Software Heritage and Boa

Stefano Zacchiroli

Université de Paris & Inria — zack@upsilon.cc, @zacchiro

Envisioning Boa 2.0 workshop
online conference

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Universal archive

preserve all software source code

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all software source code

Universal archive

preserve all software source code

Research infrastructure

enable analysis of all software source code

An international, non profit initiative built for the long term

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

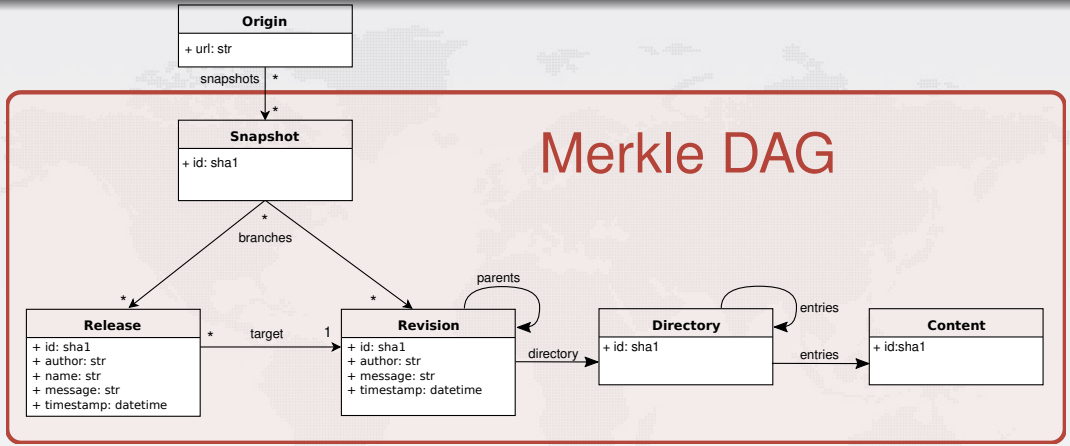
Platinum sponsors

Gold sponsors

Silver sponsors

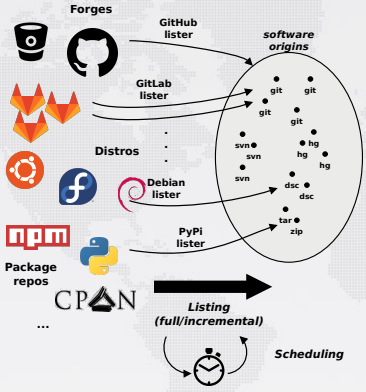
Bronze sponsors

Data model

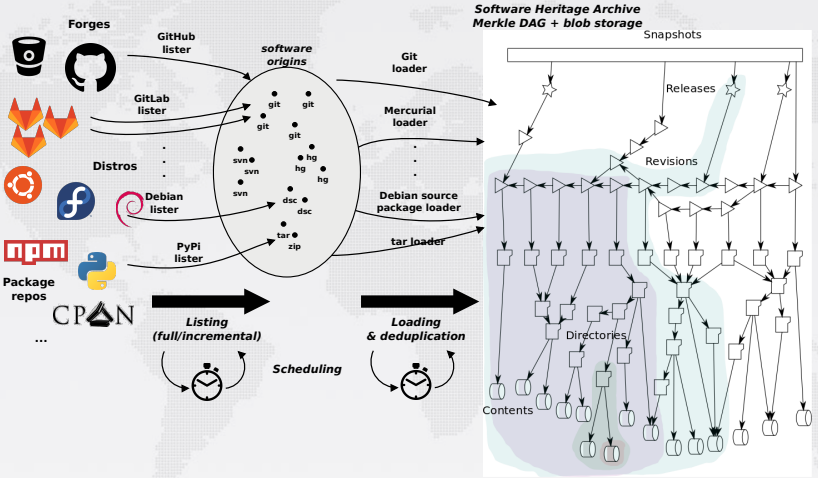


A global graph linking together fully deduplicated source code artifact (files, commits, directories, releases, etc.) to the places that distribute them (e.g., Git repositories), providing a unified view on the entire *Software Commons*.

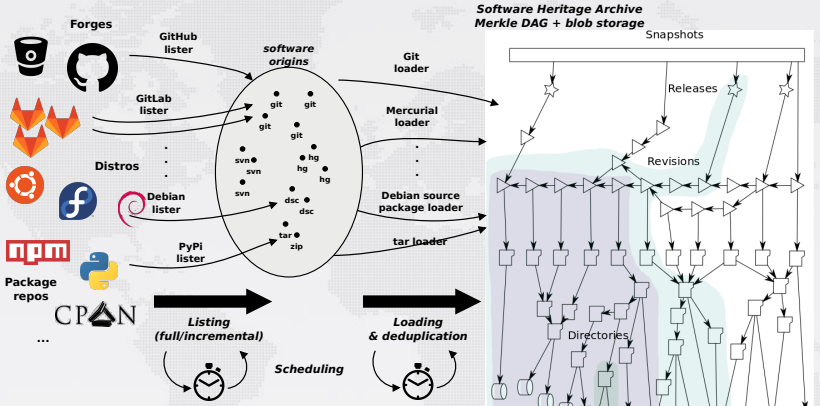
Automation, and storage



Automation, and storage



Automation, and storage



Full development history permanently archived in a uniform data model.

2021-01-archive-growth.png

2021-01-archive-growth.png

Selected research highlights

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code
Communication of the ACM, October 2018

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

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

Antoine Pietri, Guillaume Rousseau, Stefano Zacchiroli
Forking Without Clicking: on How to Identify Software Repository Forks
MSR 2020: 17th Intl. Conf. on Mining Software Repositories. IEEE

Roberto Di Cosmo, Guillaume Rousseau, Stefano Zacchiroli
Software Provenance Tracking at the Scale of Public Source Code
Empirical Software Engineering, 2020

Working together

Analyses of the Software Heritage archive are still performed in ad-hoc ways that require quite some knowledge of the archive data model and caveats.

A Boa-like runtime for Software Heritage

- a filtering language to determine which artifacts and parts of the archive to analyze
- an implementation language to describe experiments on artifacts (general purpose with a dedicated library, DSL, or otherwise)
- a runtime for executing experiments, exploiting DAG node sharing to avoid redoing unneeded expensive computations
- infrastructure connectors that allow to run experiments (locally/cluster/cloud)

Sharing archive access

- we are archiving everything *anyway*, for digital preservation purposes
- an archive mirror could be leveraged by Boa as corpus for researchers

Wrapping up

- Software Heritage archives source code artifacts for posterity and it is the most comprehensive publicly accessible archive of its kind
- Boa harvests source code artifacts and provides an environment that enables researchers to run experiments on them
- Software Heritage provides valuable research datasets, but its researcher-facing part is not as developed as Boa's
- there is mutual benefit in sharing access to source code artifacts and porting Boa's runtime to the Software Heritage archive and data model

Learn more

www.softwareheritage.org

Contacts

Stefano Zacchiroli / zack@upsilon.cc / @zacchiro / @zacchiro@mastodon.xyz