Software Heritage
The Great Library of Source Code

David Douard & Antoine R. Dumont

Software Engineer - Software Heritage
info@softwareheritage.org

27 juin 2019
PSES - Choisy-le-Roy

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Software is everywhere

Software is everywhere

Software embodies a growing part of…

… our scientific, *technical* and Cultural Heritage!

Source Code: *executable* and *human readable* knowledge

“The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Source Code: *executable* and *human readable* knowledge

"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

```
Hello World
```

Source Code: *executable* and *human readable* knowledge

"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Source Code: *executable* and *human readable* knowledge

"The source code for a work means the preferred form of the work for making modifications to it."

GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

~ 50 years, a lightning fast growth

Apollo 11 (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

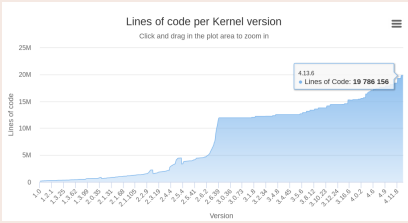
~ 50 years, a lightning fast growth

Apollo 11 (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel (in your pockets!)



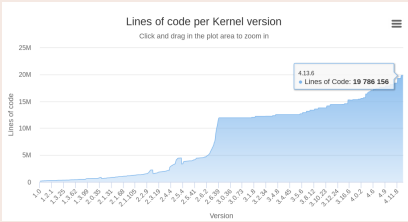
~ 50 years, a lightning fast growth

Apollo 11 (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel (in your pockets!)



Harold Abelson, Structure and Interpretation of Computer Programs (1985)

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

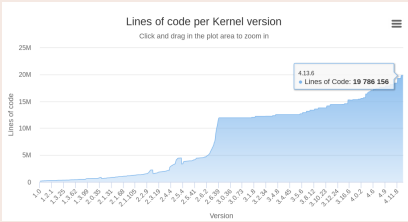
~ 50 years, a lightning fast growth

Apollo 11 (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West."

Margaret Hamilton

Linux Kernel (in your pockets!)



Harold Abelson, Structure and Interpretation of Computer Programs (1985)

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

Len Shustek, Computer History Museum (2006)

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

Software is fragile

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

Software is fragile

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

Where is the archive…

where do we go if (a repository on) GitHub or GitLab.com goes away?

We are at a turning point

Preserve the past

Only a few years left to recover the history of software technology

We are at a turning point

Preserve the past

Only a few years left to recover the history of software technology

Improve the future

We need a universal platform for all the future software source code

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

Collect, preserve and share the *source code* of *all the software* that is publicly available.

Past, present and future

Preserving the past, *enhancing* the present, *preparing* the future.

Core principles

Core principles

Open approach

- 100% Free and Open Source Software
- transparency

In for the long haul

- replication
- non profit

Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs)

We DO archive

- file content (= blobs)
- revisions (= commits), with full metadata
- releases (= tags), ditto
- where (origin) & when (visit) we found any of the above

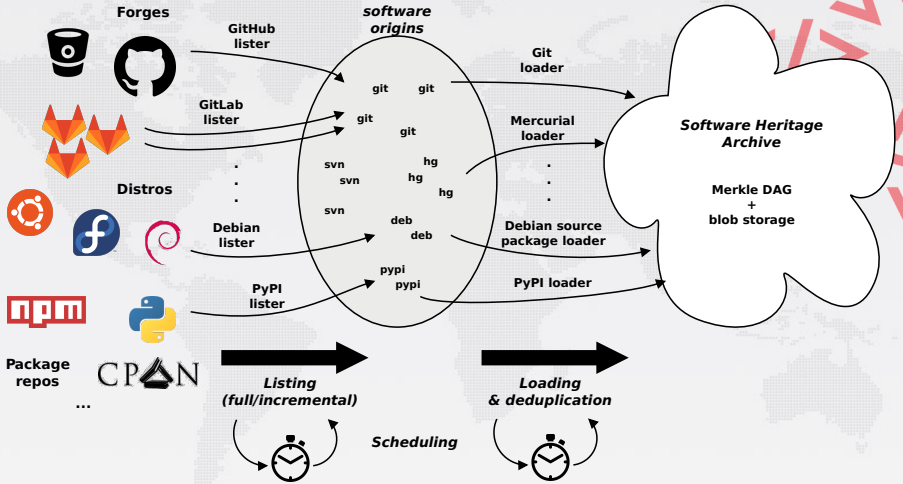
… in a VCS-/archive-agnostic canonical data model

