practice in the Federal University of Bahia (UFBA) relative to the development and socio-technical sustainability of RS from the perspective of researchers who develop it. We resort to interviews with leaders of research groups. Interviews are often one-on-one verbal conversations, and can also be carried out online. Although time-intensive, in interviews we could ask follow-up questions or clarify what the respondent meant. Related work mostly focus on survey studies [1, 2, 6] with academics and developers.

This paper is organized as follows. Section 2 presents related work. Section 3 details the design of our research. Section 4 presents a pilot study to test the interview guide and refine the research methods before conducting a more comprehensive study. Section 5 discusses the pilot interview, highlights important issues and limitations. Section 6 presents the conclusions.

## 2 RELATED WORK

Carver and colleagues presented findings from a survey of 1,149 researchers, primarily from the United States, about sustainability challenges they face in developing and using RS [1]. Our study is concerned with interviewing leaders of research groups focusing on their knowledge about RS and software engineering practices they use. We plan to build upon the findings of a pilot study to conduct interviews with other research group leaders, followed by a survey with graduate students.

Wiese, Polato, and Pinto investigated the "most pressing problems, challenges, issues, irritations, or other *pain points*" encountered when developing scientific software. They presented a taxonomy of 2,110 problems that are either (1) technical-related, (2) social-related, or (3) scientific-related [6]. In our work, we are concerned with those three dimensions and defined specific questions to address them in the interview instrument.

Jay and colleagues [2] reported a study that includes both RS engineers and domain researchers to understand how scientists publish code. The researchers interviewed domain scientists to identify the barriers to publishing their code. Then they interviewed RS engineers to understand how they would address those barriers. Finally, they synthesized the results from the interviews with RS engineers into a series of survey questions sent to a larger group of domain researchers. In our pilot study, we conducted a semi-structured interview with a domain researcher to understand how his research group develops and publishes code. The pilot study supported the improvement of our research design for future interviews to be performed with other research group leaders.

## 3 GENERAL RESEARCH DESIGN

### 3.1 Context and goal

The lack of sustainability of RS is a crucial issue in academic research, as it affects the reliability and reproducibility of studies.

*The authors contributed equally to this research.

Despite its importance, software sustainability is frequently neglected in research projects, often due to lack of awareness and knowledge among researchers about the best practices for developing sustainable RS.

The general goal of our study is to identify the current practices, challenges and opportunities related to sustainability in RS and promote RS sustainability in academic research groups, supporting evaluations, providing guidelines and recommendations for improving RS sustainability.

For the interview study, the profile for participants is "researcher leading research groups with direct involvement in software development" from the academic community affiliated with UFBA.

## 3.2 Workflow

For each interaction with a research group leader, we defined a following 4-step workflow. In this paper, we address and present the preliminary results of step 2.

(1) Present a talk to introduce the fundamental concepts of Open Science, Research Software, and Sustainability.
(2) Interview the leading investigator to collect information about his knowledge of Open Science, RS and software engineering practices, the challenges and supporting factors to develop RS, and possible efforts to make it sustainable.
(3) Analyze, based on well-defined assessment criteria and guidelines, the sustainability of at least one RS project developed by the group.
(4) Present the evaluation report to the leading investigator and collect his feedback about the assessment of RS and the research method and instruments.

## 3.3 Talk

We prepared the slides for a 40-minute talk to introduce concepts about open science, reproducibility, RS, sustainability, and software development to an audience of researchers and graduate and undergraduate students. The talk contents emphasized the importance of software for modern science and the need to adopt the best software development practices and discussed how these practices can help researchers to develop and evolve sustainable RS.

## 3.4 Interview

*3.4.1 Research Questions.* The research questions are designed to understand the knowledge, opinions and interests of researchers on Open Science and RS and its sustainability. Also, the questions aim to identify the perceived barriers and supporting factors that influence adopting sustainable software practices in scientific research. In this paper, we present and discuss the results of **R3–R6**, the questions that address sustainable RS.

**RQ1: What Open Science practices do researchers use?** - This question investigates the general practices used during research even if they are not explicitly related to RS development.

