Securing the (Open Source) Software Supply Chain

challenges and opportunities

Roberto Di Cosmo

Director, Software Heritage
Inria and Université de Paris Cité

February 14th 2023

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30+ years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20+ years* of Free and Open Source Software
- *10+ 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

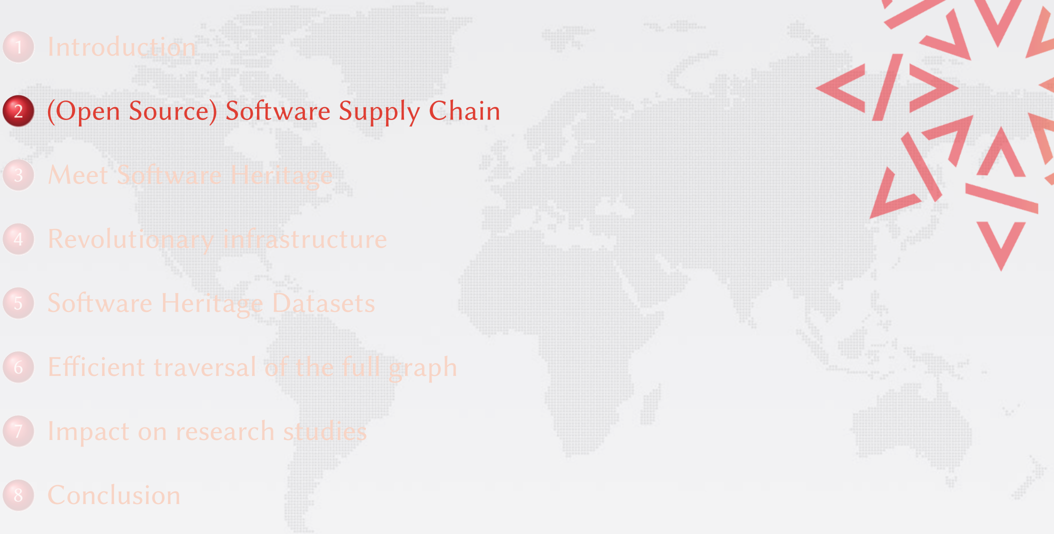
2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

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

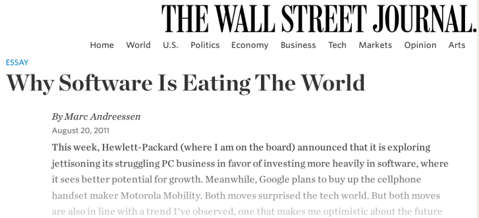
Outline



1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

Open Source is growing…

Software is eating the world



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts

ESSAY

Why Software Is Eating The World

By Marc Andreessen

August 20, 2011

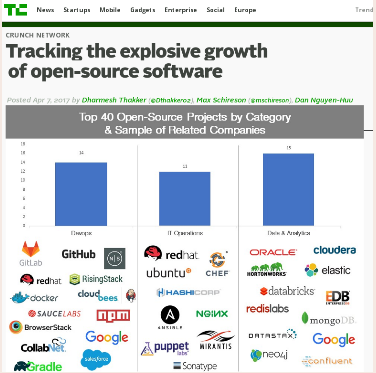
This week, Hewlett-Packard (where I am on the board) announced that it is exploring jettisoning its struggling PC business in favor of investing more heavily in software, where it sees better potential for growth. Meanwhile, Google plans to buy up the cellphone handset maker Motorola Mobility. Both moves surprised the tech world. But both moves are also in line with a trend I've observed, one that makes me optimistic about the future

Software companies outperform

or buy out traditional companies

Marc Andreessen, 2011

Open Source is eating the Software World



TC News Startups Mobile Gadgets Enterprise Social Europe Trend

CRUNCH NETWORK

Tracking the explosive growth of open-source software

Posted Apr 7, 2017 by Dharmesh Thakker (@Dthakker02), Max Schireson (@mschireson), Dan Nguyen-Huu

Reuse is the new rule

80% to 90% of a new application is … just reuse! (Sonatype survey, 2017)

… concerns are growing too … KYSW is coming!

Where does reused software come from?

Sourceforge
Debian
CPAN
Gitorious
Maven
Inria
Bitbucket
GitHub
BerliOs
CTAN
Gitlab
CRAN
GoogleCode
Adullact

Do *you* know where it comes from?

- the software you ship
- the software you use
- the software you acquire
- the software that
  - has that bug
  - has that vulnerability

KYSW: Know Your SoftWare

Like KYC in banking, KYSW is now essential all over IT…

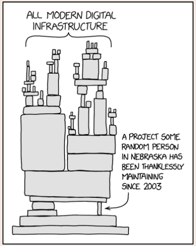
Sec. 4. Enhancing Software Supply Chain Security

ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software

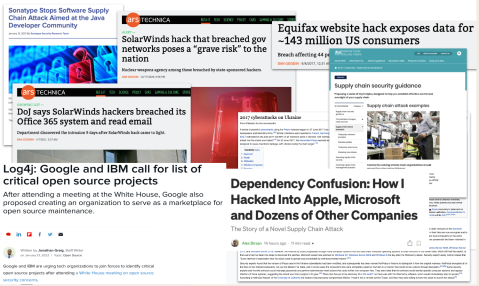
May 2021 POTUS Executive Order

Software supply chain and its issues

Complex digital infrastructure


ALL MODERN DIGITAL INFRASTRUCTURE
A PROJECT SOME RANDOM PERSON IN NEBRASKA HAS BEEN THANKLESSLY MAINTAINING SINCE 2003

Software supply chain in the news



Sonatype Stops Software Supply Chain Attack Aimed at the Java Developer Community

January 13, 2021 By Sonatype Security Research Team

ALERT —

SolarWinds hack that breached gov networks poses a “grave risk” to the nation

Nuclear weapons agency among those breached by state-sponsored hackers.

DAN GOODIN - 12/17/2020, 9:56 PM

Equifax website hack exposes data for ~143 million US consumers

Breach affecting 44 pe

DAN GOODIN - 9/8/2017, 12:31 AM

Supply chain security guidance

Proposing a series of 12 principles, designed to help you establish effective control and oversight of your supply chain.

Supply chain attack examples

Constantly evolving attacks mean organisations should

GROWING LIST —

DoJ says SolarWinds hackers breached its Office 365 system and read email

Department discovered the intrusion 9 days after SolarWinds hack came to light.

DAN GOODIN - 1/7/2021, 3:27 AM

2017 cyberattacks on Ukraine

From Wikipedia, the free encyclopedia

Contents [hide]
1 Approach
2 Attack
3 Attribution
4 Affected companies

Log4j: Google and IBM call for list of critical open source projects

After attending a meeting at the White House, Google also proposed creating an organization to serve as a marketplace for open source maintenance.

Written by Jonathan Greig, Staff Writer on January 13, 2022 | Topic: Open Source

Google and IBM are urging tech organizations to join forces to identify critical open source projects after attending a White House meeting on open source security concerns.

