Software Heritage

A revolutionary infrastructure for Open Science and Open Source

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

March 26, 2025

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

Open Science

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *35+ years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *25+ years* of Free and Open Source Software
- *15+ years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*, European Union

Software *Source Code* is Precious Knowledge

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

"Programs must be written for people to read, and only incidentally for machines to execute."

Apollo 11 source code (excerpt)

```
P63SPOT3        CA      BIT6            # IS THE LR ANTENNA IN POSITION 1 YET
                EXTEND
                RAND    CHAN33
                EXTEND
                BZF     P63SPOT4        # BRANCH IF ANTENNA ALREADY IN POSITION 1

                CAF     CODE500         # ASTRONAUT:    PLEASE CRANK THE
                TC      BANKCALL        #               SILLY THING AROUND
                CADR    GOPERF1
                TCF     GOTOPOOH        # TERMINATE
                TCF     P63SPOT3        # PROCEED       SEE IF HE'S LYING

P63SPOT4        TC      BANKCALL        # ENTER         INITIALIZE LANDING RADAR
                CADR    SETPOS1

                TC      POSTJUMP        # OFF TO SEE THE WIZARD ...
                CADR    BURNBABY
```

Covid Sim (excerpt)

```
/**
 * @brief The basic unit of the simulation and is associated to a geographical location.
 *
 * Interventions (e.g., school closures) are tracked at this level. It contains a list of its
 * members (people), places (schools, universities, workplaces etc.), road networks, links to
 * airports etc.
 */
struct Microcell
{
        /* Note use of short int here limits max run time to USHRT_MAX*ModelTimeStep - e.g. 65536*0.25=16384 days=44 yrs.
           Global search and replace of 'unsigned short int' with 'int' would remove this limit, but use more memory.
        */

        int n; // Number of people in microcell
        int adunit; // admin unit microcell belongs to
        int* members; // array of members/hosts of microcell

        int* places[MAX_NUM_PLACE_TYPES]; // list of places (of various place types) within microcell
        unsigned short int NumPlacesByType[MAX_NUM_PLACE_TYPES]; // number of places (of various place types) within mircocell
        unsigned short int keyworkerproph, move_trig, place_trig, socdist_trig, keyworkerproph_trig;
        unsigned short int move_start_time, move_end_time;
        unsigned short int place_end_time, socdist_end_time, keyworkerproph_end_time;
        TreatStat moverest, treat, vacc, socdist, placeclose;
        unsigned short int treat_trig, vacc_trig;
        unsigned short int treat_start_time, treat_end_time;
        unsigned short int vacc_start_time;
        IndexList* Airportlist;
};
```

Len Shustek, Computer History Museum 2006

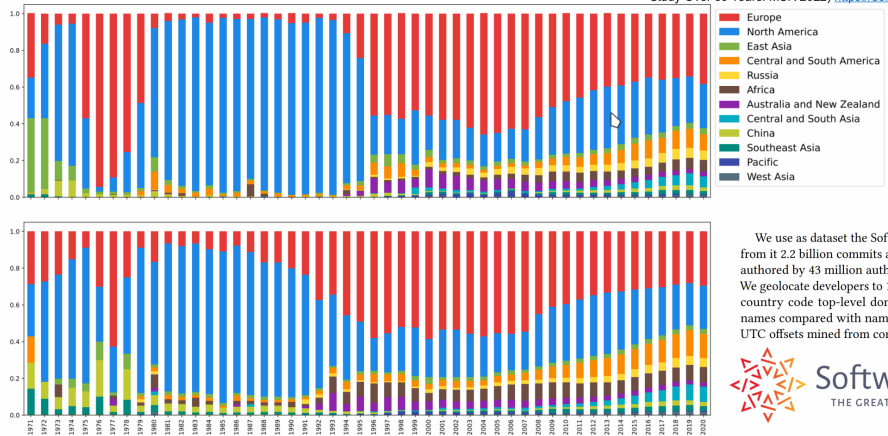
"Source code provides a view into the mind of the designer."

