

SOFTWARE HERITAGE AND CRAN

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

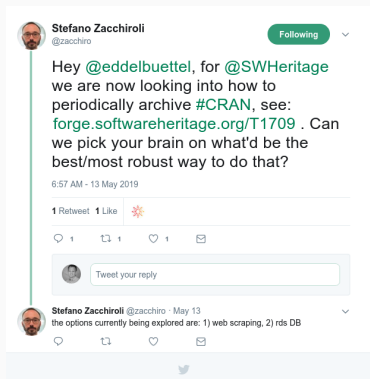
Dirk Eddebuettel^{1,5,6} Stefano Zacchiroli^{2,3,4,5}

useR! 2019 Toulouse, France 11 July 2019

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

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

```
4015 */
4016 static int do_sched_cfs_period_timer(struct cfs_bandwidth *cfs_b, int overrun)
4017 {
4018     u64 runtime, runtime_expires;
4019     int throttled;
4020
4021     /* no need to continue the timer with no bandwidth remaining */
4022     if (cfs_b->quota == RUNTIME_INF)
4023         goto out_deactivate;
4024
4025     throttled = !list_empty(&cfs_b->throttled_cfs_rq);
4026     cfs_b->nr_periods += overrun;
4027
4028     /*
4029      * idle depends on !throttled (for the case of a large backlog, see 2).
4030      * we're going inactive then everything else can be scheduled.
4031      */
4032     if (cfs_b->idle && !throttled)
4033         goto out_deactivate;
4034
4035     __refill_cfs_bandwidth_runtime(cfs_b);
4036
4037     if (!throttled) {
4038         /* mark as potentially idle for the upcoming period */
4039         cfs_b->idle = 1;
4040         return 0;
4041     }
4042
4043     /* account preceding periods in which throttling occurred */
4044     cfs_b->nr_throttled += overrun;
4045
4046     runtime_expires = cfs_b->runtime_expires;
4047
4048     /*
4049      * This check is repeated as we are building onto the run bandwidth again.
4050      * we unthrottle. This can potentially race with unthrottling logic.
4051      * trying to acquire new bandwidth from the cfs_rq will fail as 2000.
4052      */
4053 }
```

Software [is our] Heritage

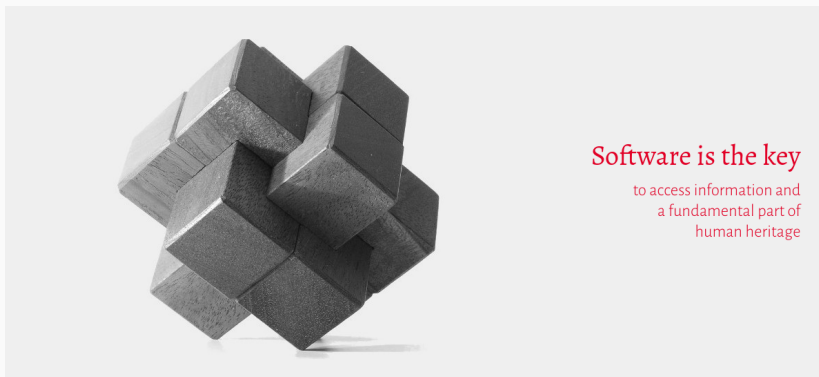
Source: <https://www.softwareheritage.org/>

Software is fragile

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



Source: <https://www.softwareheritage.org/>



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.

[Discover our mission](#)

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 reseach 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 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](https://twitter.com/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](https://twitter.com/zacchiro)

UNESCO



United Nations
Educational, Scientific and
Cultural Organization

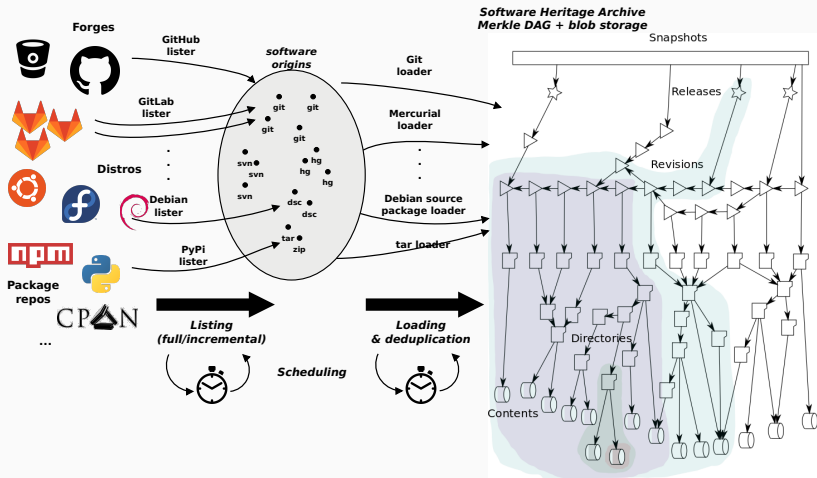


Software source code represents unique knowledge of humanity's recent history. It is therefore crucial to work together collectively so that the knowledge embedded in software source code is properly preserved, valued and shared with all. This lies at the core of UNESCO's cooperation with Inria to support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