**RQ2: What do researchers think about Open Science?** - This question investigates the researcher's perspective on the concepts and practices related to Open Science in order to analyze their knowledge about the subject.

**RQ3: What do researchers think about Sustainable Research Software?** - This question investigates the researcher's perspective on the practices related to Sustainable RS in order to analyze their use of these practices.

**RQ4: What are the barriers that prevent the adoption of practices that increase Research Software sustainability?** - This question investigates the reasons that make it difficult to adopt practices that help RS to be sustainable.

**RQ5: What are the supporting factors for adopting practices that help Research Software be sustainable?** - This question investigates the reasons in favor of the adoption of practices that help RS be sustainable.

**RQ6: Is it beneficial for researchers to make their software sustainable?** - This question aims to evaluate the researcher's interest in making their software sustainable.

*3.4.2 Participants.* For the interview study, we plan to use convenient sampling because of the general lack of availability and narrow time limit of researchers with such profile.

*3.4.3 Interview Guide.* In order to collect data for this study, semi-structured interviews will be conducted with the participants. We designed an interview guide to ensure that the topics are covered during each interview while still allowing flexibility in the conversation. The questions encourages an organic exchange of ideas while following a structured framework that facilitated later comparisons among participants' statements. The interview guide was developed based on the research questions and included demographic questions and open-ended questions to allow participants to share their thoughts. The guide also includes directives that provide information about what is expected from each question and follow-up questions to encourage participants to elaborate on their answers. The interview questions used in the study can be found in Appendix A.

*3.4.4 Data Analysis.* The analysis of the interview data involves a coding process to identify patterns. We resort to a qualitative coding procedure [4] combining two strategies (1) a concept-driven way, i.e. based on what we already know; (2) a data-driven way, i.e. by letting the coding emerge from the responses. We defined four main categories to facilitate data analysis and interpretation based on our research questions and themes of interest namely "Challenges", "Opportunities", "Open Science Practices", "Open Science Knowledge".

First, the interview will be transcribed precisely to document the participant's exact words and preserve the integrity of the data. Then, the coding labels will be defined based on the data and our understanding of the subject.

# 4 THE PILOT INTERVIEW STUDY

The purpose of the pilot study was to test the interview guide and refine the research methodology before conducting a larger study. The pilot study followed the design presented in Section 3 but only the interview results are reported in this paper.

For the pilot study, a single participant was interviewed. The participant was selected based on their alignment with the criteria established for participants inclusion and easy access: a senior researcher in Applied Physics with over 30 years of research experience. The participant's academic background includes a technical course related to computing. His interest in RS development is motivated by a desire to utilize computing as a tool to advance knowledge in his field.

The study started with a virtual meeting held on November 29, 2022 to perform an open-ended interview with the researcher following the interview guide. The interview was recorded and fully transcribed. Through an iterative process, the transcripts were reviewed and relevant codes were assigned to specific segments of the text. Each code is related to one of the four main categories ("Challenges", "Opportunities", "Open Science Practices", "Open Science Knowledge"). In each iteration the coding structure was being refined to develop a coding framework. The total of 32 coding labels identified during the coding process are presented on Appendix B. Unless otherwise stated, the quotes presented in this paper are from the transcription of the pilot interview.

The findings derived from the coding process provide valuable insights into the participant's perspective on Research Software and its Sustainability. A light discussion followed the interview, based on the answers for questions related to RQ3: "What do you think about sustainable RS?"

## 4.1 Coding Framework

The coding structure defined after the analysis of the segments were organized in four categories. The description of coding labels identified more often are listed below.

*4.1.1 Challenges.* This category groups factors and challenges related to adopting sustainable RS practices. The three with more citations during the interview are:

- **Institution Support**: Refers to the absence of formal structures and mechanisms within academic institutions and agencies to regulate and motivate researchers to engage in sustainable RS practices;
- **Concerns about openness**: Reflects the apprehensions expressed by the researcher regarding potential risks and uncertainties associated with making their research open and accessible before its the final version;
- **Developers' expertise**: Refers to the research developers' knowledge and understanding of best code development practices.