Dependency Confusion: How I Hacked Into Apple, Microsoft and Dozens of Other Companies

The Story of a Novel Supply Chain Attack

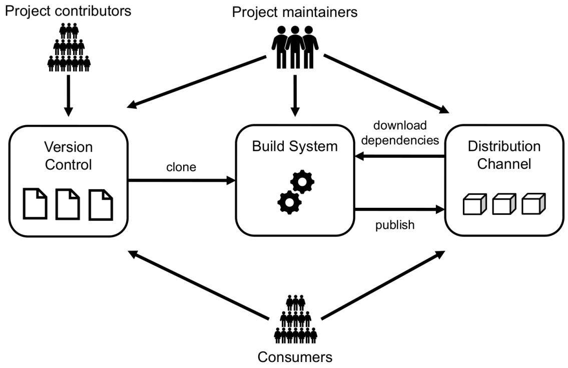
Alex Birsan 14 hours ago · 11 min read ★

Software Supply Chain attacks

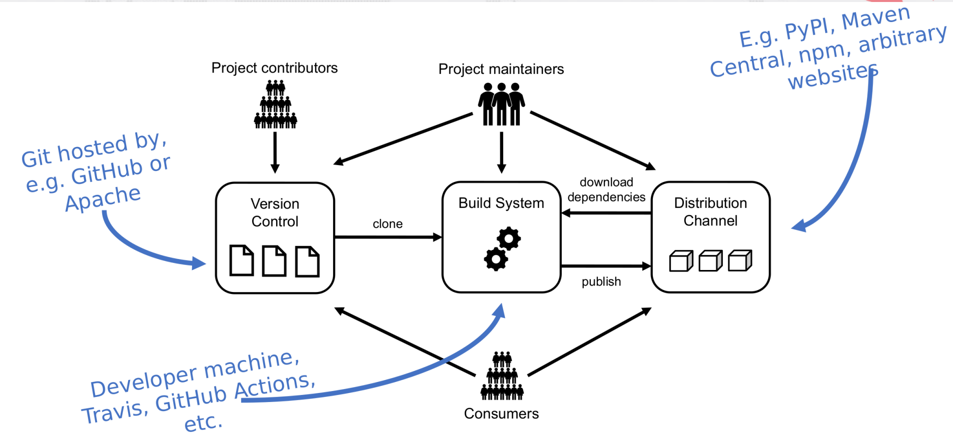
Malicious code injection into software components to compromise downstream users

March 2022	node-ipc and peacenotwar (CVE-2022-23812)
Dec 2021	Apache Log4j Remote Code Execution (Log4Shell, CVE-2021-44228)
Nov 2018	Attack on NPM package event-stream

Software supply chain in a picture



Software supply chain in a picture



Project contributors
Project maintainers
Version Control
clone
Build System
download dependencies
publish
Distribution Channel
Consumers
Git hosted by, e.g. GitHub or Apache
E.g. PyPI, Maven Central, npm, arbitrary websites
Developer machine, Travis, GitHub Actions, etc.

A long road ahead

Vertical approach

improve security of *each component* separately

Horizontal approach

explore *the whole supply chain*

A few key challenging properties

findability	needs qualified metadata
availability	needs an archive and a system of identifiers
integrity	needs crypto
traceability	needs a global provenance database
reproducibility	needs groundbreaking tools

We need a *global coordinated effort…*

and a *common, open, shared* infrastructure to track *all (Open Source) software*!

Outline

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

Software Heritage, in a nutshell

www.softwareheritage.org

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

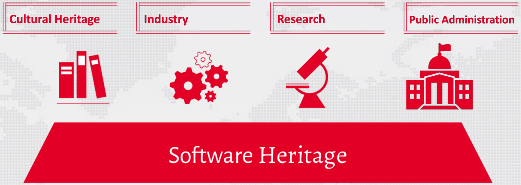
preserve and share all software source code

Research infrastructure

enable analysis of all software source code

Universal software archive, principled http://bit.ly/swhpaper

One infrastructure
open and shared



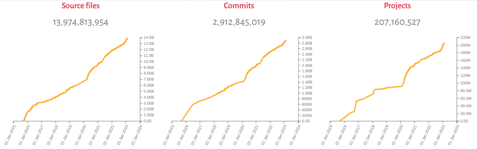
Cultural Heritage

Industry

Research

Public Administration

Software Heritage



Largest archive

Technology

- transparency and FOSS
- replicas all the way down

Content (billions!)

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

An international, non profit initiative built for the long term

Sharing the vision

UNESCO
United Nations
Educational, Scientific and
Cultural Organization

IEEE Software · GitLab · Engineering · eclipse · OW2 · fsfe · nlnet Foundation · SIGSOFT · Liberté • Égalité • Fraternité République Française · INAPI · The Linux Foundation · ADULLACT · Informatics Europe · Computer History Museum · software freedom conservancy · Software Freedom Law Center · AdaCore · gandi.net · TDF · Free Software Foundation · creative commons · SIF · open source initiative · openinventionnetwork

And many more ...
www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Inria

Diamond sponsor
cea

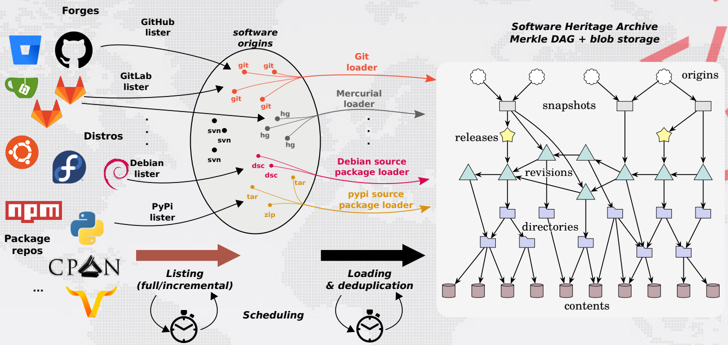
Platinum sponsors
cnrs · Société Générale · intel · Microsoft · Ministère de l'Enseignement Supérieur et de la Recherche · Huawei

Gold sponsors
openinventionnetwork · Sorbonne Université · Université de Paris

Silver sponsors
AdaCore · Ministère des Armées · DGA · République Française · GitHub · Google · Università di Pisa

Bronze sponsors
Red Hat · Scuola Normale Superiore · Alma Mater Studiorum

A peek under the hood: a universal archive

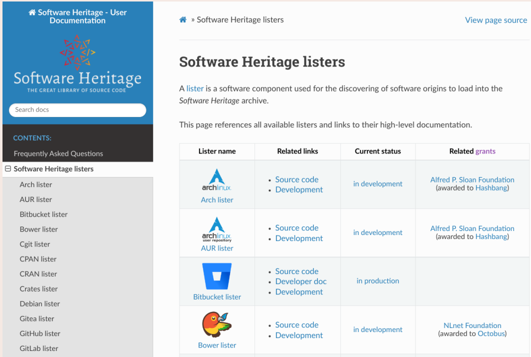


Global development history permanently archived in a uniform data model

- over 14 billion unique source files from over 210 million software projects
- ~1PB (compressed) blobs, ~30 B nodes, ~400 B edges

A peek under the hood: listers and loaders

Supported listers (index)



Software Heritage - User Documentation

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Search docs

CONTENTS:

Frequently Asked Questions

⊟ Software Heritage listers

- Arch lister
- AUR lister
- Bitbucket lister
- Bower lister
- Cgit lister
- CPAN lister
- CRAN lister
- Crates lister
- Debian lister
- Gitea lister
- GitHub lister
- GitLab lister

» Software Heritage listers

View page source

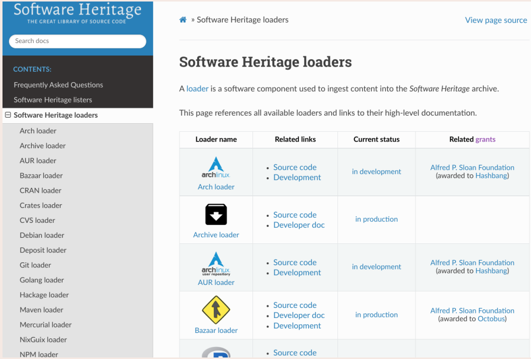
Software Heritage listers

A lister is a software component used for the discovering of software origins to load into the *Software Heritage* archive.

