PhD position — Conceptual and empirical models of long-term collaboration in free and open source software

Key information

- Scientific fields: computer science, data science, computational social science
- Research laboratory: LTCI, Télécom Paris, Polytechnic Institute of Paris, Palaiseau, France
- Advisors: Stefano Zacchiroli, Théo Zimmermann, Roberto Di Cosmo
- Start date: 2026Q4 (October–December)
- Duration: 3 years (thesis defense expected in 2029Q4)

Summary

This PhD project aims to develop a **multi-level conceptual and empirical framework** for understanding long-term collaboration in Free and Open Source Software (FOSS). It combines large-scale data analysis—leveraging the Software Heritage archive—with theory-building to model collaboration as a **sociotechnical system** operating across two levels:

- **Micro level:** individual contributor trajectories (entry, progression, barriers, retention)
- **Macro level:** ecosystem-wide dynamics (project structures, technology evolution, contribution patterns)

A key contribution is to **bridge these levels** by identifying feedback mechanisms between individual behaviors and global structures. The project ultimately seeks to produce both **theoretical models and actionable insights**, enabling improved sustainability, inclusiveness, and coordination in open-source ecosystems.

Context and motivation: towards a sociotechnical theory of long-term collaboration on open source

Free and Open Source Software (FOSS) has become a foundational layer of modern digital infrastructure. Over time, it has evolved from small, cohesive communities into a vast, heterogeneous ecosystem composed of millions of contributors, projects, organizations, and platforms. Collaboration in this ecosystem is inherently long-term, distributed, and mediated by complex engineering infrastructure such as version control systems, issue trackers, and code review platforms.

Despite extensive empirical work on FOSS, current research remains fragmented. Most large-scale studies focus on a single platform (e.g., GitHub), which in-

troduces significant biases and provides only a partial view of the ecosystem. As a result, we still lack a comprehensive understanding of long-term trends in collaboration, technology adoption, and project sustainability across the full diversity of FOSS.

The Software Heritage archive provides a unique opportunity to overcome these limitations. By collecting and preserving the full development history of publicly available source code across multiple forges and decades, it enables, for the first time, a unified and longitudinal analysis of FOSS at ecosystem scale.

At the same time, there is a growing need for **conceptual and sociotechnical frameworks** that go beyond measurement to explain how collaboration emerges, evolves, and sometimes fails over long time horizons. This PhD aims to address this gap by combining large-scale empirical analysis with theory-building.

The central hypothesis is that FOSS collaboration can be understood as a **multi-level sociotechnical system**, in which ecosystem-level structures and trends interact with individual contributor trajectories. To study this system, we articulate two complementary levels of modeling—micro and macro—and focus on the mechanisms that connect them.

Research objective: a multi-level conceptual framework for FOSS collaboration

The core objective of this thesis is to construct, formalize, and validate a **multi-level conceptual framework** for long-term collaboration in FOSS.

This framework will:

- Identify the key entities, roles, and interactions involved in collaboration
- Model the mechanisms through which coordination occurs over time
- Explain how local behaviors aggregate into global structures
- Support the design of interventions aimed at improving sustainability and inclusiveness

The framework is structured around two complementary levels of analysis.

Micro-level model: contributor trajectories and local coordination

At the micro level, collaboration is modeled as the evolution of individual contributors within and across projects.

We model collaboration as **trajectories**: from first contribution to sustained participation (or disengagement), through a sequence of roles and interactions.

Key dimensions include:

- **Entry and progression:** how contributors move from newcomer to regular contributor to maintainer

- **Critical transitions:** onboarding, first accepted contribution, increasing responsibility, or dropout
- **Barriers along pathways:** friction points such as lack of feedback, complex tooling, unclear norms, or gatekeeping
- **Differential impact:** how these barriers affect contributors unevenly (e.g., by experience, geography, or underrepresented status)

This perspective enables the identification of **typical pathways and failure modes**, and how they relate to **diversity and retention**. In particular, it will help characterize where and for whom participation breaks down, and which project practices are associated with more inclusive and sustainable trajectories. Furthermore, by not focusing on individual projects separately, but looking at trajectories in the whole Software Heritage archive, we can identify complete contributor pathways and take into account the influence and barriers of the various projects a contributor participates in.