We DON'T archive

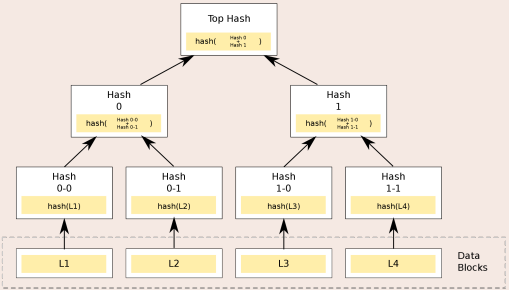
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

Long term vision: play our part in a *"semantic wikipedia of software"*

Data flow


Forges
GitHub lister
GitLab lister
Distros
Debian lister
PyPI lister
Package repos
CPAN
...
software origins
git
git
git
git
svn
svn
svn
hg
hg
hg
deb
deb
pypi
pypi
Git loader
Mercurial loader
Debian source package loader
PyPI loader
Software Heritage Archive
Merkle DAG + blob storage
Listing (full/incremental)
Scheduling
Loading & deduplication

Merkle tree (R. C. Merkle, Crypto 1979)

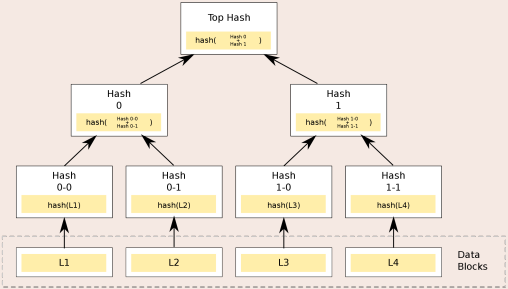


Combination of

- tree
- hash function

Merkle trees

Merkle tree (R. C. Merkle, Crypto 1979)



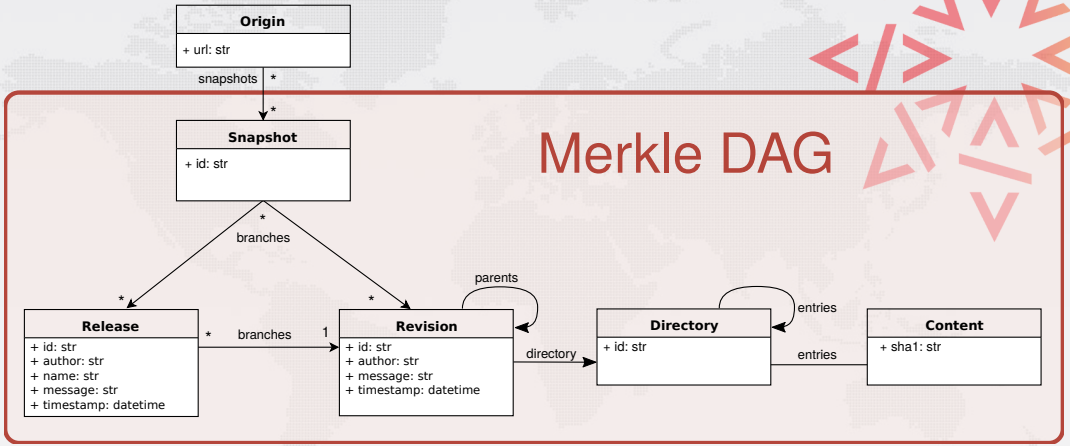
Combination of

- tree
- hash function

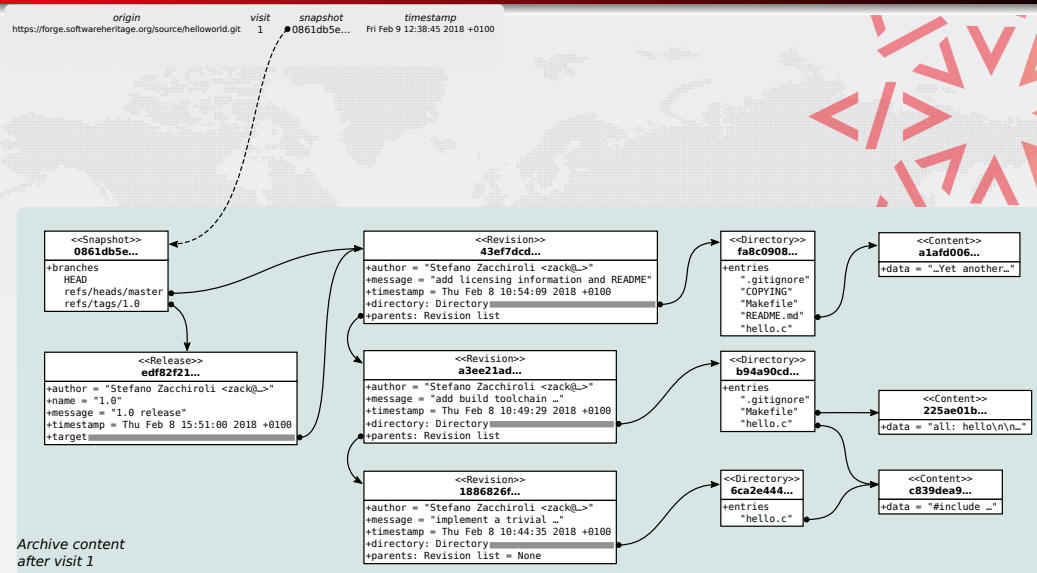
Classical cryptographic construction

- fast, parallel signature of large data structures
- widely used (e.g., Git, blockchains, IPFS, …)
- built-in deduplication