This page references all available listers and links to their high-level documentation.

Lister name	Related links	Current status	Related grants
Arch lister	• Source code • Development	in development	Alfred P. Sloan Foundation (awarded to Hashbang)
AUR lister	• Source code • Development	in development	Alfred P. Sloan Foundation (awarded to Hashbang)
Bitbucket lister	• Source code • Developer doc • Development	in production	
Bower lister	• Source code • Development	in development	NLnet Foundation (awarded to Octobus)

Supported loaders (index)



Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Search docs

CONTENTS:

Frequently Asked Questions

Software Heritage listers

⊟ Software Heritage loaders

- Arch loader
- Archive loader
- AUR loader
- Bazaar loader
- CRAN loader
- Crates loader
- CVS loader
- Debian loader
- Deposit loader
- Git loader
- Golang loader
- Hackage loader
- Maven loader
- Mercurial loader
- NixGuix loader
- NPM loader

» Software Heritage loaders

View page source

Software Heritage loaders

A loader is a software component used to ingest content into the *Software Heritage* archive.

This page references all available loaders and links to their high-level documentation.

Loader name	Related links	Current status	Related grants
Arch loader	• Source code • Development	in development	Alfred P. Sloan Foundation (awarded to Hashbang)
Archive loader	• Source code • Developer doc	in production	
AUR loader	• Source code • Development	in development	Alfred P. Sloan Foundation (awarded to Hashbang)
Bazaar loader	• Source code • Developer doc • Development	in production	Alfred P. Sloan Foundation (awarded to Octobus)
	• Source code		

Many contributed from external experts

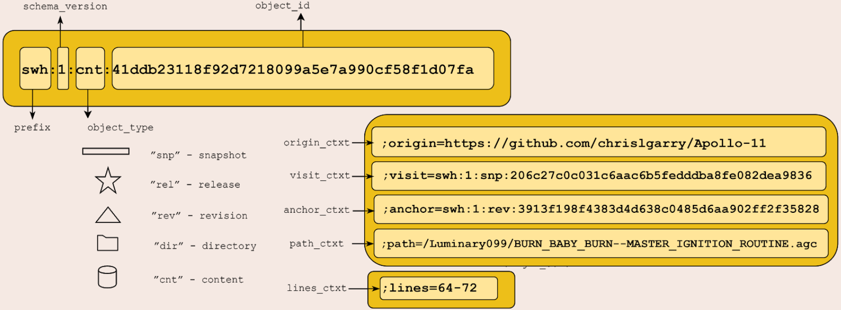
thanks to support of Alfred P. Sloan and NLNet foundations

Intrinsic Identifiers for software artefacts

Software Heritage Identifiers (SWHID)

link to full docs

25+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



schema_version

object_id

swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

prefix

object_type

▭ "snp" - snapshot

☆ "rel" - release

△ "rev" - revision

🗀 "dir" - directory

⛁ "cnt" - content

origin_ctxt → ;origin=https://github.com/chrislgarry/Apollo-11

visit_ctxt → ;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836

anchor_ctxt → ;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828

path_ctxt → ;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc

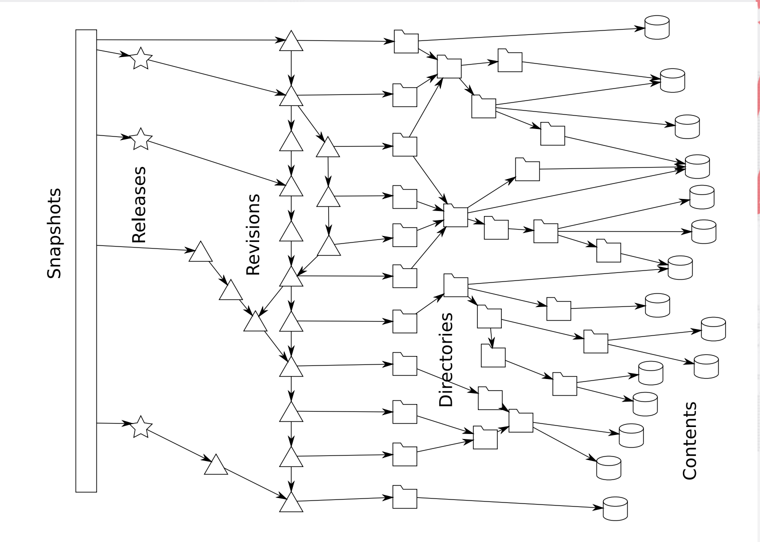
lines_ctxt → ;lines=64-72

Emerging standard : Linux Foundation SPDX 2.2; IANA registered; WikiData P6138

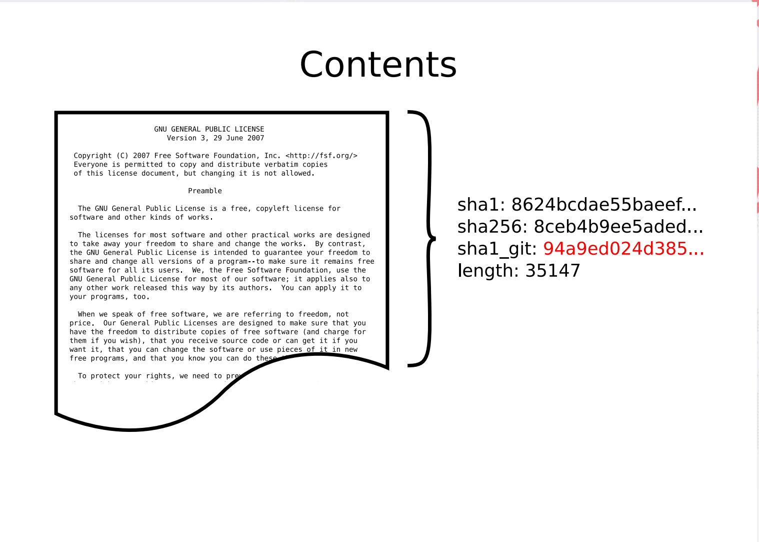
Full fledged *source code references* for reproducibility

Examples: Apollo 11 AGC excerpt, Quake III rsqrt; Guidelines available, see ICMS 2020