*4.1.2 Opportunities.* This category groups factors that support the adoption of practices to help RS to be more sustainable:

- **Positive reaction to practices**: Refers to the researcher's positive response upon being introduced to RS sustainability practices;
- **Addressing errors**: Refers to the possibility of research replication and identifying errors;
- **Collaboration**: Recognition of the potential contributions and improvements through collaboration.

*4.1.3 Open Science Practices.* This category groups practices related to Open Science currently applied by the research group. The eight practices related to sustainable RS with more citations during the interview are:

- **Public code**: Making RS code accessible and available;
- **Versioning**: Management of software releases;
- **Licensing**: Process of defining the permissions and restrictions for the use, modification, and distribution of the RS;
- **Reproducibility**: Refers to what extent the research findings and outcomes can be reproduced using the provided RS;
- **Documentation**: Existence of detailed information that facilitates the software's understanding and utilization by other researchers;
- **Software Reuse**: Practicing designing and developing RS with the intention of allowing its utilization and adaptability by other people;
- **Testing and Validation**: Process of designing and implementing test cases, executing them on the software, and verifying whether the software meets its intended purpose;
- **Credits to Software Collaborators**: Recognition of the contributions made by individuals that provided code contributions or ideas to the RS.

*4.1.4 Open Science Knowledge.* This category groups the categories of knowledge the participant has about practices for sustainable RS:

- **General knowledge**: Refers to the general knowledge about sustainable RS;
- **Licensing knowledge**: Refers to the knowledge about software licenses;
- **Software Contribution Skills**: Refers to knowledge about how to contribute to an open source software;
- **Automated testing**: Refers to knowledge about improving the tests by making they automated.

## 4.2 Results from the Interview

*4.2.1 What does the researcher think about Sustainable Research Software?* The researcher expresses his admiration for efforts to make the software more sustainable, acknowledging its significant impact on their work and the wider research community. The individual emphasizes the need for guidance and support in the early stages of adopting sustainable practices due to their research group developers' need for more expertise. The group members would value having some support in understanding and implementing these practices effectively.

Overall, the researcher's opinions strongly endorse sustainable RS for the research group. He recognizes the value in their work and the desire to contribute to the broader research community. Their belief in the benefits of sustainability, coupled with their enthusiasm for making software accessible and their acknowledgment of the importance of replication, further underscores the researcher's support for sustainable RS practices.

*4.2.2 What are the barriers that prevent the adoption of practices to help RS be sustainable?* The researcher expressed concerns regarding the need for an institutional hosting server for the research's code and data, and policies to support the development and maintenance of sustainable RS. Moreover, staff support and guidelines are absent for prioritizing and encouraging sustainable practices in their research.

The findings also highlight the considerations and dilemmas related to the desire for openness, the protection of their intellectual contributions, and the fear of compromising the integrity and credibility of their research if shared prematurely.

The lack of familiarity of RS developers with best practices in code development is also considered a barrier. Their academic background includes a mathematician and a biologist, and their primary expertise relies on their respective research domains rather than software engineering.

*4.2.3 What are the supporting factors for adopting practices that help RS be sustainable?* The researcher initially lacked knowledge about these practices but showed enthusiasm and accepted the presented practices once we introduced them. The response indicates a recognition of the benefits and importance of adopting sustainability practices in their RS development.

Another supporting factor identified is the commitment to promoting reproducibility in their research. With the data and code publicly available, the research can be replicated and potentially expose any errors or inconsistencies, contributing to scientific knowledge's overall reliability and robustness.

By adopting sustainable practices, researchers also create an opportunity for other experts in the field to collaborate on software development, provide feedback and suggest improvements. Moreover, disseminating research findings can extend the impact beyond the research community, benefiting the general public.

*4.2.4 Is it beneficial for the researchers to make their software sustainable?* Their positive attitude towards sustainable RS reflects a genuine appreciation for the benefits it offers to their work. The researcher wanted to see their software made available to everyone, acknowledging its potential to help numerous researchers. They emphasize the practical applications of research findings in clinical settings, where the software could be used to evaluate and treat people effectively.