The archive: a (giant) Merkle DAG



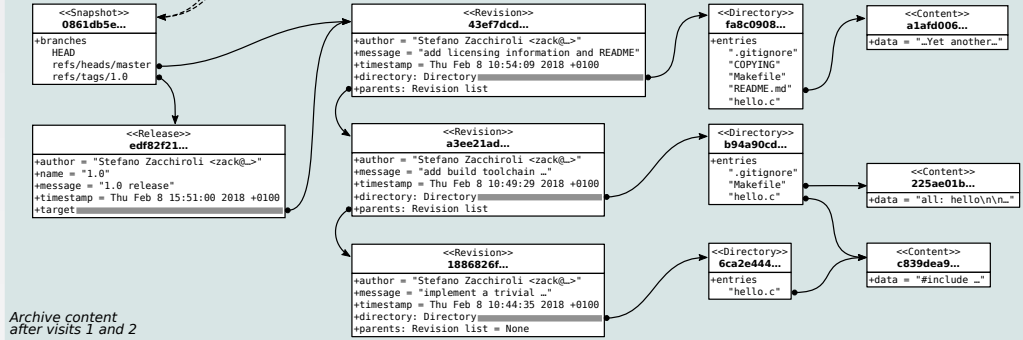
The archive: a (giant) Merkle DAG



Archive content after visit 1

The archive: a (giant) Merkle DAG

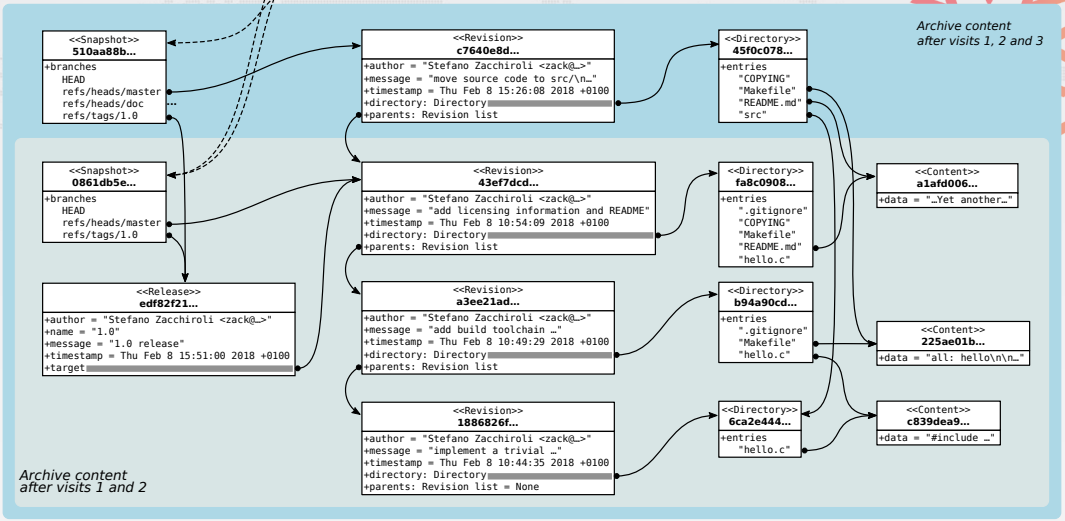
origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e…	Fri Feb 9 12:38:45 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	2	0861db5e…	Fri Feb 9 13:29:00 2018 +0100



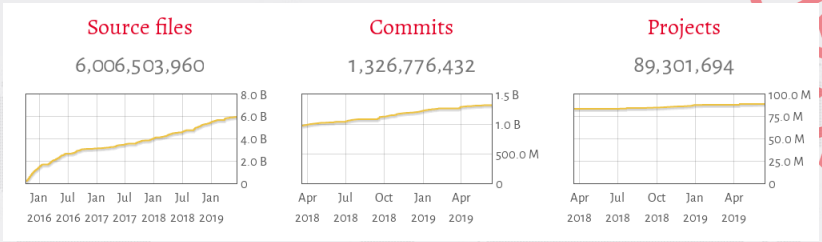
Archive content after visits 1 and 2

The archive: a (giant) Merkle DAG

origin	visit	snapshot	timestamp
https://forge.softwareheritage.org/source/helloworld.git	1	0861db5e…	Fri Feb 9 12:38:45 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	2	0861db5e…	Fri Feb 9 13:29:00 2018 +0100
https://forge.softwareheritage.org/source/helloworld.git	3	510aa88b…	Fri Feb 9 15:52:50 2018 +0100