(Open) Source Code comes from all over the world

MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

Davide Rossi and Stefano Zacchiroli

Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022) https://doi.org/10.1145/3524842.3528471

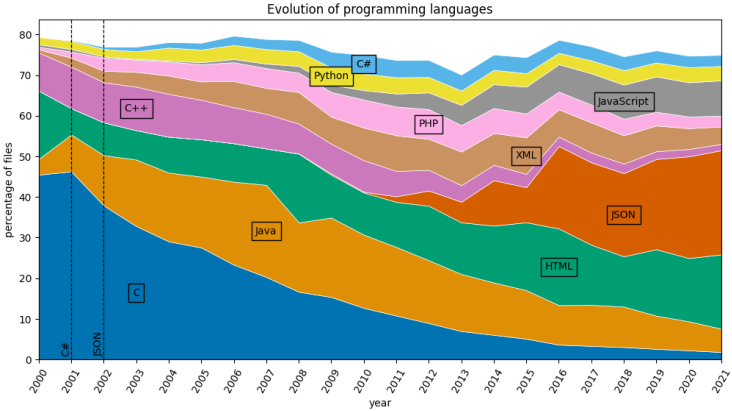


We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

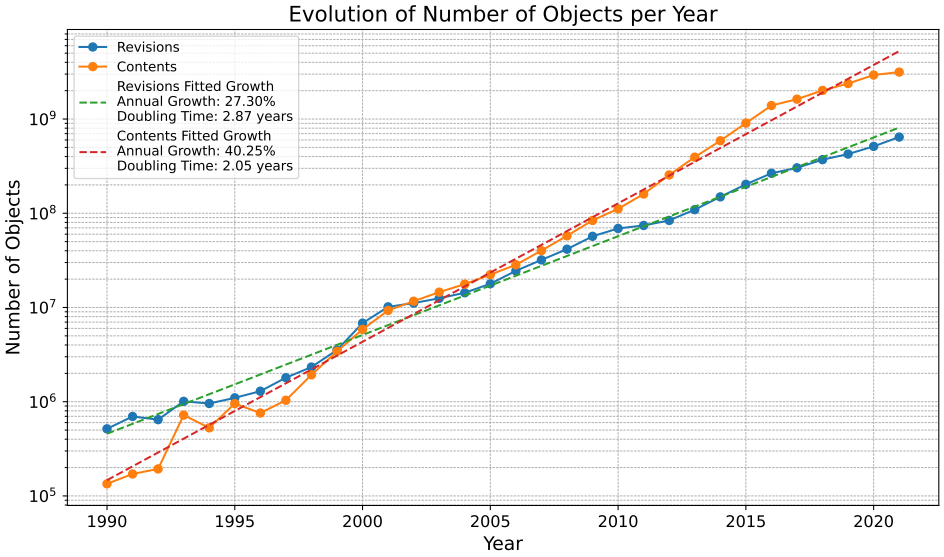
(Open) Source Code is written in many languages



Evolution of the activity for programming, markup, and data languages from 2000 to 2021.

A. Desmazières, R. Di Cosmo, V. Lorentz
50 years of programming language evolution through the Software Heritage Looking Glass
MSR 2025. To appear.

(Open) Source Code grows at an exponential rate

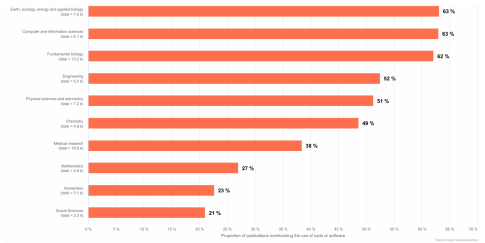


(Open Source) Software is *precious technical and scientific knowledge*

Yuval Noah Harari (on COVID 19)

"*The real antidote [to epidemic] is* scientific knowledge *and* global cooperation."

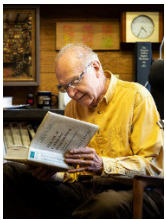
Software powers modern research



20%+ articles use software, all disciplines

2023 French Open Science Monitor

We can still talk to the early inventors



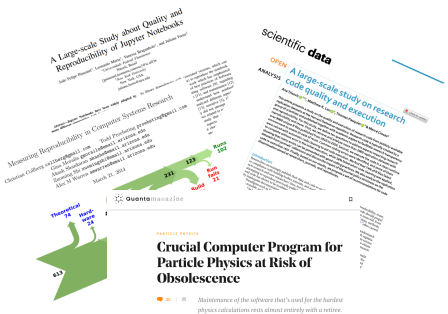
"Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to."