Additionally, the researcher highlights the importance of replication in research and sustainable software's role in ensuring accurate and reliable results, as he recognizes that errors and bugs are inherent in software. They are keen on addressing and fixing them. The researcher also recognizes the benefits of automated tests in ensuring that new software versions are reliable and addressing previous bugs, thus saving valuable time and effort. The researcher understood the benefits of making RS sustainable and collaborating to advance scientific knowledge.

## 5 DISCUSSION

The pilot study assessed the feasibility of an investigation with research groups from a Brazilian university. The interview for the pilot study included questions to allow the participant to share his opinion and recommendations for improving future interviews.

The positive feedback from the participant underscores the value of the interview process and how it was conducted. The researcher expressed that the format provides an opportunity to reflect on their own practices and motivates them to explore topics related to RS sustainability practices. Also, the discussions contribute to researchers' understanding of the subject. The participant's comments suggest that it can be a valuable tool for promoting awareness, knowledge sharing, and behavior change within the research community. From our perspective, the pilot study confirmed that the semi-structured interview format with researchers is feasible and preferred over the survey, as it brings some flexibility and allows rich discussions.

On the other hand, the participant noted that the length could be an issue as researchers only have little time available. The interview took 65 minutes. Additionally, the participant suggested more questions related to how the interviewee thinks the institution could help to promote RS practices in academia.

From our perspective as interviewers, we recognize that enhancing the interview experience and preventing it from being long could be achieved by sharing relevant materials containing key concepts before the interview. For instance, while talking about licenses, the researcher was not aware that there were licenses in the research software developed by his own group; he only realized it when we asked about the process of publishing the software. This lack of awareness may explain the need for follow-up questions and clarification of concepts.

The research group leader reported about lack of institutional supporting infrastructure for research groups that develop their own software. According to Katz et al. [3], universities, funding agencies and other kinds of organizations that belong to the scientific ecosystem should be prepared and invest resources in addressing software research, open science policies, practices, training and scientific infrastructures.

We recognize that institutional infrastructure is required to support RS, for instance, institutional digital repositories are necessary to researchers and the research community. Moreover, universities may invest resources to provide a team of RS engineers to support research groups. We plan to review the interview instrument to address this specific concern.

*Limitations.* The pilot study assessed the feasibility of an investigation with feedback from the main investigator of one research group from a Brazilian university. The preliminary findings are limited to one researcher's opinions, in specific time, place, and condition, and do not imply any generalization nor provide a comprehensive view of the challenges and practices across different research groups or domains. Future interviews should include researchers from different study fields with some software development experience.

## 6 CONCLUSION

Universities and other kinds of organizations that belong to the scientific ecosystem are should be prepared and invest resources in addressing software research, open science policies, practices, training and scientific infrastructures [3].

In this paper, we describe a pilot study that resorted to semi-structured interview as a method to investigate research software sustainability in research groups from UFBA. Such interview supported our exploration of participant's experiences, opinions and challenges, and offered a dynamic and interactive space for discussion. We highlight the importance of a talk about Open Science and RS prior to the interview to introduce the fundamental concepts and examples of the practices. Moreover, it may enhance the interview experience and prevent it from being long.

The study and its findings helped us structuring a preliminary coding framework. We intend to use it to perform interviews with other research groups from different domains to understand the current practices adopted, the challenges and supporting factors to promote the development of sustainable RS at UFBA.

We hope that the preliminary results of our pilot study also inspire researchers to reflect upon the sustainability of their software and the software engineering practices that can help them to achieve it.