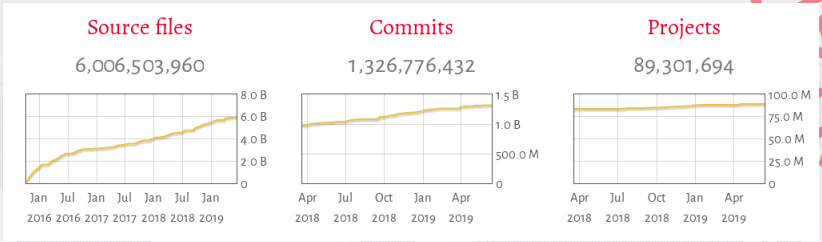


Archive coverage — archive.softwareheritage.org



GitHub · debian · GitLab · Google code · {P} · python Package Index · GITORIOUS · GNU · HAL archives-ouvertes.fr · Inria inventeurs du monde numérique

Archive coverage — archive.softwareheritage.org



GitHub debian GitLab Google code {P} python Package Index GITORIOUS GNU HAL archives-ouvertes.fr Inria inventeurs du monde numérique

- 410 TB (uncompressed) blobs, 20 B nodes, 280 B edges
- The *richest* public source code archive, … and growing daily!

Web user interface

Browser-based interface to browse the Software Heritage archive
https://archive.softwareheritage.org/browse/

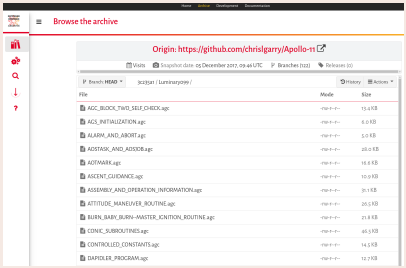
Features

- all REST API features, but good looking :-)
  - browsing: snapshots → revisions → directories → contents …
  - access to metadata and crawling information
- origin search, as full text indexing of origin URLs
- bulk download, via integration with the Vault

Visiting the archive: the Apollo 11 source code

Margaret Hamilton today

The Apollo 11 source code in SWH



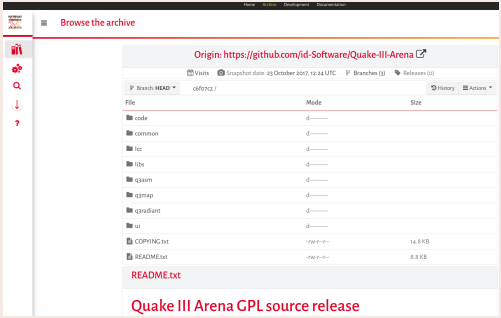
Some pointers

- Entry point
- Burn, baby, burn!