Donald E. Knuth
Len Shustek
CACM, January 2021

We need a *dedicated infrastructure* to preserve and share *all* this knowledge!

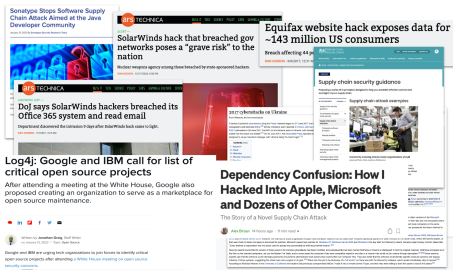
How are we managing our (open source) software ?

Reproducibility, maintenance in Academia



(articles: here, here, here and here)

Security, integrity, traceability in Industry



Can they track the software that they

- ship, use, acquire
- has that bug or vulnerability

awareness is raising at the level of public policy

Outline

Open Science

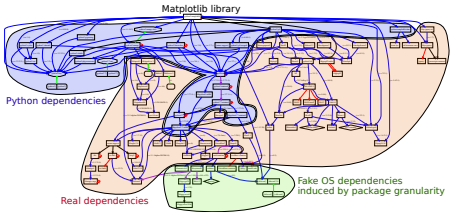
Source code is *special* (software is *not* data)

Software *evolves* over time

- projects may last decades
- the *development history* is key to its *understanding*

Complexity

- *millions* of lines of code
- large *web of dependencies*
  - easy to break, difficult to maintain
  - *research software* a thin top layer
- sophisticated *developer communities*



Precious, endangered *executable* and *human readable* knowledge

key people passing away, platforms (GoogleCode, Gitorious, etc.) closing down …

no organised effort to catalog and archive it

A plurality of needs (in increasing order of difficulty)

Archive

Research software artifacts must be properly archived

make sure we can *retrieve* them (*reproducibility*)

Reference

Research software artifacts must be properly referenced

make sure we can *identify* them (*reproducibility*)

Describe

Research software artifacts must be properly described

make it easy to *discover* and *reuse* them (*visibility*)

Cite/Credit

Research software artifacts must be properly cited *(not the same as referenced!)*

to give *credit* to authors (*evaluation*!)

Outline

Open Science

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 and share all software source code

Research infrastructure

enable analysis of all software source code

A *universal* archive as a shared infrastructure

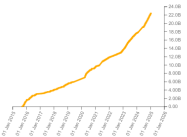
One infrastructure
open and shared

Inria

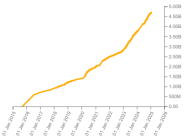
unesco

The largest archive ever built

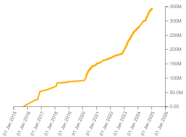
Source files
22,548,654,226



Commits
4,743,656,085



Projects
344,672,354



Directories
17,848,569,305

Authors
86,143,084

Releases
102,056,508

Bitbucket 2,578,475 origins	56,975 origins	git 33,350 origins
27,377 origins	debian 141,834 origins	88,561 origins
GitHub 249,868,189 origins	gitiles 24,378 origins	GitLab 5,736,223 origins
git 3,791 origins	Gogs 394 origins	1,887,337 origins
Guix 68,391 origins	GNU 354 origins	heptapod 1,340 origins
launchpad 654,755 origins	Maven 312,179 origins	NixOS 48,905 origins
npm 4,003,267 origins	5,438 origins	Packagist 376,882 origins

