SOFTWARE HERITAGE AND CRAN

CONNECT THE GREAT LIBRARY OF SOURCE CODE WITH
THE COMPREHENSIVE R ARCHIVE NETWORK

Dirk Eddelbuettel¹,⁵,⁶ Stefano Zacchiroli²,³,⁴,⁵

useR! 2019 Toulouse, France 11 July 2019

¹University of Illinois ²Université Paris Diderot ³Inria ⁴Software Heritage ⁵Debian Project ⁶R Project
**Disclaimer**

**Quick Clarifications**

- I (Dirk) am not affiliated with Software Heritage
- So whatever I say do not blame Software Heritage for it
- I have known about Software Heritage for years via Debian
- I meant to get involved for years too and ... ... am now informally helping within Google Summer of Code
- Big thanks to Software Heritage for what they do
- And for access to their collection of talks to bootstrap this one
- Most of what follows ended up coming from https://www.softwareheritage.org/
SOFTWARE HERITAGE: MOTIVATION
Software [is our] Heritage

Source: https://www.softwareheritage.org/
Motivation

Software is fragile

unlike words carved in stone it can be deleted or get corrupted

Source: [https://www.softwareheritage.org/](https://www.softwareheritage.org/)
Motivation

Software is the key to access information and a fundamental part of human heritage.

Source: https://www.softwareheritage.org/
Software Heritage
preserves software source code for present and future generations

Source: https://www.softwareheritage.org/
We are building the universal software archive

Collect
Preserve
Share

We collect and preserve software in source code form, because software embodies our technical and scientific knowledge and humanity cannot afford the risk of losing it.

Software is a precious part of our cultural heritage. We curate and make accessible all the software we collect, because only by sharing it we can guarantee its preservation in the very long term.

Source: https://www.softwareheritage.org/
SOFTWARE HERITAGE: WHO AND HOW
Executives

**Roberto Di Cosmo (Founder, CEO)**

After teaching for almost a decade at Ecole Normale Supérieure in Paris, Roberto Di Cosmo became full professor in Computer Science at University Paris Diderot. He is currently on leave at Inria to lead the Software Heritage project.

His research interests span a wide spectrum from foundational aspects of logical systems to functional programming, parallel and distributed programming. He created and directed the European research project Mancoosi to improve the quality of large collections of software quality, and is investigating now the scientific problems posed by the general adoption of Free Software, with a particular focus on static analysis of large software collections.

A long term Free Software advocate, contributing to its adoption since 1998, he has created the [Free Software thematic group of Systematic](https://www.softwareheritage.org/people) in October 2007, which has helped fund over 40 research and development projects, and he is now director of IRILL, a research structure dedicated to Free and Open Source Software quality.

- Email: roberto@softwareheritage.org
- Twitter: @rdicosmo

**Stefano Zacchiroli (Founder, CTO)**

Stefano Zacchiroli holds a PhD in Computer Science from the University of Bologna, Italy, and is Associate Professor of Computer Science at University Paris Diderot, France. He is currently on leave at INRIA and a research fellow at IRILL, a research institute dedicated to the study of Free/Open Source Software (FOSS).

His research interests span formal methods, their applications to improve software quality and packages upgrades, as well as Free Software evolution.

He has been an official member of the Debian Project since 2001, where he worked on many tasks, from package maintenance to distribution-wide Quality Assurance. He has been elected to serve as Debian Project Leader (DPL) for 3 terms in a row, during the period 2010-2013. He is a Board Director of the Open Source Initiative (OSI) and a recipient of the 2015 O'Reilly Open Source Award.

- Email: zack@softwareheritage.org
- Twitter: @zacchiro
Software source code represents unique knowledge of humanity's recent history. It is therefore crucial to work together collectively so that the knowledge embedded in software source code is properly preserved, valued and shared with all. This lies at the core of UNESCO's cooperation with Inria to support the creation of Software Heritage, the global archive of software source code.

—Moez Chakchouk, Assistant Director-General for Communication and Information, UNESCO

Source: https://www.softwareheritage.org/support/testimonials/
https://www.softwareheritage.org/2019/06/24/unesco/
Data Flow into DAG

Source: Software Heritage, with permission
Merkle Trees

• Combination of
  • tree
  • hash function
• Classical cryptographic construction
  • fast, parallel signature of large data structures
  • widely used (e.g., Git, blockchains, IPFS, …)
  • built-in deduplication

Source: Software Heritage, with permission
Overview

The long term goal of the Software Heritage initiative is to collect all publicly available software in source code form together with its development history, replicate it massively to ensure its preservation, and share it with everyone who needs it. The Software Heritage archive is growing over time as we crawl new source code from software projects and development forges. We will incrementally release archive search and browse functionalities — as of now you can check whether source code you care about is already present in the archive or not.

Content

A significant amount of source code has already been ingested in the Software Heritage archive. It currently includes:

![Logos of various sources](image)

Size

As of today the archive already contains and keeps safe for you the following amount of objects:

<table>
<thead>
<tr>
<th>Source files</th>
<th>Directories</th>
<th>Commits</th>
<th>Authors</th>
<th>Projects</th>
<th>Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,028,249,303</td>
<td>5,241,278,479</td>
<td>1,339,869,315</td>
<td>24,883,628</td>
<td>89,387,814</td>
<td>11,442,050</td>
</tr>
</tbody>
</table>

Source: [https://archive.softwareheritage.org/](https://archive.softwareheritage.org/) (as of June 29, 2019)
Archive Coverage: Scale and Growth

Scale

- 200 TB (compressed) blobs
- 6 TB database (as a graph: 10 B nodes + 100 B edges)
- The *richest* public source code archive, ... and growing daily!

Source: Software Heritage, with permission
Stefano Zacchiroli @zacchiro

Observing the memory/CPU/disk usage patterns of #GNU #sort sorting a ~7 TB file is hypnotizing.

6:04 AM - 5 May 2019

2 Retweets 6 Likes

in case you were wondering:
$ time zcat dir_to_file.gz | tr ' ' "n" | sort -u --parallel 32 -S100G -T tmp | pigz -p 32 -c > dir_to_file.nodes.gz
real 4424m12.496s
user 17201m13.913s
sys 841m59.555s

putting the big back into big data
SOFTWARE HERITAGE: ADDING R
Significant Parts Already Available via GitHub

With thanks to Gábor Csárdi for the CRAN mirror at GitHub!
This Year

One of the Google Summer of Code Projects

- Archit Agrawal works on “Increasing Archive Coverage”
- This includes R / CRAN ingestion (among others)
- Use CRAN metadata for ‘state’ and change from last saved state
- Download, parse, index, ... tarballs off CRAN
- Seed with CRAN archive section, then update live as CRAN grows
- Will start “any day now”
GET INVOLVED
Main Website

- https://www.softwareheritage.org/

Archive

- https://archive.softwareheritage.org/browse/search/
- https://archive.softwareheritage.org/api/

Development

- https://forge.softwareheritage.org/
- https://docs.softwareheritage.org/devel/
THANKS
Thanks

- To the *useR! 2019* organizers for giving us a last-minute slot
- To everybody
  - writing,
  - releasing, and
  - maintaining software worth archiving
- To everybody hosting: GitHub, CRAN, ...