

Empirical Software Engineering Research with Software Heritage

Stefano Zacchiroli

Télécom Paris, Polytechnic Institute of Paris
`stefano.zacchiroli@telecom-paris.fr`

28 September 2022



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 1 Datasets
- 2 Accessing source code artifacts
- 3 Software provenance and evolution
- 4 Software forks
- 5 Diversity, equity, and inclusion



Graph dataset

Use case: large scale analyses of the most comprehensive corpus on the development history of free/open source software.



Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli

The Software Heritage Graph Dataset: Public software development under one roof

MSR 2019: 16th Intl. Conf. on Mining Software Repositories. IEEE

preprint: <http://deb.li/swhmsr19>

Dataset

- Relational representation of the full graph as a set of tables
- Available as open data: <https://doi.org/10.5281/zenodo.2583978>
- Chosen as subject for the **MSR 2020 Mining Challenge**

Formats

- Local use: PostgreSQL dumps, or Apache Parquet files (~1 TiB each)
- Live usage: Amazon Athena (SQL-queriable), Azure Data Lake

```
SELECT COUNT(*) AS c, word FROM (  
  SELECT LOWER(REGEXP_EXTRACT(FROM_UTF8(  
    message), '^w+')) AS word FROM revision)  
WHERE word != ''  
GROUP BY word ORDER BY COUNT(*) DESC LIMIT 5;
```

```
SELECT COUNT(*) AS c, word FROM (  
  SELECT LOWER(REGEXP_EXTRACT(FROM_UTF8(  
    message), '^\\w+')) AS word FROM revision)  
WHERE word != ''  
GROUP BY word ORDER BY COUNT(*) DESC LIMIT 5;
```

Count	Word
71 338 310	update
64 980 346	merge
56 854 372	add
44 971 954	added
33 222 056	fix



Stefano Zacchiroli

A Large-scale Dataset of (Open Source) License Text Variants

MSR 2022 (best dataset paper award)

preprint: <https://arxiv.org/abs/2204.00256>

Dataset

- 6.5 million unique full texts of FOSS license variants
- Detected using filename patterns across the entire SWH archive
 - LICENSE, COPYRIGHT, NOTICE, etc.
- Metadata: file lengths measures, detected MIME type, detected SPDX license (via ScanCode), example origin repository, oldest public commit of origin

Use cases

- Empirical studies on FOSS licensing, including phylogenetics
- Training of automated license classifiers
- NLP analyses of legal texts


- 1 Datasets
- 2 Accessing source code artifacts
- 3 Software provenance and evolution
- 4 Software forks
- 5 Diversity, equity, and inclusion



The Software Heritage Filesystem (SwhFS)

The **Software Heritage Filesystem (SwhFS)** is a user-space POSIX filesystem that enables browsing parts of the Software Heritage archive as if it were locally available.

- code: forge.softwareheritage.org/source/swh-fuse
- documentation: docs.softwareheritage.org/devel/swh-fuse

 **Thibault Allançon, Antoine Pietri, Stefano Zacchiroli**
The Software Heritage Filesystem (SwhFS): Integrating Source Code Archival with Development
ICSE 2021: The 43rd International Conference on Software Engineering
<https://arxiv.org/abs/2102.06390>

The Software Heritage Filesystem (SwhFS) — example

```
$ mkdir swarfs
$ swarf fs mount swarfs/ # mount the archive
$ cd swarfs/

$ cat archive/swh:1:cnt:c839dea9e8e6f0528b468214348fee8669b305b2
#include <stdio.h>

int main(void) {
    printf("Hello, World!\n");
}

$ cd archive/swh:1:dir:1fee702c7e6d14395bbf5ac3598e73bcbf97b030
$ ls | wc -l
127
$ grep -i antenna THE_LUNAR_LANDING.s | cut -f 5
# IS THE LR ANTENNA IN POSITION 1 YET
# BRANCH IF ANTENNA ALREADY IN POSITION 1
```



Paolo Boldi, Antoine Pietri, Sebastiano Vigna, Stefano Zacchiroli

Ultra-Large-Scale Repository Analysis via Graph Compression

SANER 2020, 27th Intl. Conf. on Software Analysis, Evolution and Reengineering. IEEE

Research question

Is it possible to efficiently perform software development history analyses at the scale of Software Heritage archive on a single, relatively cheap machine?