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

Donors, members, sponsors

Diamond sponsors

Platinum sponsors

Gold sponsors

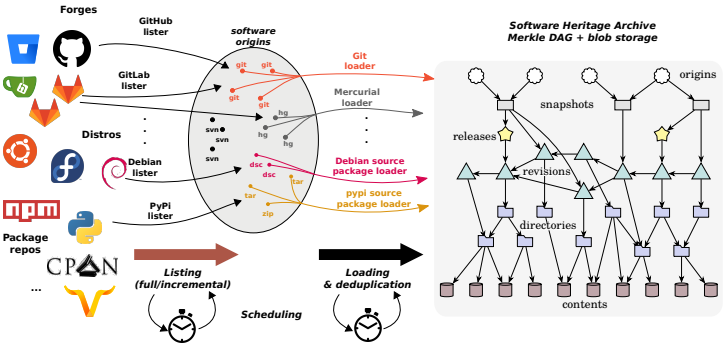
Silver sponsors

Bronze sponsors

And many more ...

www.softwareheritage.org/support/testimonials

The archive, under the hood



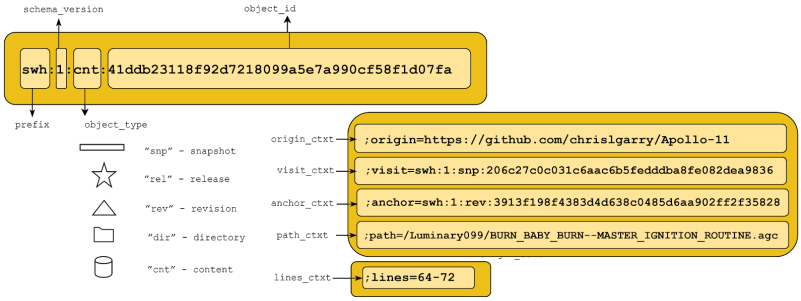
Global development history permanently archived in a uniform data model

- over 22 billion unique source files from over 340 million software projects
- ~2PB (compressed) blobs, ~50 B nodes, ~800 B edges

The Software Hash identifier (SWHID)

Software Heritage Identifiers (SWHID)

see swhid.org



50+B
intrinsic,
decentralised,
cryptographic

Full fledged *source code references* for traceability, integrity and reproducibility

- Linux Foundation SPDX 2.2
- IANA-registered "swh:"
- WikiData property P6138

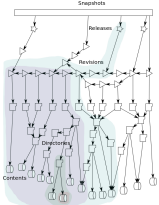
Examples: Apollo 11 AGC excerpt, Quake III rsqrt
Guidelines available, see the HOWTO

Breaking news: standardisation, see swhid.org - DIS 18670

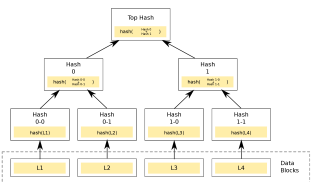
A revolutionary infrastructure

Modern "Library of Alexandria", *international, non profit, long term* initiative
addressing the needs of *industry, research, culture and society as a whole*

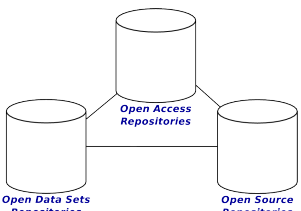
Software Graph



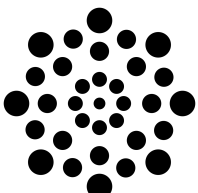
Software Blockchain



Open Science pillar



Big Code



One infrastructure, *shared*: more efficient, less waste …

… supporting a broad range of applications

Outline

Open Science

A walkthrough

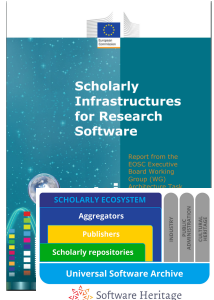
- Browse (e.g. Apollo 11 [excerpt], your work may be already there !)
- Trigger archival, use the updateswh browser extension, configure the webhooks
- Get and use SWHIDs (full specification available online)
- Cite software with biblatex-software package from CTAN
  - Overleaf ACMART template available
- Example in journals: article from IPOL
- Example with Parmap: devel on Github, archive in SWH, curated deposit in HAL
- Extracting all the software products for Inria, for CNRS, for CNES, for LIRMM or for Rémi Gribonval using HalTools
- Curated deposit in SWH via HAL, see for example: LinBox, SLALOM, Givaro, NS2DDV, SumGra, Coq proof, ...
- Example use in research articles:
  - compare Fig. 1 and conclusions in the 2012 version and the updated version
  - SWHID in a replication experiment

