**Scientific Collaboration Networks on Software**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Team Member: **Swetha Kolloju**

POTENTIAL ADVISOR: **Dr.Sarah Bratt**

Date: **04-01-2025**

Contents

[1. Executive Summary 3](#_Toc194588092)

[2. Literature Review/Market research 3](#_Toc194588093)

[3. Research Project Deliverables 4](#_Toc194588094)

[4. Project Timeline & Gannt Chart 5](#_Toc194588095)

[5. Ethics 5](#_Toc194588096)

[6. Approvals 7](#_Toc194588097)

[7. Appendix 7](#_Toc194588098)

[A. Advisor Engagement 7](#_Toc194588099)

[1) Project Team Responsibilities 7](#_Toc194588100)

[2) Faculty Advisor Responsibilities 8](#_Toc194588101)

[B. Ground Rules 8](#_Toc194588102)

# Executive Summary

This project examines the factors influencing software adoption within the field of evolutionary biology, specifically focusing on phylogenetic research. The study analyzes 1.2 million software mentions to understand how software tools are selected by researchers over time. By looking at the clustering of software use across different geographic, linguistic, and academic communities, the project aims to uncover how these factors shape the research process, community dynamics, and career trajectories within the field. The ultimate goal is to provide insights into how software choices impact scientific progress and the mobility of researchers.

The need for this research arises from the limited understanding of how software preferences, often shaped by geography, language, and academic lineage, influence the development of research communities and scientific outcomes. Despite the growing importance of open-source tools in scientific research, there is a lack of systematic studies that address how the use of different software packages may hinder or accelerate the pace of scientific discovery. By exploring these patterns, this study will highlight the potential inefficiencies or benefits of software clustering in evolutionary biology, providing valuable knowledge for optimizing research practices and institutional mobility. Understanding these dynamics will have broader implications for scientific communities beyond evolutionary biology.

The project will involve analyzing software usage data, identifying clustering patterns, and studying the influence of geographic, linguistic, and genealogical factors on software adoption. Data collection will be assisted by the professor, while the primary analysis and development will be carried out by me. By the end of the semester, we expect to have completed a comprehensive report outlining the clustering of software communities, their impact on scientific progress, and the career implications for researchers. Table 1 shows a preliminary division of tasks, where I will focus on the data, Network analysis and synthesis of findings, while my professor will assist with understanding of data, data collection, Network analysis etc.

|  |  |
| --- | --- |
| Team Member | Feature responsibility |
| *Swetha* | *Data and Network Analysis* |

Table 1 Preliminary Subsystem Responsibilities

# Literature Review/Market research

Research on software adoption in scientific fields has expanded, especially concerning open-source tools and their impact on collaboration. While studies have examined the role of geographic and technical factors in software selection, little attention has been given to how linguistic factors influence these choices. The open science movement has made scientific tools more accessible, yet regional and language-based clustering of software use still persists, creating potential barriers to collaboration and innovation. This research seeks to fill this gap by exploring how software usage in evolutionary biology, particularly in phylogenetic studies, is influenced by linguistic, geographic, and academic factors.

Existing research shows that the open-source movement has significantly increased software accessibility, allowing tools to be shared globally via platforms like GitHub. However, many studies overlook how shared linguistic backgrounds influence the adoption of certain tools. For example, researchers in English-speaking countries may gravitate toward software with English-based documentation, which can limit access for non-English-speaking researchers. Despite the global nature of scientific research, the persistence of linguistic barriers may create insular software communities that slow the exchange of ideas and limit the tools available to researchers in non-English-speaking regions.

Our project will investigate whether software is used in evolutionary biology clusters along linguistic and geographic lines. We will explore whether software preferences within these communities’ foster innovation and collaboration or contribute to inefficiencies. By examining these patterns, we aim to understand whether greater linguistic and software diversity correlates with more innovative research outcomes.

Interviews with potential users in evolutionary biology reveal significant pain points related to language barriers, software interoperability, and regional software preferences. This research will provide valuable insights into how these factors shape research practices and career trajectories, aiming to create a more inclusive and efficient scientific community.

# Research Project Deliverables

This project will examine how scientific software is adopted and used in evolutionary biology, focusing on patterns influenced by geography, language, and academic networks. The final outcome will include a research paper, statistical analyses, and visual representations of key findings.