Visiting the archive: the Quake 3 source code

John Carmack

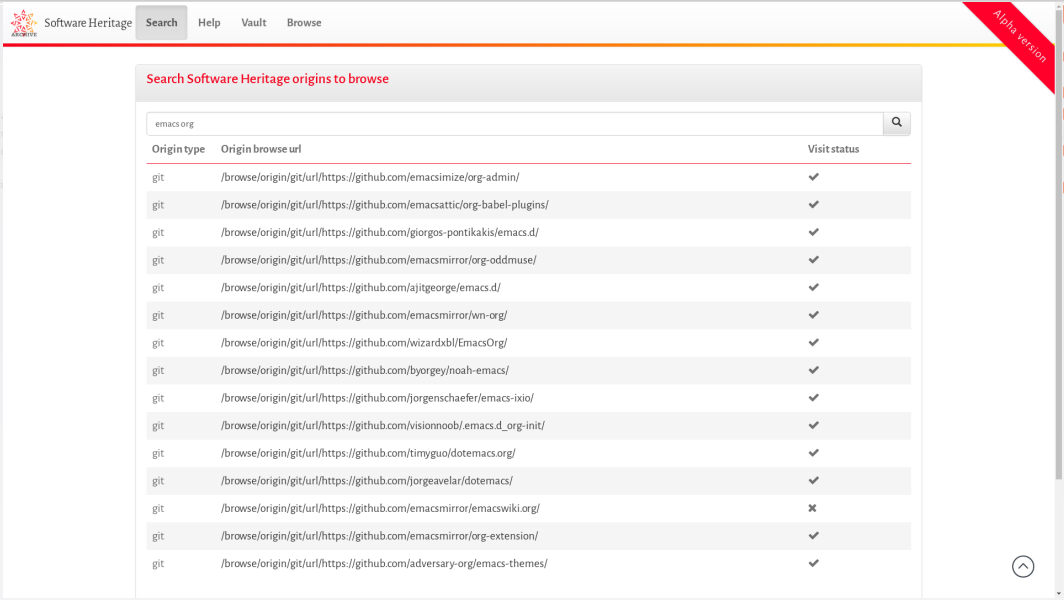
The Quake 3 source code in SWH



Some pointers

- Entry point
- What the f…

Web UI — origin search



Web UI — available visits

Software Heritage Search Help Vault Browse

Alpha version

SWH origin: https://github.com/hylang/hy

Go to origin

SWH origin visits

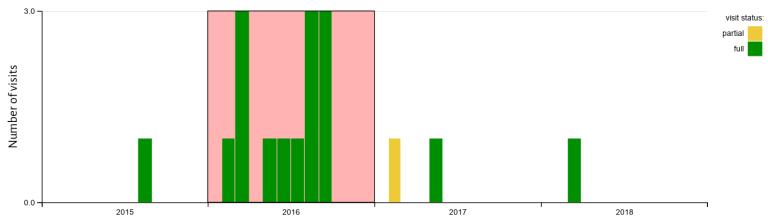
Visits overview

- Total number of visits: **17**
- Last full visit: ✔ 02 March 2018, 20:03 UTC
- First full visit: ✔ 04 August 2015, 22:26 UTC
- Last visit: ✔ 02 March 2018, 20:03 UTC

Visits history

Show full visits with different snapshots | Show all full visits | Show all visits