Outline

Open Science

A few adoption indicators

Policy



- Recommendations in ANR 2023 guidelines (p. 17)
- HAL+SWH in the Open Science software booklet

Projects

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

The CodeMeta Project

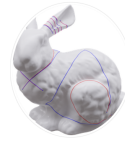
eosc | FAIR-IMPACT
Expanding FAIR solutions across EOSC

Users and collaborations

What are they “referencing”?

source	n	percentage
Not available	2868	46.22
GitHub	1151	18.55
software heritage	387	6.24
zenodo	142	2.29
r package	70	1.13
cran	56	0.90
r package version	54	0.87
gitlab	35	0.56

Graphics Replicability Stamp Initiative



b/Surf: Interactive Bézier Splines on Surface Meshes

Claudio Mancinelli, Giacomo Nazzaro, Fabio Pellacini, Enrico Puppo
IEEE Transactions on Visualization and Computer Graphics (TVCG)

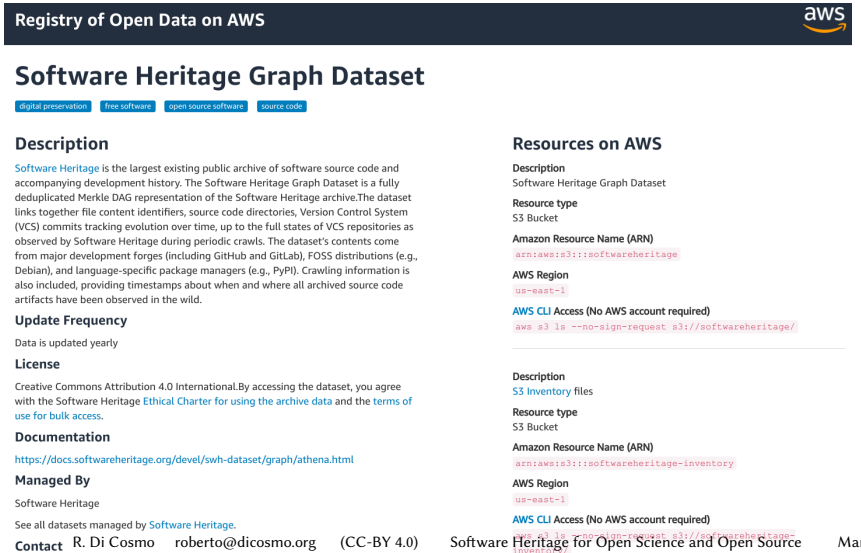
doi | Repository

Outline

Open Science

The full graph in the AWS Open Data collection

https://registry.opendata.aws/software-heritage/



Registry of Open Data on AWS

aws

Software Heritage Graph Dataset

digital preservation | free software | open source software | source code

Description

Software Heritage is the largest existing public archive of software source code and accompanying development history. The Software Heritage Graph Dataset is a fully deduplicated Merkle DAG representation of the Software Heritage archive.The dataset links together file content identifiers, source code directories, Version Control System (VCS) commits tracking evolution over time, up to the full states of VCS repositories as observed by Software Heritage during periodic crawls. The dataset's contents come from major development forges (including GitHub and GitLab), FOSS distributions (e.g., Debian), and language-specific package managers (e.g., PyPI). Crawling information is also included, providing timestamps about when and where all archived source code artifacts have been observed in the wild.

Update Frequency

Data is updated yearly

License

Creative Commons Attribution 4.0 International.By accessing the dataset, you agree with the Software Heritage Ethical Charter for using the archive data and the terms of use for bulk access.

Documentation

https://docs.softwareheritage.org/devel/swh-dataset/graph/athena.html

Managed By

Software Heritage

See all datasets managed by Software Heritage.