Macro-level model: ecosystem-wide coordination and structural dynamics

At the macro level, the FOSS ecosystem is modeled as a **large-scale, evolving system of projects, ecosystems, and technologies**.

The objective is to produce a **systematic, longitudinal characterization of the "state of open source"**, grounded in reproducible measurements over the full Software Heritage archive.

Key dimensions include:

- **Population structure:** number and types of projects, active contributors, and their evolution over time
- **Contribution patterns:** distribution of work (e.g., concentration, inequality, bus factor)
- **Technology trends:** evolution of programming languages, tools, and platforms
- **Mobility and connectivity:** how contributors move across projects and how projects are interlinked
- **Temporal dynamics:** long-term trends, cycles, and responses to external shocks

This model aims to identify **robust empirical regularities** and to quantify biases in platform-specific studies by providing a comprehensive, ecosystem-wide perspective.

A key output will be a **reproducible measurement framework enabling periodic "state of open source" reports**, for consistent tracking of the evolution of FOSS collaboration over time.

Bridging micro and macro: mechanisms of sociotechnical coordination

A central contribution of this thesis is to articulate how micro- and macro-level dynamics influence each other over time.

On the one hand, **macro-level structures and trends** shape individual trajectories. Changes in programming languages, the emergence of new ecosystems, shifts in platform usage, or the adoption of governance practices (e.g., codes of conduct) affect how contributors enter projects, interact, and remain engaged.

On the other hand, **micro-level events can have long-term systemic effects**. For example, maintainer dropout, contributor concentration, or the loss of key expertise can impact project sustainability and propagate to the broader ecosystem over time.

These interactions define a set of **sociotechnical feedback mechanisms**, in which individual behaviors and structural conditions co-evolve. Some of these mechanisms can be studied using causal inference approaches (e.g., assessing the impact of contributor dropout or external shocks), while others emerge from the aggregation of trajectories at scale.

By explicitly modeling these relationships, the thesis aims to connect individual experiences of collaboration with ecosystem-level outcomes, and to explain how long-term coordination is maintained—or disrupted—in FOSS.

From conceptual models to interventions

The ultimate goal of this framework is not only descriptive but also actionable.

By identifying the mechanisms that underpin successful or fragile forms of collaboration, the thesis will support the design of **theory-informed interventions**, such as:

- Improving contributor onboarding and retention strategies
- Detecting early warning signals of project fragility
- Informing the design of collaborative tools and platforms

These interventions will be grounded in the conceptual framework and evaluated using empirical data, closing the loop between theory and practice.

Methodological approach

The development and validation of the proposed conceptual framework rely on a combination of complementary approaches.

The thesis will leverage the Software Heritage archive to conduct large-scale empirical analyses of FOSS collaboration across projects, platforms, and decades. Collaboration structures will be modeled using graph-based representations (e.g., contributor–project networks), while contributor trajectories will be analyzed

using temporal and sequential methods to capture patterns of engagement and disengagement.

Where appropriate, econometric techniques will be used to estimate causal effects, particularly for studying the impact of shocks (e.g., contributor dropout, external events) on projects and ecosystems. These approaches will be complemented by other methods to ensure robustness and relevance.

Finally, conceptual models may be operationalized through simulation or generative approaches to explore how micro-level behaviors give rise to macro-level phenomena and to evaluate potential interventions.

Technical challenges and enabling infrastructure

Studying FOSS collaboration at this scale raises significant computational and methodological challenges.

A key component of the thesis will be the development of tools and representations that enable efficient analysis of Software Heritage data. This includes:

- Scalable data processing pipelines
- Intermediate representations adapted to data science workflows
- Methods for identity resolution across platforms
- Techniques for handling heterogeneity and temporal biases

These contributions are essential to make large-scale sociotechnical analysis feasible and reproducible.

Expected contributions

This thesis will contribute:

- A **multi-level conceptual framework** for long-term collaboration in FOSS
- Empirical characterizations of contributor trajectories and ecosystem dynamics at unprecedented scale
- Methodological advances for large-scale analysis of software repositories
- Practical insights and tools to support the sustainability of FOSS ecosystems

---

Alignment with PEPR eNSEMBLE – PC PILOT (Thème 3)

This project aligns with **Thème 3: "Cadres conceptuels pour la collaboration à long terme"**, as it directly addresses the need for new conceptual and analytical frameworks to understand evolving collaborative practices in complex sociotechnical environments.