# A INTERVIEW INSTRUMENT

(1) Which institute are you affiliated with?
(2) What is your academic formation?
(3) What is your research area?
(4) Are you a researcher?
(5) How many years of career as a researcher?
(6) Do you allow virtual access to scientific articles published in academic journals?
(7) Do you define the license when articles are made available?
(8) Are experimental elements (resources, algorithms, methods) available?
(9) Do you define license when elements are made available?
(10) Are preliminary versions of research manuscripts published openly?
(11) Can the data and other materials used in the research be reused without your permission?
(12) Are the tools and software used in your scientific research open source?
(13) What do you think about making your experiment data and details available so that other researchers can reproduce it and verify your research?
(14) In your opinion, what is Open Science?
(15) Which Open Science practices are you aware of and/or use?
(16) In your research practice, have you ever encountered obstacles to your daily work?
(17) If you need to query data from more than two or three years ago, will you still be able to understand and use it?
(18) Do you know or have you used any tool that supports Open Science?
(19) How much effort do you put into verifying that your inputs or outputs can be reused or shared?
(20) In your opinion, to which public should Science be open?
(21) In your opinion, what are the reasons for using Open Science practices?
(22) In your opinion, what are the reasons for not using Open Science practices?
(23) What could be done at the research institution to adopt some practices?
(24) Would you as a researcher adopt Open Science for your research and development environment?
(25) Could you tell me a about how the software to be used in the research is chosen?
(26) What do you think about using open source software in your research?
(27) What do you think about using software developed by your group in other research?
(28) In your opinion, is it possible for an external researcher to reproduce your research?
(29) In your opinion what can be done to ensure that the software does what is expected of it?
(30) Are the softwares used in the research cited?
(31) What is the background of the people who develop software in your group?
(32) On average, how many people are involved in the software development process?
(33) Are the costs related to the development/maintenance of the software accounted for in the research budget?
(34) Have you ever had problems with your computer that led to loss of important data or information about your search?
(35) Would you like to receive input on the software from people outside the group?
(36) In your opinion, would your research benefit from using software with sustainable practices?
(37) What do you think about adopting some practices to make software sustainable?
(38) Can you name two tools used?

# B INTERVIEW CODING FRAMEWORK

- **Challenges**: Publishing constraints, Financial constraints, Time constraints, Motivation constraints, Institution support, Concerns about openness, Developers' expertise, Access challenges, Funding for developing software, Research software Sustainability constraints
- **Opportunities**: Positive reaction to practices, Addressing errors, Collaboration
- **Open Science practices**: Public code, Public data, Versioning, Licensing software, Workflow tools, Reproducibility, Availability, Documentation, Software Reuse, Sharing knowledge, Testing and Validation, Pair programming, Code review, Credits to software collaborators
- **Open Science knowledge**: Publicizing data, Licensing knowledge, General knowledge, Software Contribution Skills, Automated testing

# REFERENCES


[1] J.C. Carver, N. Weber, K. Ram, S. Gesing, and D.S. Katz. 2022. A survey of the state of the practice for research software in the United States. *PeerJ Comput. Science* (2022). https://doi.org/10.7717/peerj-cs.963

[2] Caroline Jay, Rawan Sanyour, and Robert Haines. 2022. *"Not everyone can use Git"': Research Software Engineers' recommendations for scientist-centred software support (and what researchers really think of them).* Technical Report. University of Manchester. DOI: 10.48420/17313215.v1.

[3] D.S. Katz et al. 2021. Addressing Research Software Sustainability via Institutes. *arXiv:2103.03690 [cs]* (March 2021). https://doi.org/10.1109/BoKSS52540.2021.00013 arXiv: 2103.03690.

[4] Margrit Schreier. 2012. *Qualitative Content Analysis in Practice* (1st ed.). SAGE Publications Ltd, USA.

[5] C. Venters, S. Akinli kocak, S. Betz, I. Brooks, R. Capilla, R. Chitchyan, L. Duboc, R. Heldal, A. Moreira, S. Oyedeji, B. Penzenstadler, and J. Porras. 2021. Software Sustainability: Beyond the Tower of Babel. https://doi.org/10.6084/m9.figshare.14370611.v1

[6] Igor Wiese, Ivanilton Polato, and Gustavo Pinto. 2020. Naming the Pain in Developing Scientific Software. *IEEE Software* 37, 4 (2020), 75–82. https://doi.org/10.1109/MS.2019.2899838

[7] Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. Good enough practices in scientific computing. *PLOS Computational Biology* 13, 6 (06 2017), 1–20. https://doi.org/10.1371/journal.pcbi.1005510