Resources on AWS

Description
Software Heritage Graph Dataset

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::softwareheritage`

AWS Region
`us-east-1`

AWS CLI Access (No AWS account required)
`aws s3 ls --no-sign-request s3://softwareheritage/`

Description
S3 Inventory files

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::softwareheritage-inventory`

AWS Region
`us-east-1`

AWS CLI Access (No AWS account required)
`aws s3 ls --no-sign-request s3://softwareheritage-inventory/`

Contact

A peek at the dataset

Accessing graph leaves (a.k.a. contents)

```
$ aws s3 ls --no-sign-request s3://softwareheritage/
       PRE content/
       PRE graph/
```

File contents can be accessed using their SHA1 checksum

```
$ aws s3 cp --no-sign-request \
  s3://softwareheritage/content/8624bcdae55baeef00cd11d5dfcfa60f68710a02 .
```

Notice that file contents are compressed:

```
$ zcat 8624bcdae55baeef00cd11d5dfcfa60f68710a02 | head
  GNU GENERAL PUBLIC LICENSE
     Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.
```

A peek at the dataset, cont'd

Annual dumps of (inner nodes of) the full graph

```
$ aws s3 ls --no-sign-request s3://softwareheritage/graph/
    PRE 2018-09-25/
    PRE 2019-01-28-popular-3k-python/
    PRE 2019-01-28-popular-4k/
    PRE 2020-05-20/
    PRE 2020-12-15/
    PRE 2021-03-23-cpython-3-5/
    PRE 2021-03-23-popular-3k-python/
    PRE 2021-03-23/
    PRE 2022-04-25/
```

How to use and cite

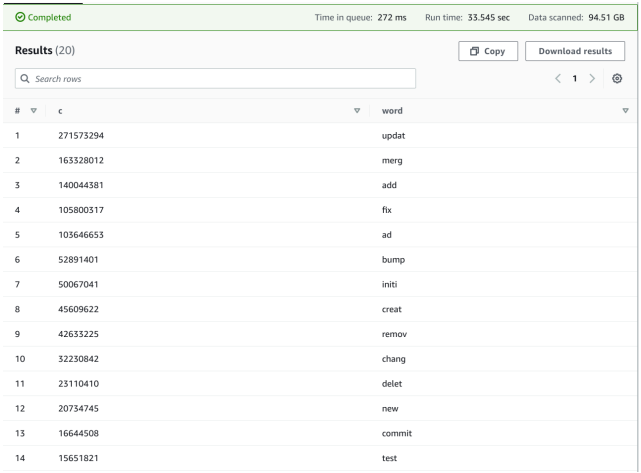
- online full documentation, and read Antoine Pietri's PhD Thesis
- Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli. *The Software Heritage Graph Dataset: Public software development under one roof.* MSR 2019. (bibtex)

Example: most popular commit verbs (stemmed)

Query using Amazon Athena

```
SELECT COUNT(*) AS C, word FROM (
   SELECT word_stem(lower(split_part(
    trim(from_utf8(message)),' ', 1)))
   AS word FROM revision
   WHERE length(message) < 1000000)
WHERE word != ''
GROUP BY word
ORDER BY C
DESC LIMIT 20;
```

Results



Completed — Time in queue: 272 ms — Run time: 33.545 sec — Data scanned: 94.51 GB

Results (20) — Copy — Download results

Search rows — 1

#	c	word
1	271573294	updat
2	163328012	merg
3	140044381	add
4	105800317	fix
5	103646653	ad
6	52891401	bump
7	50067041	initi
8	45609622	creat
9	42633225	remov
10	32230842	chang
11	23110410	delet
12	20734745	new
13	16644508	commit
14	15651821	test

Outline

Open Science

Going beyond SQL

State-of-the-art graph compression from social networks

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

Results

Full graph structure (35 B nodes, 500 B edges) in 300 GiB RAM

- traversal time is tens of ns per edge
- bidirectional traversals implemented
- beware: metadata access is still *off RAM*

~~Java~~ Rust and gRPC APIs available …

docs.softwareheritage.org/devel/swh-graph/grpc-api.html

Find all origins containing a given content

```
grpc_cli call localhost:50091 swh.graph.TraversalService.Traverse "\
src: 'swh:1:cnt:8722d84d658e5e11519b807abb5c05bfbfc531f0', direction: BACKWARD, \
mask: {paths: ['swhid','ori.url']}, return_nodes: {types: 'ori'}"
```

Gives a list of origins including "https://github.com/rdicosmo/parmap", encoded as "swh:1:ori:8903a90cff8f07159be7aed69f19d66d33db3f86" (beware: this is not a SWHID!)