Zoom on the Software Heritage Merkle DAG



Zoom on the Software Heritage Merkle DAG

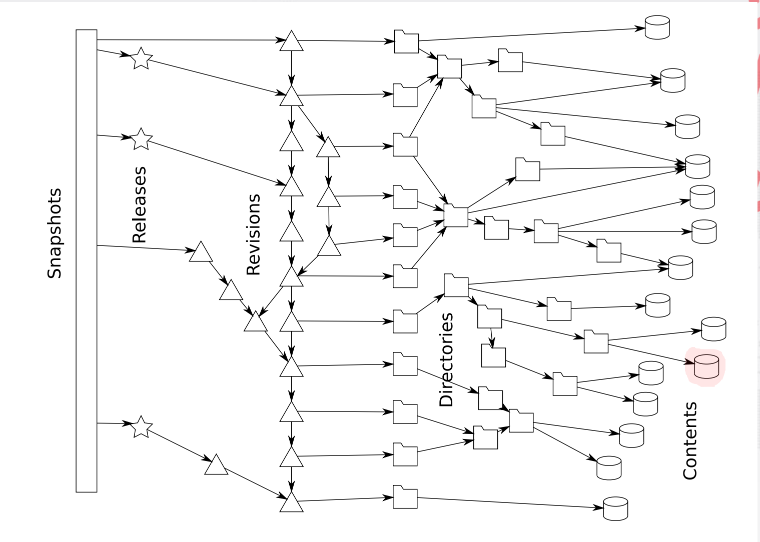


Contents

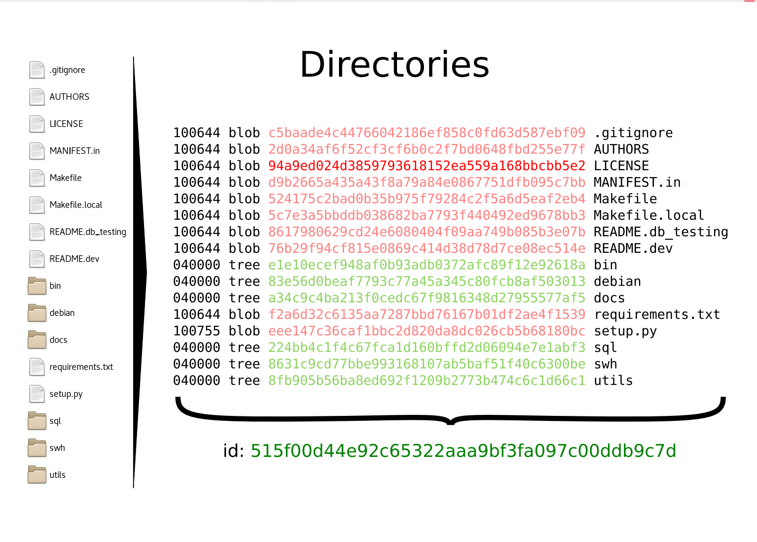
```
                    GNU GENERAL PUBLIC LICENSE
                       Version 3, 29 June 2007

 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.

                            Preamble

  The GNU General Public License is a free, copyleft license for
software and other kinds of works.

  The licenses for most software and other practical works are designed
to take away your freedom to share and change the works.  By contrast,
the GNU General Public License is intended to guarantee your freedom to
share and change all versions of a program--to make sure it remains free
software for all its users.  We, the Free Software Foundation, use the
GNU General Public License for most of our software; it applies also to
any other work released this way by its authors.  You can apply it to
your programs, too.

  When we speak of free software, we are referring to freedom, not
price.  Our General Public Licenses are designed to make sure that you
have the freedom to distribute copies of free software (and charge for
them if you wish), that you receive source code or can get it if you
want it, that you can change the software or use pieces of it in new
free programs, and that you know you can do these

  To protect your rights, we need to pre
```

sha1: 8624bcdae55baeef...
sha256: 8ceb4b9ee5aded...
sha1_git: 94a9ed024d385...
length: 35147

Zoom on the Software Heritage Merkle DAG



Zoom on the Software Heritage Merkle DAG



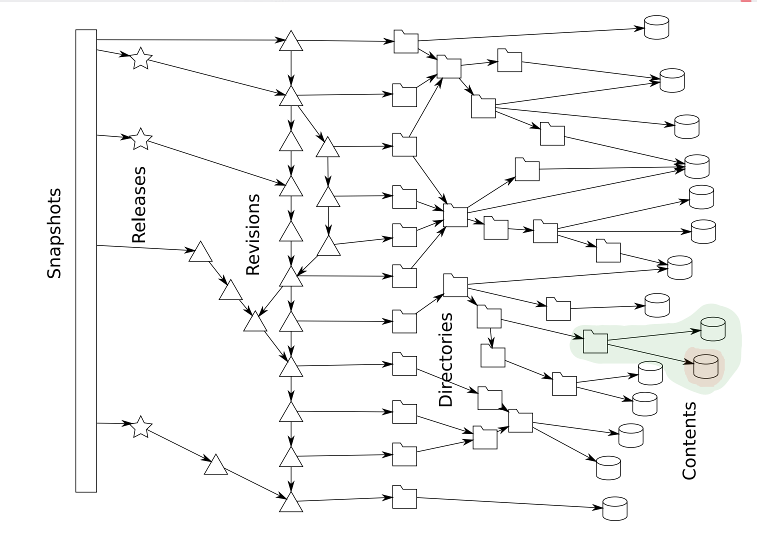
.gitignore
AUTHORS
LICENSE
MANIFEST.in
Makefile
Makefile.local
README.db_testing
README.dev
bin
debian
docs
requirements.txt
setup.py
sql
swh
utils

Directories

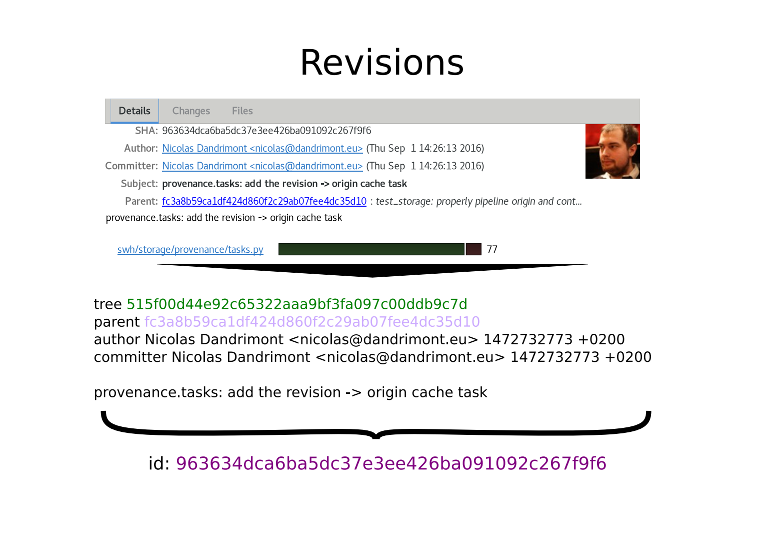
```
100644 blob c5baade4c44766042186ef858c0fd63d587ebf09 .gitignore
100644 blob 2d0a34af6f52cf3cf6b0c2f7bd0648fbd255e77f AUTHORS
100644 blob 94a9ed024d3859793618152ea559a168bbcbb5e2 LICENSE
100644 blob d9b2665a435a43f8a79a84e0867751dfb095c7bb MANIFEST.in
100644 blob 524175c2bad0b35b975f79284c2f5a6d5eaf2eb4 Makefile
100644 blob 5c7e3a5bbddb038682ba7793f440492ed9678bb3 Makefile.local
100644 blob 8617980629cd24e6080404f09aa749b085b3e07b README.db_testing
100644 blob 76b29f94cf815e0869c414d38d78d7ce08ec514e README.dev
040000 tree e1e10ecef948af0b93adb0372afc89f12e92618a bin
040000 tree 83e56d0beaf7793c77a45a345c80fcb8af503013 debian
040000 tree a34c9c4ba213f0cedc67f9816348d27955577af5 docs
100644 blob f2a6d32c6135aa7287bbd76167b01df2ae4f1539 requirements.txt
100755 blob eee147c36caf1bbc2d820da8dc026cb5b68180bc setup.py
040000 tree 224bb4c1f4c67fca1d160bffd2d06094e7e1abf3 sql
040000 tree 8631c9cd77bbe993168107ab5baf51f40c6300be swh
040000 tree 8fb905b56ba8ed692f1209b2773b474c6c1d66c1 utils
```

id: 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d

Zoom on the Software Heritage Merkle DAG



Zoom on the Software Heritage Merkle DAG



Revisions

SHA: 963634dca6ba5dc37e3ee426ba091092c267f9f6
Author: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)
Committer: Nicolas Dandrimont <nicolas@dandrimont.eu> (Thu Sep 1 14:26:13 2016)
Subject: **provenance.tasks: add the revision -> origin cache task**
Parent: fc3a8b59ca1df424d860f2c29ab07fee4dc35d10 : *test_storage: properly pipeline origin and cont...*