Timeline

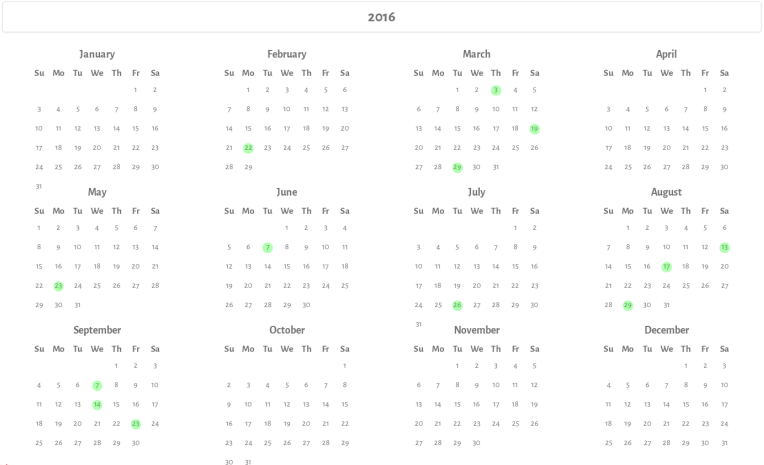


Calendar

2016

Web UI — calendar

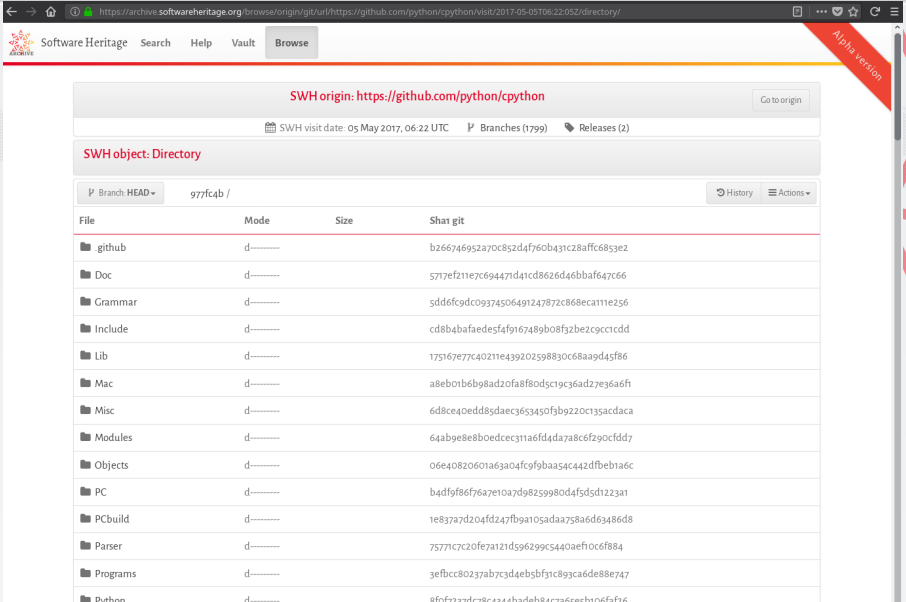
Calendar



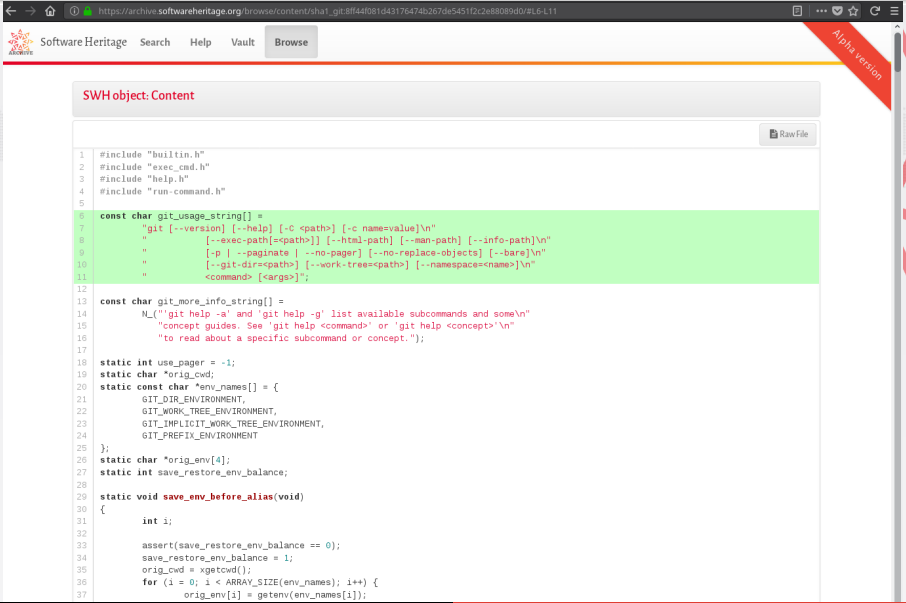
List

- 22 February 2016, 16:56 UTC
- 03 March 2016, 17:25 UTC
- 19 March 2016, 01:53 UTC
- 29 March 2016, 06:26 UTC
- 23 May 2016, 11:22 UTC
- 07 June 2016, 04:42 UTC
- 26 July 2016, 14:59 UTC
- 13 August 2016, 03:45 UTC
- 16 August 2016, 22:40 UTC
- 29 August 2016, 19:38 UTC
- 07 September 2016, 09:19 UTC
- 14 September 2016, 11:04 UTC