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

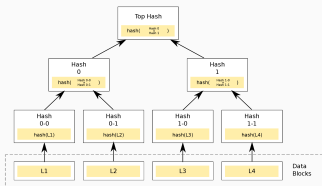
Source: <https://www.softwareheritage.org/support/testimonials/>
<https://www.softwareheritage.org/2019/06/24/unesco/>

DATA FLOW INTO DAG



Source: Software Heritage, with permission

Merkle Trees



Source: Software Heritage, with permission

- 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

ARCHIVE COVERAGE

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:



GitLab



Google code



Size

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

Source files	Directories	Commits	Authors	Projects	Releases
6,028,249,303	5,241,278,479	1,339,869,315	24,883,628	89,387,814	11,442,050

Source: <https://archive.softwareheritage.org/> (as of June 29, 2019)

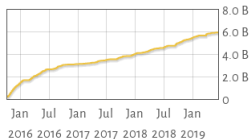
ARCHIVE COVERAGE: SCALE AND GROWTH

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!

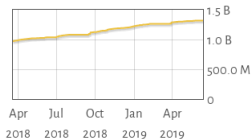
Source files

6,006,503,960



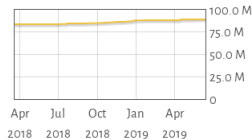
Commits

1,326,776,432



Projects

89,301,694



Source: Software Heritage, with permission



Stefano Zacchiroli

@zacchiro

Following

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



2



2



6



Tweet your reply



Stefano Zacchiroli @zacchiro · May 8

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



1



2



putting the *big*
back into big data

SOFTWARE HERITAGE: ADDING R

SIGNIFICANT PARTS ALREADY AVAILABLE VIA GITHUB

Revision history - HEAD - origin: <https://github.com/cran/nnet> - Software Heritage archive - Google Chrome

<https://archive.softwareheritage.org/browse/origin/https://github.com/cran/nnet/0a/>

Home Archive Development Documentation [Donate](#)

Software Heritage

Browse archived revisions history for origin <https://github.com/cran/nnet>

Visits Snapshot date: 02 May 2019, 14:11 UTC Branches (36) Releases (0)

Branch: HEAD

Actions

sort by: revision date DFS DFS post-ordering BFS

Revision	Author	Date	Message	Commit Date
a6a4b10	Brian Ripley	02 February 2016, 14:52 UTC	version 7.3-12	02 February 2016, 14:52 UTC
ce4cb17	Brian Ripley	30 August 2015, 08:59 UTC	version 7.3-11	30 August 2015, 08:59 UTC
d3cd10a	Brian Ripley	29 June 2015, 17:15 UTC	version 7.3-10	29 June 2015, 17:15 UTC
2472025	Brian Ripley	11 February 2015, 09:30 UTC	version 7.3-9	11 February 2015, 09:30 UTC
1278b12	Brian Ripley	28 March 2014, 09:12 UTC	version 7.3-8	28 March 2014, 09:12 UTC
e420020	Brian Ripley	01 July 2013, 13:42 UTC	version 7.3-7	01 July 2013, 13:42 UTC
2b21afb	Brian Ripley	20 March 2013, 11:20 UTC	version 7.3-6	20 March 2013, 11:20 UTC
a5472b6	Brian Ripley	11 October 2012, 13:01 UTC	version 7.3-5	11 October 2012, 13:01 UTC
b965de8	Brian Ripley	27 June 2012, 09:58 UTC	version 7.3-4	27 June 2012, 09:58 UTC
86623c9	Brian Ripley	28 October 2009, 07:17 UTC	version 7.3-1	28 October 2009, 07:17 UTC
23b68c7	Brian Ripley	07 May 2009, 11:20 UTC	version 7.3-0	07 May 2009, 11:20 UTC

Newer Older

Software Heritage — Copyright (C) 2015–2019, The Software Heritage developers. License: [GNU AGPLv3+](#).
The source code of Software Heritage itself is available on our [development forge](#).
The source code files archived by Software Heritage are available under their own copyright and licenses.
[Terms of use](#), [Archive access](#), [API](#), [Contact](#), [JavaScript license information](#)

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

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, ...