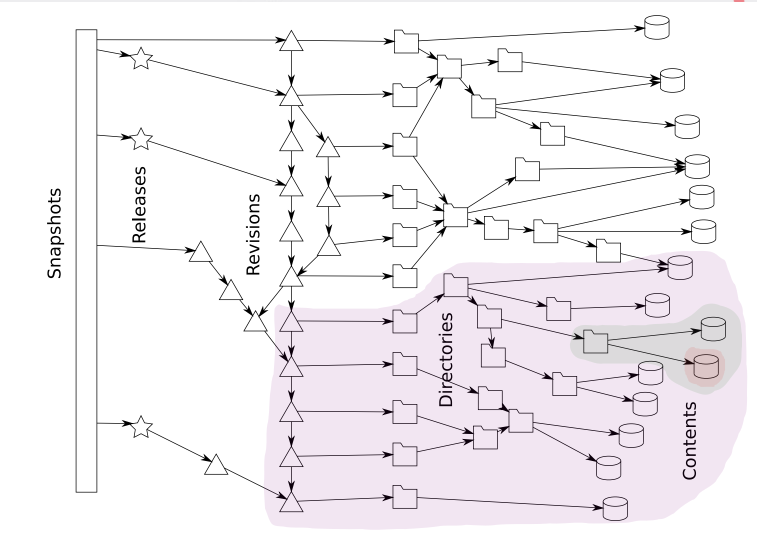
provenance.tasks: add the revision -> origin cache task

swh/storage/provenance/tasks.py 77

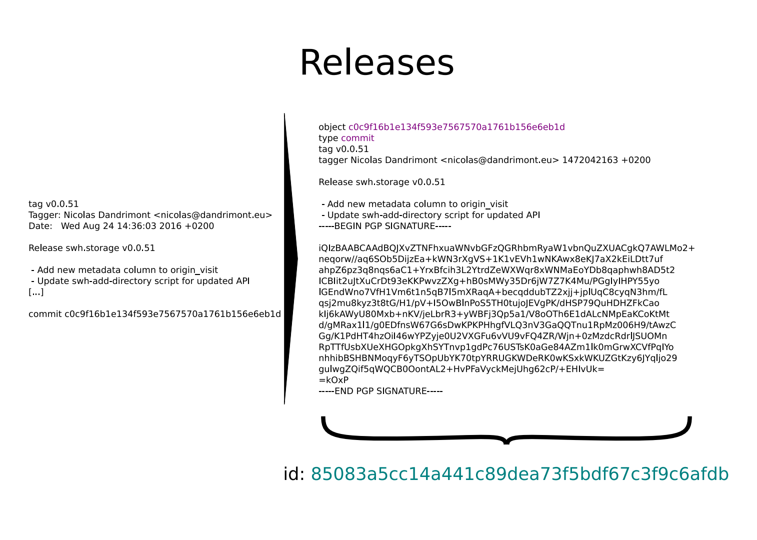
```
tree 515f00d44e92c65322aaa9bf3fa097c00ddb9c7d
parent fc3a8b59ca1df424d860f2c29ab07fee4dc35d10
author Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200
committer Nicolas Dandrimont <nicolas@dandrimont.eu> 1472732773 +0200

provenance.tasks: add the revision -> origin cache task
```

id: 963634dca6ba5dc37e3ee426ba091092c267f9f6

Zoom on the Software Heritage Merkle DAG



Zoom on the Software Heritage Merkle DAG



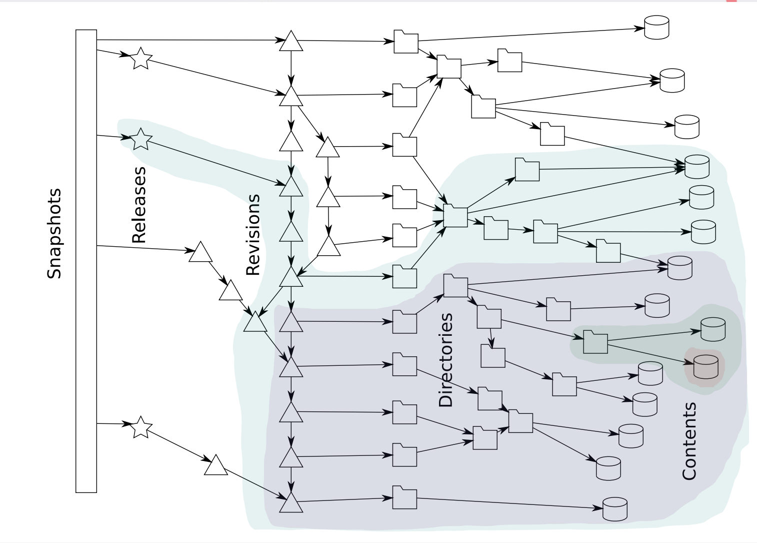
Releases

```
tag v0.0.51
Tagger: Nicolas Dandrimont <nicolas@dandrimont.eu>
Date:   Wed Aug 24 14:36:03 2016 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
[...]

commit c0c9f16b1e134f593e7567570a1761b156e6eb1d
```