First, the thesis explicitly focuses on **long-term, distributed collaboration mediated by multiple tools and platforms**, which is at the core of the PILOT program's concerns. By moving beyond platform-specific analyses and leveraging the Software Heritage archive, the project captures the **multi-tool, multi-organization nature of collaboration**, a key challenge highlighted in the call.

Second, the proposed **multi-level conceptual framework** (micro trajectories + macro ecosystem dynamics) directly contributes to the development of **new theoretical constructs for analyzing coordination mechanisms over time**. In particular, the modeling of roles, interactions, contributor pathways, and structural dynamics addresses the call's emphasis on:

- asymmetries in roles and participation,
- evolving organizational forms,
- and the need to understand coordination in complex artifact ecologies.

Third, the project strongly embodies the **sociotechnical perspective** promoted by PILOT. It explicitly connects:

- **social dimensions** (onboarding, diversity, contributor retention, governance practices),
- with **technical infrastructures** (version control systems, platforms, programming-language-specific infrastructure), and studies their co-evolution through feedback mechanisms. This matches the objective of articulating **social needs with technological capabilities**.

Finally, the thesis goes beyond analysis by aiming to produce **actionable, theory-informed interventions** (e.g., improving onboarding, detecting fragility, informing tool design), which is fully consistent with PILOT's goal of informing **software engineering and design practices** through conceptual advances.

Candidate profile

We are seeking a data enthusiast who loves both the theoretical challenge of designing efficient algorithms and the practical work of implementing them at scale. The ideal candidate should be excited about extracting insights from massive datasets, particularly graph-structured data like the Software Heritage archive. This position suits someone comfortable at the intersection of algorithms, empirical analysis, and data science.

Required qualifications:

- Master's degree in Computer Science, Software Engineering, Data Science, or a related field
- Strong foundation in algorithms and data structures, particularly graph algorithms

- Programming proficiency and experience with large-scale data analysis
- Strong technical writing skills in English
- Fluency in spoken English; French is not required

Desired skills:

- *Technical:* Python, Rust (or willingness to learn), data science packages (e.g., Pandas or similar), distributed computing frameworks
- *Soft skills:* Analytical thinking, intellectual curiosity, self-motivation, teamwork, communication skills
- *Experience:* Open source development, empirical software engineering, or mining software repositories

Work environment

The PhD candidate will be hosted at the LTCI laboratory of Télécom Paris, as a member of the ACES team (Architecture, Code, and Software Engineering). Close and frequent collaboration with the Software Heritage team at Inria is expected, given the central role of the Software Heritage archive in this research. Powerful computational resources suitable for large-scale data analysis will be available.

Contact

- Stefano Zacchiroli: `<stefano.zacchiroli@telecom-paris.fr>`
- Théo Zimmermann: `<theo.zimmermann@telecom-paris.fr>`

References

- Jergensen, C., Sarma, A., & Wagstrom, P. (2011, September). The onion patch: migration in open source ecosystems. In Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering (pp. 70-80).
- Roberto Di Cosmo, Stefano Zacchiroli. Software Heritage: Why and How to Preserve Software Source Code. iPRES 2017.
- Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli. The Software Heritage Graph Dataset: Public software development under one roof. MSR 2019.
- Bianca Trinkenreich, Mariam Guizani, Igor Wiese, Anita Sarma, Igor Steinmacher. Hidden Figures: Roles and Pathways of Successful OSS Contributors. Proc. ACM Hum. Comput. Interact. 4(CSCW2): 180:1-180:22 (2020).
- Paolo Boldi, Antoine Pietri, Sebastiano Vigna, Stefano Zacchiroli. Ultra-Large-Scale Repository Analysis via Graph Compression. SANER 2020.
- Davide Rossi, Stefano Zacchiroli. Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022.

- Annalí Casanueva, Davide Rossi, Stefano Zacchiroli, Théo Zimmermann. The Impact of the COVID-19 Pandemic on Women's Contribution to Public Code. Empir. Softw. Eng. 30(1): 25 (2025).