Shortest provenance path of a content in a given origin

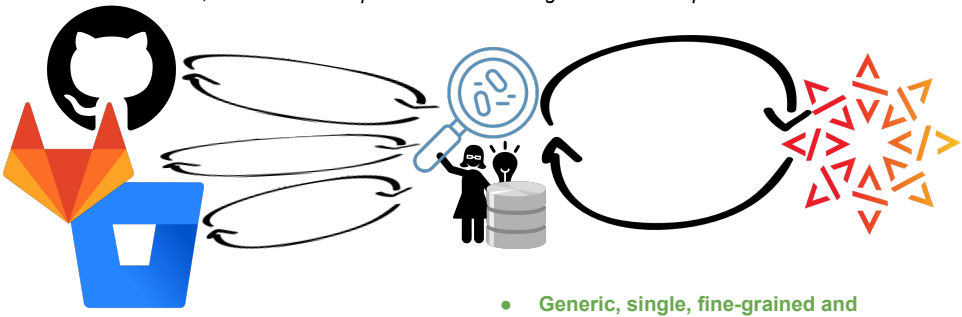
```
grpc_cli call localhost:50091 swh.graph.TraversalService.FindPathBetween "\
src: 'swh:1:ori:8903a90cff8f07159be7aed69f19d66d33db3f86', \
dst: 'swh:1:cnt:8722d84d658e5e11519b807abb5c05bfbfc531f0', \
mask: {paths: ['swhid']}" | egrep 'swhid'
connecting to localhost:50091
  swhid: "swh:1:ori:8903a90cff8f07159be7aed69f19d66d33db3f86"
  swhid: "swh:1:snp:1527a93b039d70f6a781b05d76b77c6209912887"
  swhid: "swh:1:rev:82df563aecf86b9164eee7d10d40f2d8cbd1c78d"
  swhid: "swh:1:dir:484db39bb2825886191837bb0960b7450f9099bb"
  swhid: "swh:1:dir:4d15e44b378fe39dd23817abee756cd47ad14575"
  swhid: "swh:1:cnt:8722d84d658e5e11519b807abb5c05bfbfc531f0"
Rpc succeeded with OK status
```

Outline

Open Science

Mining Android Applications on Software Heritage

RQ: how to build a specific dataset for a given research question?



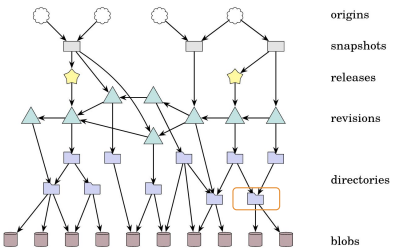
- **Specific and limited API**
- **Hardly reproducible**

- **Generic, single, fine-grained and unlimited API**
- **Growing number of source codes**
- **Easy to update the dataset**

(from the Inria/IRISA DiverSE team)

Using the SWH merkle dag to identify android repositories

Identify android application repositories = Find the AndroidManifest.xml among the sources

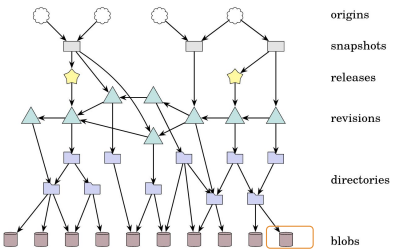


SWH Merkle DAG, Antoine Pietri

1) Iterate over the graph nodes until you find a directory node containing a file named "AndroidManifest.xml".