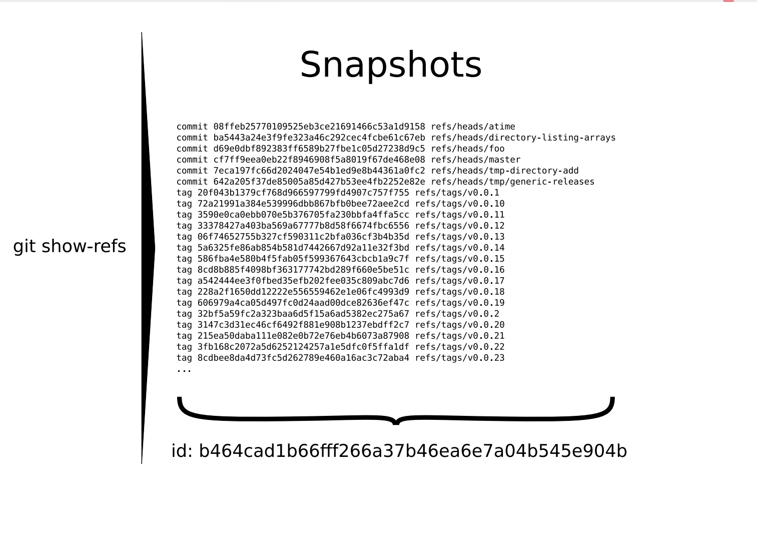
```
object c0c9f16b1e134f593e7567570a1761b156e6eb1d
type commit
tag v0.0.51
tagger Nicolas Dandrimont <nicolas@dandrimont.eu> 1472042163 +0200

Release swh.storage v0.0.51

 - Add new metadata column to origin_visit
 - Update swh-add-directory script for updated API
-----BEGIN PGP SIGNATURE-----

iQIzBAABCAAdBQJXvZTNFhxuaWNvbGFzQGRhbmRyaW1vbnQuZXUACgkQ7AWLMo2+
neqorw//aq6SOb5DijzEa+kWN3rXgVS+1K1vEVh1wNKAwx8eKJ7aX2kEiLDtt7uf
ahpZ6pz3q8nqs6aC1+YrxBfcih3L2YtrdZeWXWqr8xWNMaEoYDb8qaphwh8AD5t2
lCBIit2uJtXuCrDt93eKKPwvzZXg+hB0sMWy35Dr6jW7Z7K4Mu/PGgIyIHPY55yo
lGEndWno7VfH1Vm6t1n5qB7l5mXRaqA+becqddubTZ2xjj+jplUqC8cyqN3hm/fL
qsj2mu8kyz3t8tG/H1/pV+I5OwBlnPoS5TH0tujoJEVgPK/dHSP79QuHDHZFkCao
kIj6kAWyU80Mxb+nKV/jeLbrR3+yWBFj3Qp5a1/V8oOTh6E1dALcNMpEaKCoKtMt
d/gMRax1l1/g0EDfnsW67G6sDwKPKPHhgfVLQ3nV3GaQQTnu1RpMz006H9/tAwzC
Gg/K1PdHT4hzOiI46wYPZyje0U2VXGFu6vVU9vFQ4ZR/Wjn+0zMzdcRdrlJSUOMn
RpTTfUsbXUeXHGOpkgXhSYTnvp1gdPc76USTsK0aGe84AZm1lk0mGrwXCVfPqIYo
nhhibBSHBNMoqyF6yTSOpUbYK70tpYRRUGKWDeRK0wKSxkWKUZGtKzy6JYqIjo29
gulwgZQif5qWQCB0OontAL2+HvPFaVyckMejUhg62cP/+EHIvUk=
=kOxP
-----END PGP SIGNATURE-----
```

id: 85083a5cc14a441c89dea73f5bdf67c3f9c6afdb

Zoom on the Software Heritage Merkle DAG



Zoom on the Software Heritage Merkle DAG



```
commit 08ffeb25770109525eb3ce21691466c53a1d9158 refs/heads/atime
commit ba5443a24e3f9fe323a46c292cec4fcbe61c67eb refs/heads/directory-listing-arrays
commit d69e0dbf892383ff6589b27fbe1c05d27238d9c5 refs/heads/foo
commit cf7ff9eea0eb22f8946908f5a8019f67de468e08 refs/heads/master
commit 7eca197fc66d2024047e54b1ed9e8b44361a0fc2 refs/heads/tmp-directory-add
commit 642a205f37de85005a85d427b53ee4fb2252e82e refs/heads/tmp/generic-releases
tag 20f043b1379cf768d966597799fd4907c757f755 refs/tags/v0.0.1
tag 72a21991a384e539996dbb867bfb0bee72aee2cd refs/tags/v0.0.10
tag 3590e0ca0ebb070e5b376705fa230bbfa4ffa5cc refs/tags/v0.0.11
tag 33378427a403ba569a67777b8d58f6674fbc6556 refs/tags/v0.0.12
tag 06f74652755b327cf590311c2bfa036cf3b4b35d refs/tags/v0.0.13
tag 5a6325fe86ab854b581d7442667d92a11e32f3bd refs/tags/v0.0.14
tag 586fba4e580b4f5fab05f599367643cbcb1a9c7f refs/tags/v0.0.15
tag 8cd8b885f4098bf363177742bd289f660e5be51c refs/tags/v0.0.16
tag a542444ee3f0fbed35efb202fee035c809abc7d6 refs/tags/v0.0.17
tag 228a2f1650dd12222e556559462e1e06fc4993d9 refs/tags/v0.0.18
tag 606979a4ca05d497fc0d24aad00dce82636ef47c refs/tags/v0.0.19
tag 32bf5a59fc2a323baa6d5f15a6ad5382ec275a67 refs/tags/v0.0.2
tag 3147c3d31ec46cf6492f881e908b1237ebdff2c7 refs/tags/v0.0.20
tag 215ea50daba111e082e0b72e76eb4b6073a87908 refs/tags/v0.0.21
tag 3fb168c2072a5d6252124257a1e5dfc0f5ffa1df refs/tags/v0.0.22
tag 8cdbee8da4d73fc5d262789e460a16ac3c72aba4 refs/tags/v0.0.23
...
```

id: b464cad1b66fff266a37b46ea6e7a04b545e904b

A quick tour

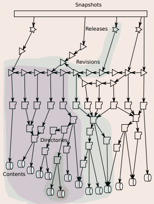
- Browse (e.g. Apollo 11, and your work may be already there !)
- Trigger archival, use the updateswh browser extension (GitHub action available too)
- Get and use SWHIDs (full specification available online)
- Cite software with biblatex-software package from CTAN
  - Overleaf ACMART template available
- 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

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

A revolutionary infrastructure for industry

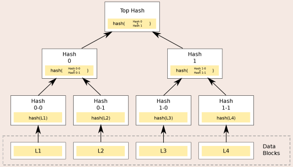
The *graph* of Software Development



All of the software development in a single graph!

- lookup by content hash
- wayback machine for software development
    - http://archive.softwareheritage.org/
- … and much more

The *blockchain* of Software Development



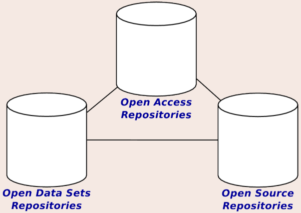
All of a software development… in a single Merkle graph!

Widely used crypto (e.g., Git, blockchains, IPFS, …)

- built-in deduplication
- intrinsic, unforgeable identifiers at all levels
- simplifies traceability (licensing, supply chain management)

A revolutionary infrastructure for research and innovation

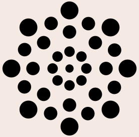
A *pillar* of Open Science



The *reference archive* of Research Software for Open Science

- curated deposit of research software
  - in collaboration with HAL, CCSD and Inria IES
  - now open *to all researchers*!
- intrinsic identifiers for reproducibility

Reference platform for *Big Code*



- unique observatory of all software development
- big data, machine learning paradise: classification, trends, coding patterns, code completion…

Industry use cases (selection)

Open Source complete and corresponding source code distribution (Intel)

Software Heritage members can:

- archive source code in Software Heritage, distribute only the SWHID

Traceability and integrity (OIN for the *Linux System Definition*)

Software Heritage members can:

- archive source code in Software Heritage
- track it and verify its integrity using its SWHID