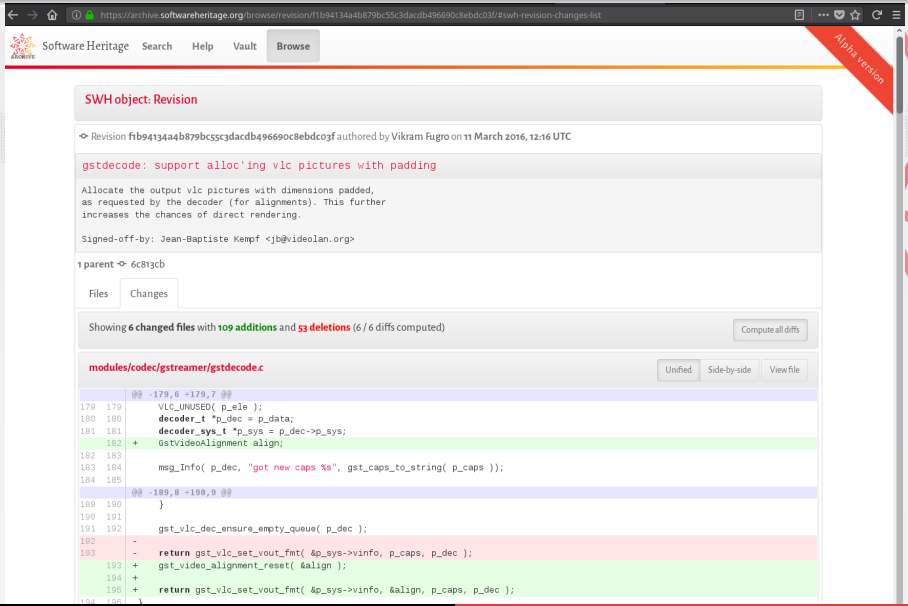
Web UI — directory browsing



Web UI — syntax highlighting and selection



Web UI — revisions as diffs



Software stack

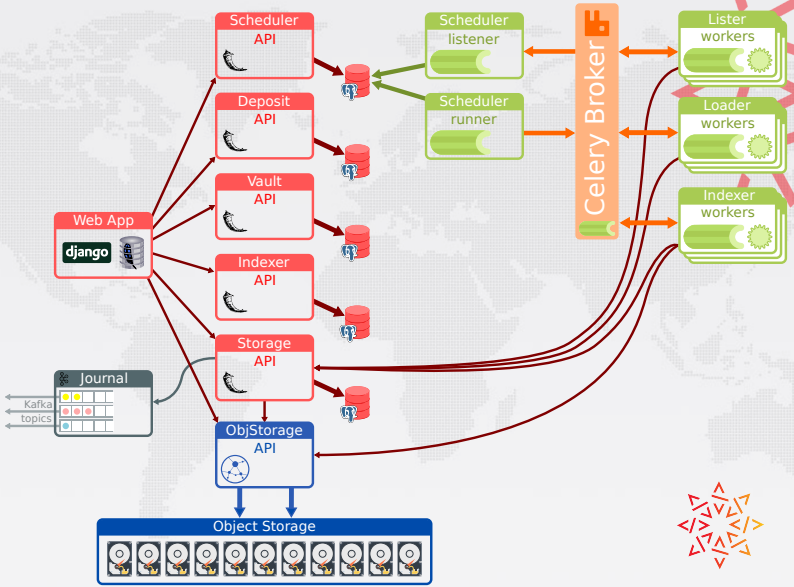
3rd party

- Debian, Puppet, Ceph
- PostgreSQL for metadata storage, with barman & pglogical
- Celery (RabbitMQ backend) for task scheduling
- Python3 and psycopg2 for the backend
- Django, Bootstrap, D3.js for Web stuff

in house

- *ad hoc* object storage (to avoid imposing tech to mirrors)
- data model implementation, listers, loaders, scheduler
- ~60 Git repositories (~20 Python packages, ~30 Puppet modules)
- ~60 kSLOC Python / ~12 kSLOC SQL / ~4 kSLOC Puppet
- licence choice: GPLv3 (backend) / AGPLv3 (frontend)

Deployment architecture



Hardware stack

in house

- 2x hypervisors with ~20 VMs
- 1x high performance database server
- 2x dedicated storage server using
- 2x high density storage array (60 * 6TB => 300TB usable each)
- 3x nodes for a kafka+elasticsearch cluster

on Azure

- full object storage mirror
- full mirror of the database containing the graph
- workers for content indexing
- workers for download bundle preparation

Software development

classic FOSS development

- language: English
- development mailing list
  `https://sympa.inria.fr/sympa/info/swh-devel`
- IRC
  #swh-devel / FreeNode
- Forge
  `https://forge.softwareheritage.org`
- Git, tasks, code review, etc.

for more information

`https://www.softwareheritage.org/community/developers/`

Yes, now you can!

Wayback machine for source code

retrieve the source code as it was

Reference catalog

use intrinsic identifiers for software

Open science

deposit scientific software (via HAL)

Universal knowledge base

store the knowledge about source code

And much, much more is in store!

With your help?

The next steps

The Software Heritage Foundation

- independent
- long term mission
- multistakeholder

The community

- academia: Open Access, research
- industry: better software
- cultural heritage: all the software history

The mirror network

- resilience
- biodiversity

> *"Let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident."*
>
> *Thomas Jefferson*

Software Heritage

www.softwareheritage.org @swheritage

Everybody is concerned, everybody can help build

The Library of Alexandria of code

- recover the past
- structure the future

A CERN for Software

- build better software
  - for industry
  - for society as a whole

Q: do you archive *only* Free Software?

- We only crawl origins *meant* to host source code (e.g., forges)
- Most (~90%) of what we *actually* retrieve is textual content

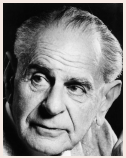
Our goal

Archive the entire Free Software Commons

- Large parts of what we retrieve is *already* Free Software, today
- Most of the rest *will become* Free Software in the long term
  - e.g., at copyright expiration

Reproducibility

Reproducibility is the key

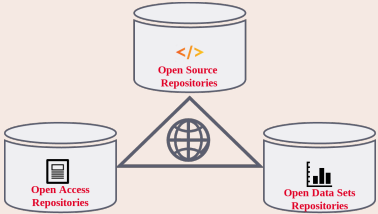


non-reproducible single occurrences are of no significance to science

Karl Popper, The Logic of Scientific Discovery, 1934

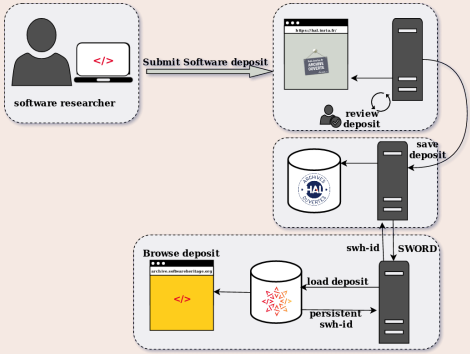
the *software source code* is special

- It embodies the logic of the data transformation.
- It must be included in strategies for *scientific knowledge preservation*.
- knowing the exact version of the software used is essential for reproducibility.