Idea

Apply state-of-the-art graph compression techniques from the field of Web graph / social network analysis.

Results

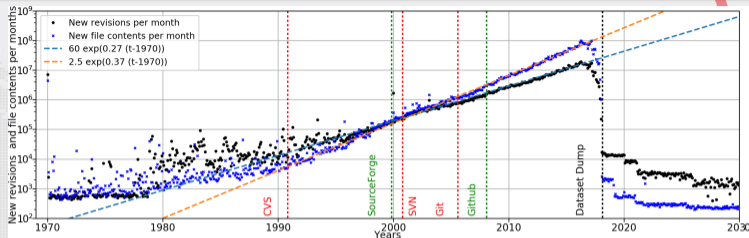
The entire archive graph (25 B nodes, 350 B edges) can be loaded in 200 GiB and then traversed at the cost of tens of ns per edge (= a few hours for a full single-thread visit).

Java and gRPC APIs available: docs.softwareheritage.org/devel/swh-graph/grpc-api.html

- 1 Datasets
- 2 Accessing source code artifacts
- 3 Software provenance and evolution**
- 4 Software forks
- 5 Diversity, equity, and inclusion



Software provenance and evolution



Key findings

- The amount of original commits in public code doubles every ~30 months and has been doing so for 20+ years; original source code files double every ~22 months
- It is possible to trace the provenance of source code artifacts at this scale in a compact relational model via the notion of isochrone graphs.



Rousseau, Di Cosmo, Zacchioli

Software Provenance Tracking at the Scale of Public Source Code

In *Empirical Software Engineering*, 2020

- 1 Datasets
- 2 Accessing source code artifacts
- 3 Software provenance and evolution
- 4 Software forks
- 5 Diversity, equity, and inclusion



Idea

- Forks can be detected via either platform metadata (e.g., GitHub keeping track of who clicked "fork" on what repo; the most common approach), or via shared version control system history.
- Thanks to deduplication and platform agnosticity, Software Heritage provide a privileged observation point on the global fork ecosystem in public code.

Research questions

- What is the right definition of "being a fork"? (methodology)
- How many forks could we miss by looking only at platform metadata?
- How many "cross-platform" forks (e.g., GitHub → GitLab) exist in the wild?

Findings

- Forks classification: based on platform metadata (“type 1” forks), sharing at least one commit (“type 2”), sharing a common root directory at some point in VCS history (“type 3”).
- Up to 16% forks could be overlooked by considering only GitHub type 1 forks (a potentially significant threat to validity!).
- Relevant independent development activity can happen on GitLab.com for projects initially just mirrored from GitHub.



Pietri, Rousseau, Zacchioli.

Forking Without Clicking: on How to Identify Software Repository Forks.


MSR 2020



Bhattacharjee et al.

An exploratory study to find motives behind cross-platform forks from Software Heritage dataset.

MSR 2020

- 
- 1 Datasets
 - 2 Accessing source code artifacts
 - 3 Software provenance and evolution
 - 4 Software forks
 - 5 Diversity, equity, and inclusion

Idea

Archived commit metadata contains public information that can be mined to study long-term trends of diversity, equity, and inclusion (DEI) traits of the global population of public code contributors.

Key findings on the gender gap

- Male authors contributed 92% of public code commits up to 2019.
- The ratio of female authors (and their contributions) has grown stably for 15 years reaching for the first time 10% of yearly contributions in 2019.
- The COVID-19 pandemic has reversed the trend.

Key findings on the geographic gap

- The early decades of public code were dominated by contributions from North America, followed by a period of alternating dominance between North America and Europe.
- Since then geographic diversity has increased constantly, with raising importance of contributions from Central and South America.
- The trend of increased female contributions is almost worldwide, with the notable exception of specific regions of Asia where it is either slower or flat.

References

- Zacchiroli. *Gender differences in public code contributions: a 50-year perspective*. IEEE Software, 2021
- Rossi and Zacchiroli. *Worldwide gender differences in public code contributions*. ICSE SEIS, 2022
- Rossi and Zacchiroli. *Geographic diversity in public code contributions*. MSR 2022