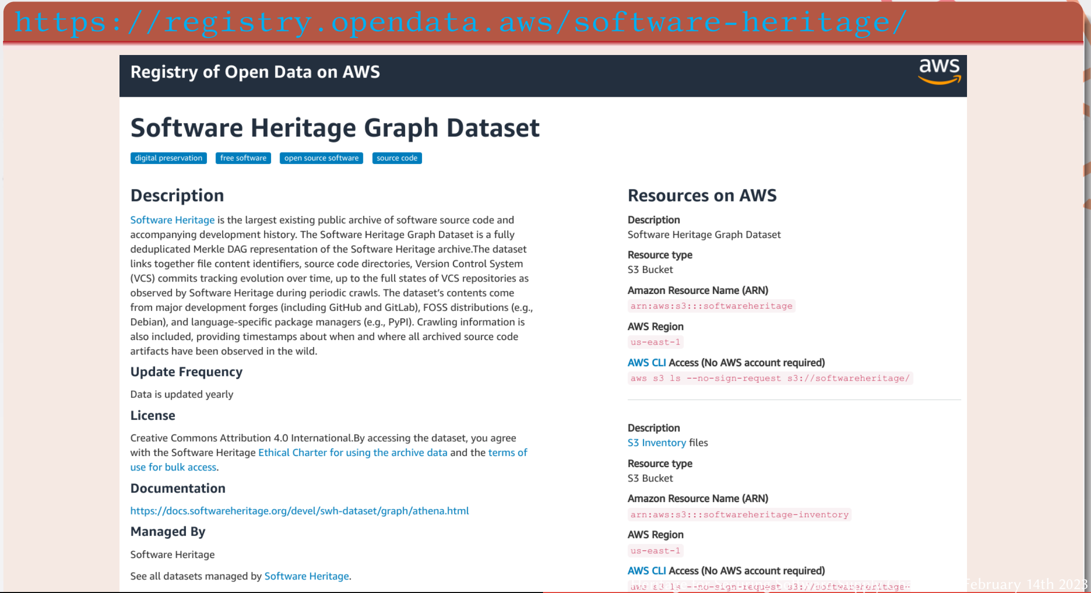
And much more!

- compliance (collaborations with Intel)
- security (large project with French Government)
- supply chain management, long term archive *add your use case here*

Outline

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. **Software Heritage Datasets**
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

The full graph in the AWS Open Data collection



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

Registry of Open Data on 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-
```

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/


  2018-09-25/
  2019-01-28-popular-3k-python/   2021-03-23-cpython-3-5/
  2019-01-28-popular-4k/          2021-03-23-popular-3k-python/
  2020-05-20/                     2021-03-23/
  2020-12-15/                     2022-04-25/
```

How to use

- online full documentation
- Antoine Pietri's PhD Thesis

How to cite

Antoine Pietri, Diomidis Spinellis, Stefano Zacchiroli. *The Software Heritage Graph Dataset: Public software development under one roof.* MSR 2019. (bibtex)

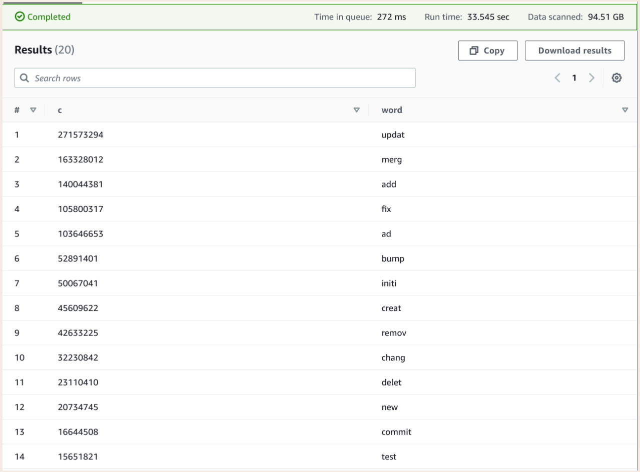
Example: most popular commit verbs (stemmed)