**Final Presentation Format**

The main deliverable will be a research paper structured with an introduction, methods, results, and discussion. The paper will be under ten pages and will include at least five peer-reviewed sources. It will also feature two or more graphics to illustrate key findings, such as network maps showing software adoption trends and statistical models analyzing career impacts. Additionally, all data and analysis scripts will be made publicly available on GitHub to ensure transparency and encourage further research.

**Analysis Plan**

Once the data is collected from Constellate and OpenAlex, several key analyses will be performed:

**Tracking Software Mentions:** Identifying how often different phylogenetic software tools appear in nearly 200,000 scientific papers from 1990 to 2024.

**Network Mapping**: Examining how researchers are connected through shared software use and academic collaborations.

**Geographic and Language Trends:** Investigating whether software adoption follows regional or linguistic patterns.

**Statistical Modeling:** Using fixed-effects regression to explore whether scientists who use multiple software tools produce more impactful research or have different career trajectories.

**Handling Challenges**

To ensure reliable results, software mentions will be manually reviewed for accuracy. If statistical models do not show strong patterns, alternative methods, such as qualitative case studies, will be considered. If data access issues arise, additional sources of open-access research papers will be explored to maintain project feasibility.

# Project Timeline & Gannt Chart

|  |  |
| --- | --- |
| Milestone | Date |
|  |  |
| Project Meeting | 03/10/25 |
| Data Collection Phase Begins | 03/24/25 |
| Data Preprocessing & Cleaning Completed | 03/28/25 |
| Software Mention Extraction Completed | 04/01/25 |
| Signed proposal submission | 04/04/25 |
| Statistical Analysis of Data | 04/08/25 |
| Author-End relations | 04/08/25 |
| Introduction to Network Analysis | 04/09/25 |
| Network Analysis on Data | 04/20/25 |
| Draft of Research Paper(Not included in Capstone) | 04/26/25 |
| Poster Demo | 05/02/25 |
| iShowcase | 05/09/25 |
| Research Paper draft for iConference (Not included in Capstone) | 05/09/25 |

Table 3: Milestone Schedule

# Ethics

*This project examines how often different phylogenetic software programs are mentioned in scientific articles written in various languages. The data comes from widely available academic sources like JSTOR, Portico, and OpenAlex. Since the study focuses on published research, it does not involve direct interaction with people or collect any personal information. However, it is important to consider ethical concerns, such as the possibility of bias in data analysis, misinterpretation of results, or unintended consequences in how the findings are used.*

|  |  |  |  |
| --- | --- | --- | --- |
| # | Question | Generally | Data Breach |
| 1 | |  |  |  | | --- | --- | --- | | Could this research be used to misrepresent software usage trends in a misleading way? |  |  | | M | N |
| 2 | |  |  | | --- | --- | |  | Could linguistic biases in software mentions be misinterpreted as evidence of exclusion or discrimination? | | M | N |
| 3 | |  |  | | --- | --- | |  | Could the dataset's limitations lead to inaccurate conclusions about software popularity? | | Y | N |
| 4 | |  |  | | --- | --- | |  | Could this research be used to unfairly criticize certain linguistic or geographic scientific communities? | | M | N |
| 5 | |  |  | | --- | --- | |  | Could the findings be used to influence funding or policy decisions in a way that disadvantages certain researchers? | | M | N |
| 6 | |  |  | | --- | --- | |  | Does the methodology ensure transparency in data collection and processing? | | Y | N |
| N | |  |  | | --- | --- | |  | Could software providers misuse the findings to promote their tools over others unfairly? | | M | N |
| 8 | |  |  | | --- | --- | |  | Are there potential biases in how software mentions are extracted and analyzed? | | Y | N |
| 9 | |  |  | | --- | --- | |  | Could this research contribute to reinforcing existing inequalities in scientific software access? | | M | N |
| 10 | |  |  | | --- | --- | |  | Are all sources properly credited and cited to maintain academic integrity? | | Y | N |
| 11 | |  |  | | --- | --- | |  | If errors exist in data extraction, could they lead to misleading results? | | Y | N |
| 12 | |  |  | | --- | --- | |  | Are there ethical concerns in using JSTOR, Portico, and OpenAlex data for this purpose? | | N | N |
| 13 | |  |  | | --- | --- | |  | Could this study be misinterpreted to suggest a hierarchy of software tools based on language? | | M | N |
| 14 | |  |  | | --- | --- | |  | Could the findings be used to justify restrictive policies on software adoption in certain regions? | | M | N |
| 15 | |  |  | | --- | --- | |  | Could this study unintentionally privilege proprietary software over open-source alternatives? | | M | N |
| 16 | |  |  | | --- | --- | |  | Are limitations of the dataset and methodology clearly stated to prevent misinterpretation? | | Y | N |