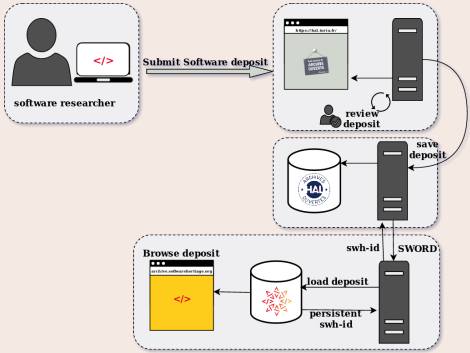
Deposit software in HAL

http://hal.inria.fr/hal-01738741



Deposit software in HAL

http://hal.inria.fr/hal-01738741

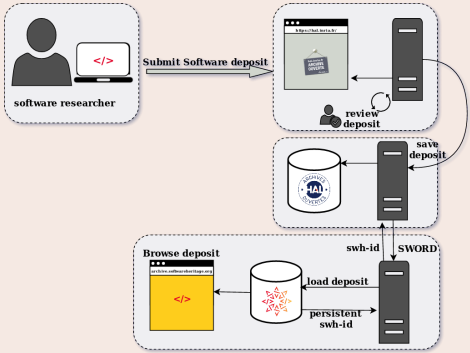


Generic mechanism:

- SWORD based
- review process
- versioning

Deposit software in HAL

http://hal.inria.fr/hal-01738741



Generic mechanism:

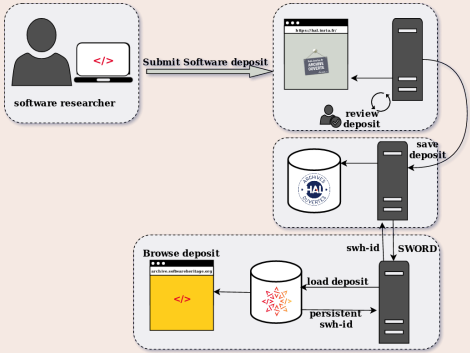
- SWORD based
- review process
- versioning

How to do it:

- today: deposit .zip or .tar.gz file (*guide*)
- tomorrow:
  - provide *SWH id* and metadata
  - include *metadata file* for automatic metadata extraction
  - …

Deposit software in HAL

http://hal.inria.fr/hal-01738741



Generic mechanism:

- SWORD based
- review process
- versioning

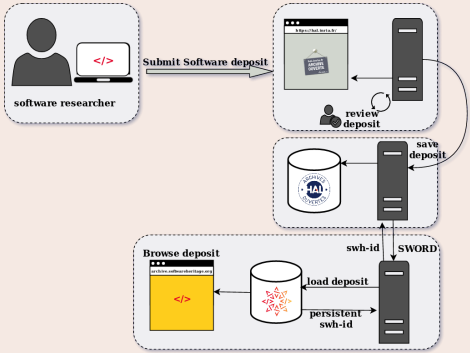
How to do it:

- today: deposit .zip or .tar.gz file (*guide*)
- tomorrow:
  - provide *SWH id* and metadata
  - include *metadata file* for automatic metadata extraction
  - …

September 2018: open to all on https://hal.archives-ouvertes.fr/

Deposit software in HAL

http://hal.inria.fr/hal-01738741



Generic mechanism:

- SWORD based
- review process
- versioning

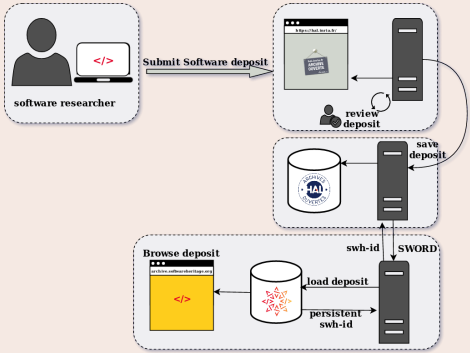
How to do it:

- today: deposit .zip or .tar.gz file (*guide*)
- tomorrow:
  - provide *SWH id* and metadata
  - include *metadata file* for automatic metadata extraction
  - …

September 2018: open to all on https://hal.archives-ouvertes.fr/

Deposit software in HAL

http://hal.inria.fr/hal-01738741



Generic mechanism:

- SWORD based
- review process
- versioning

How to do it:

- today: deposit .zip or .tar.gz file (*guide*)
- tomorrow:
  - provide *SWH id* and metadata
  - include *metadata file* for automatic metadata extraction
  - …

September 2018: open to all on https://hal.archives-ouvertes.fr/