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;
```

Total cost: approximately .5 euros

Results



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

Results (20)

#	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

1 Introduction

2 (Open Source) Software Supply Chain

3 Meet Software Heritage

4 Revolutionary infrastructure

5 Software Heritage Datasets

6 **Efficient traversal of the full graph**

7 Impact on research studies

8 Conclusion

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 (25 B nodes, 350 B edges) in 200 GiB RAM

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

Java and gRPC APIs available

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

Examples

assume graph service on localhost:50091

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

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. **Impact on research studies**
8. Conclusion

Selected research works using Software Heritage

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

Stefano Zacchiroli
Gender Differences in Public Code Contributions: a 50-year Perspective
IEEE Softw. 38(2): 45-50 (2021)

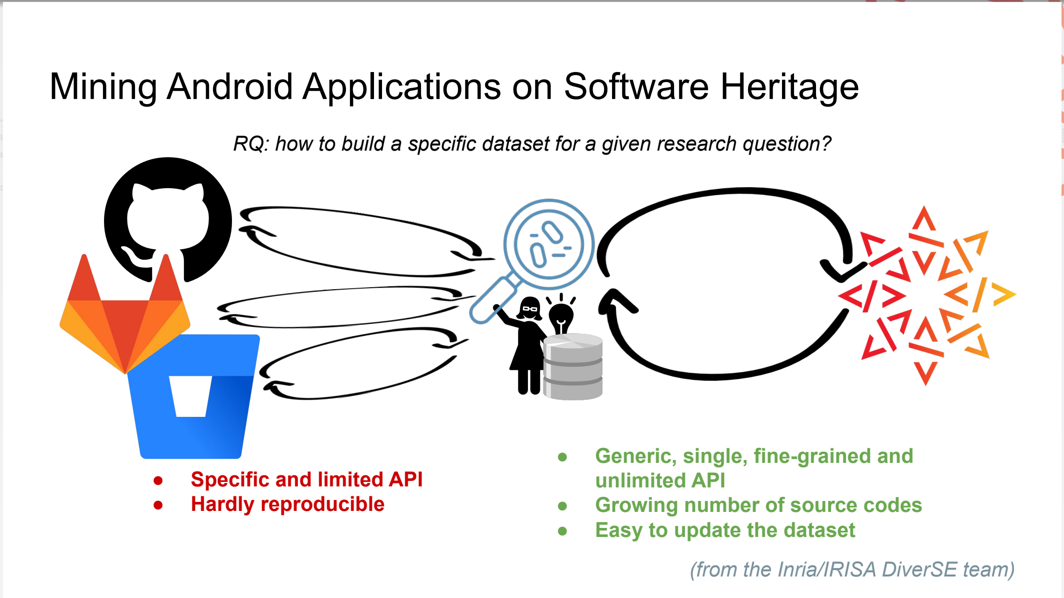
Antoine Pietri, Guillaume Rousseau, Stefano Zacchiroli
Forking Without Clicking: on How to Identify Software Repository Forks
MSR 2020: 17th Intl. Conf. on Mining Software Repositories. IEEE

Antoine Pietri, Guillaume Rousseau, Stefano Zacchiroli
Determining the Intrinsic Structure of Public Software Development History
MSR 2020: 17th 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

Roberto Di Cosmo, Guillaume Rousseau, Stefano Zacchiroli
Software Provenance Tracking at the Scale of Public Source Code
Empirical Software Engineering 25(4): 2930-2959 (2020)

From fellow researchers (Inria DiverSE team, thanks Benoît Combemale)



Mining Android Applications on Software Heritage

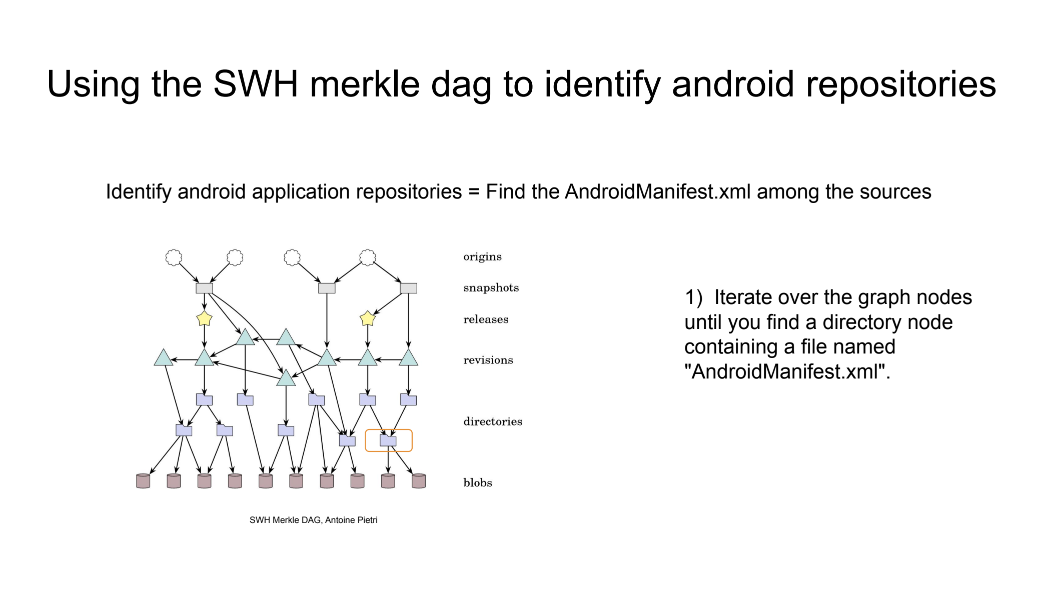
RQ: how to build a specific dataset for a given research question?

- 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)

From fellow researchers (Inria DiverSE team, thanks Benoît Combemale)



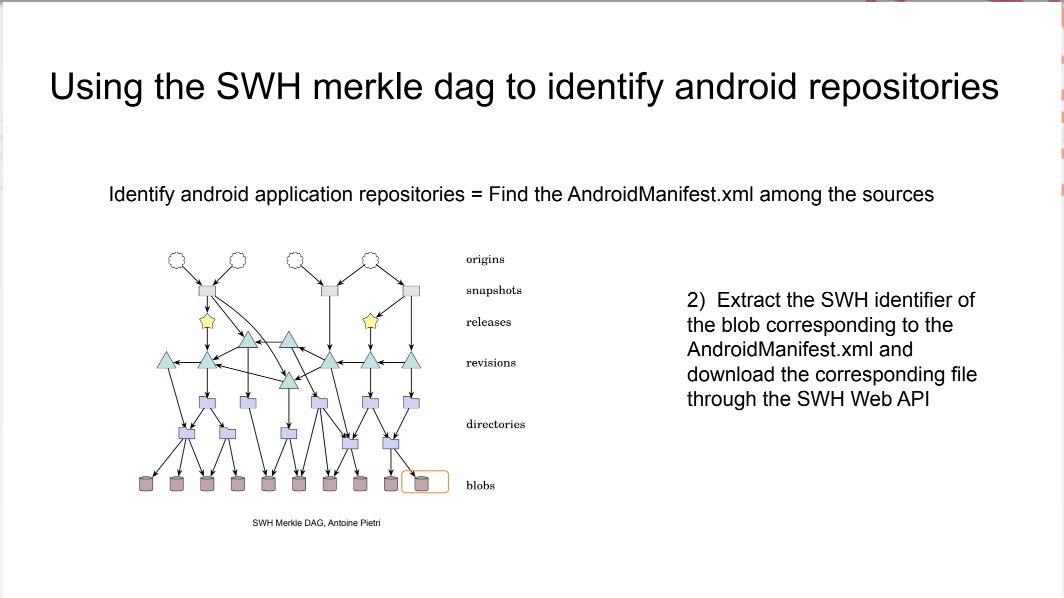
Using the SWH merkle dag to identify android repositories

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".

From fellow researchers (Inria DiverSE team, thanks Benoît Combemale)



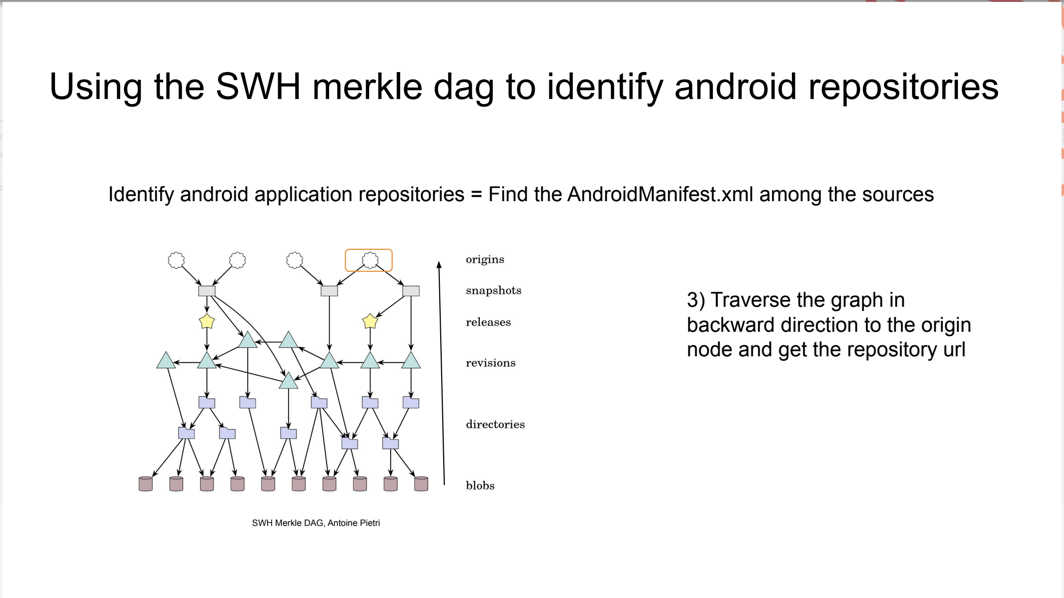
Using the SWH merkle dag to identify android repositories

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

From fellow researchers (Inria DiverSE team, thanks Benoît Combemale)



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

Outline

1. Introduction
2. (Open Source) Software Supply Chain
3. Meet Software Heritage
4. Revolutionary infrastructure
5. Software Heritage Datasets
6. Efficient traversal of the full graph
7. Impact on research studies
8. Conclusion

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!

develop the infrastructure, *use* in research, *build* tools, . . .

The floor is yours

Let’s work together! (PhD and job openings soon)

Questions?

References

- R. Di Cosmo, *A revolutionary infrastructure for Open Source*, 2021, EU Software Forum (slides) (video)
- French Ministry of Research, *Second National Plan for Open Science*
  2021, (online)
- R. Di Cosmo, *Archiving and Referencing Source Code with Software Heritage*
  International Conference on Mathematical Software 2020 (10.1007/978-3-030-52200-1_36)
- J.F. Abramatic, R. Di Cosmo, S. Zacchiroli, *Building the Universal Archive of Source Code*
  CACM, October 2018 (10.1145/3183558)