# Approvals

The signatures of the people below indicate an understanding of the purpose and content of this document by those signing it. By signing this document, you indicate that you approve of the proposed project outlined in this Statement of Work, the division of work, the Ground Rules and that the next steps may be taken to create a Product Specification and proceed with the project.

This document is based upon and supersedes the *<Project proposal – Scientific Collabaration Networks on Software> Version 1.0.* Deviations, (versus clarifications), from the PDR have been clearly noted. For any requirements not listed in this SOW, the PRD requirements shall remain in effect.

|  |  |  |  |
| --- | --- | --- | --- |
| Approver Name | Title | Signature | Date |
| *SWETHA KOLLOJU* | Team Member |  | 04/03/25 |
| *Dr SARAH BRATT* | Advisor |  |  |
| *Dr GREG CHISM* | Instructor |  |  |

# Appendix

## Advisor Engagement

## Project Team Responsibilities

* As individual team member, I will set up and facilitate a weekly call/meeting with the Faculty Advisor. I will provide weekly status updates, including upcoming deliverables, critical issues, and any adjustments to the Project Plan.
* Documents will be provided to the Faculty Advisor with adequate time for review and signature. The review time will be agreed upon, with the minimum review period being 3 days before the document due date.
* Design files will be provided to the Faculty Advisor as requested, in an agreed-upon format.
* Any support required from the Faculty Advisor will be requested with clear deadlines and an adequate timeframe for fulfillment.
* Any modifications requested by the Faculty Advisor to the Project Plan will be reviewed and agreed upon within 1 week of the request.

## Faculty Advisor Responsibilities

* The Faculty Advisor will provide knowledge and expertise to help me develop the necessary skills for the project.
* The Faculty Advisor will participate in a weekly or bi-weekly call/meeting to review the project status, upcoming deliverables, priorities, issues, and progress.
* The Faculty Advisor will provide document review, feedback, and approval, rejection, or approval with contingencies, with adequate time for me to meet the course due dates.
* The Faculty Advisor will provide feedback and guidance on design implementation decisions, design files, test plans, test procedures, and results.
* The Faculty Advisor will provide technical advice and guidance for approximately 1 hour per week.
* Modifications to the Project Plan will be resolved and documented within 1 week of the request.
* The Faculty Advisor will grade the final project using a skill-based rubric.
* The Faculty Advisor will attend the iShowcase in May.

## Ground Rules

Since I am the only team member, I will follow these ground rules to ensure the smooth progression of the project:

1. **Stay focused on objectives and goals.** At the start of each meeting or session, I will clearly define the objectives and desired outcomes. I will make sure to stay on track, even if distractions arise.
2. **“Sidebar” non-relevant issues.** If important matters arise that are not directly related to the immediate goals of the project, I will note them down and address them at a later time.
3. **Listen actively.** I will listen carefully when receiving feedback or guidance from the Faculty Advisor, considering all input before responding.
4. **Ensure all viewpoints are considered.** While I am the sole team member, I will ensure that the advisor’s viewpoints are heard, respected, and considered in the decision-making process.
5. **Respectful discussion of differences.** If there are disagreements, I will ensure that any differences are addressed respectfully and constructively, focusing on finding solutions rather than assigning blame.
6. **Look for value in new ideas.** I will keep an open mind and explore the value in any new ideas or suggestions that arise during the project.
7. **Focus on future objectives, not past issues.** I will use past experiences to inform decisions but will always focus on moving the project forward.
8. **Agree on action items and next steps.** After each discussion or meeting, I will summarize the next steps and action items and ensure clarity on what needs to be done.
9. **Accountability.** I will take full responsibility for all aspects of the project, ensuring that I meet deadlines, deliver quality work, and fulfill my responsibilities.