Using the SWH merkle dag to identify android repositories

Identify android application repositories = Find the AndroidManifest.xml among the sources

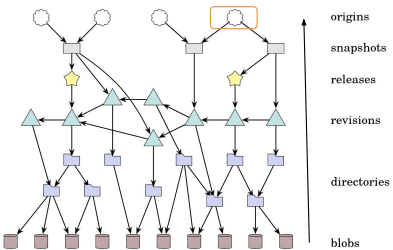


SWH Merkle DAG, Antoine Pietri

2) Extract the SWH identifier of the blob corresponding to the AndroidManifest.xml and download the corresponding file through the SWH Web API

Using the SWH merkle dag to identify android repositories

Identify android application repositories = Find the AndroidManifest.xml among the sources



SWH Merkle DAG, Antoine Pietri

3) Traverse the graph in backward direction to the origin node and get the repository url

Bottomline

Broad variety of sources in *one open dataset*

reduces usual GH bias

Reference simple *standard data format*

VCS and forge details are abstracted away

Simplifies reproducibility packages

no need to create a full copy, *just list the SWHIDs!*

Software Heritage does the heavy lifting for you

no need to scrape/download repositories all over again

Large scale studies: the CERN use case

CERN Accelerating science Sign in Directory

CERN's open source heritage: Building blocks to share

Photo by Jan Huber on Unsplash

The European Organization for Nuclear Research, known as CERN, has long been a driving force in scientific discovery and technological advancement. Beyond its groundbreaking research, CERN has also quietly championed open-source software for decades. But how to measure CERN's impact on the global open-source community?

"Member states will appreciate that CERN is not only carrying out physics research but also contributes back through open source," says Axel Naumann, Chair of CERN's Open Source Program Office (OSPO). "That said, we've been scratching our heads on how to measure the impact. It's a non-trivial task, and while we lack statistically sound information about our user base, we can look at the code that's been produced for insights."

Software Heritage Mission Archive Community Grants Support us

Home / Search / cern

Measuring the reach of CERN's code with Software Heritage

CERN, renowned for its groundbreaking physics research, also has a significant impact on the open-source software world. Software Heritage is working to preserve and share this valuable code, despite the challenges of tracking its decentralized nature.

November 7, 2024

CERN, European Organization for Nuclear Research

And much more

Cybersecurity

SWHSec (CampusCyber project) — main question: how can we leverage Software Heritage as a knowledge base to increase the security of open source software?

AI

- Code Commons: producing research datasets for *ethical* training of LLMs on Software Heritage code

October 19, 2023

Software Heritage Statement on Large Language Models for Code

1. Give back to humanity
2. Precisely identify training inputs (with SWHIDs!)
3. Opt-out

Outline

Open Science

A growing and active community

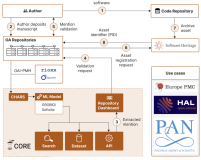
Core Team

All together, 2024 Symposium

Ambassadors

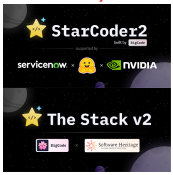
Software Heritage: 2024 progress hilights

January 2024 — SoFAIR Kickoff



Massive identification, archival and reference of software in academic publications

February 2024 — First AI contact



StarCoder2 trained on GitHub subset of Software Heritage

February 2024 — SWHSec Kickoff

SWHSec: 5 years PTCC project
Software supply chain Security

Academic interoperability

eosc
FAIR-IMPACT
FAIRCORE4EOSC

Connect with Zenodo, Episcience, Dagsthul, SwMath, OpenAire, …

Fall 2024 — CodeCommons

Building the reference software knowledge base for AI

… and much more

2024 annual report here →

Adopt and share best practices for ARDC

Archiving and referencing

For all source code used in research (*yes, even small scripts!*)

- archive and reference in Software Heritage (see detailed HOWTO)

Describing and Citing/Crediting

For software one wants to put forward, add these extra steps:

- add codemeta.json with description (see the codemeta generator)
- (french partners) reference in the HAL portal (see online HAL documentation)
- cite software using the biblatex-software package (in CTAN and TeXLive)

We can (and must)

- train students and colleagues
- engage AECs, journals, conferences, learned societies

A rally flag for a grand vision

Bring together academia, industry, governments, communities

"to build a reference, global infrastructure for open and better software"

Software Heritage is the first brick …

- vendor neutral
- open source
- a worldwide initiative
- a long term initiative

… that will enable

- archival, reference, integrity
- qualification, sharing and reuse
- a global software knowledge base
- test and deploy world class tooling

A lot more is needed

Software Heritage can be the *catalyser* of a way bigger undertaking

You can help!

use, disseminate, contribute, build&adapt research tools, …