Software Heritage
Preserving the Free Software Commons

Nicolas Dandrimont

Software Engineer
Software Heritage
nicolas@dandrimont.eu

06 june 2017
Café LoOPS
Bures-Sur-Yvette

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Free Software is everywhere

Software source code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs

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

Quake 2 source code (excerpt)

```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Net. queue in Linux (excerpt)

```
/*
 * SFB uses two B[l][n] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS  (1 << SFB_BUCKET_SHIFT) /* N bins per Level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS      (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
        u16             qlen; /* length of virtual queue */
        u16             p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

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

Our Software Commons

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. `https://en.wikipedia.org/wiki/Commons`

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

`https://en.wikipedia.org/wiki/Software_Commons`

Our Software Commons

Definition (Commons)

The commons is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. `https://en.wikipedia.org/wiki/Commons`

Definition (Software Commons)

The software commons consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons.* [...]

`https://en.wikipedia.org/wiki/Software_Commons`

Source code is *a precious part* of our commons

are we taking care of it?

Software is spread all around

Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to another over time

Software is spread all around

Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to another over time

Where is the place …

where we can find, track and search *all* source code?

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

Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

A wealth of software research on crucial issues…

- safety, security, test, verification, proof
- software engineering, software evolution
- big data, machine learning, empirical studies

Software lacks its own research infrastructure

Photo: ALMA(ESO/NAOJ/NRAO), R. Hills

A wealth of software research on crucial issues…

- safety, security, test, verification, proof
- software engineering, software evolution
- big data, machine learning, empirical studies

If you study the stars, you go to Atacama…

… where is the *very large telescope* of 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.

Our principles

Our principles

Open approach

- 100% Free 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

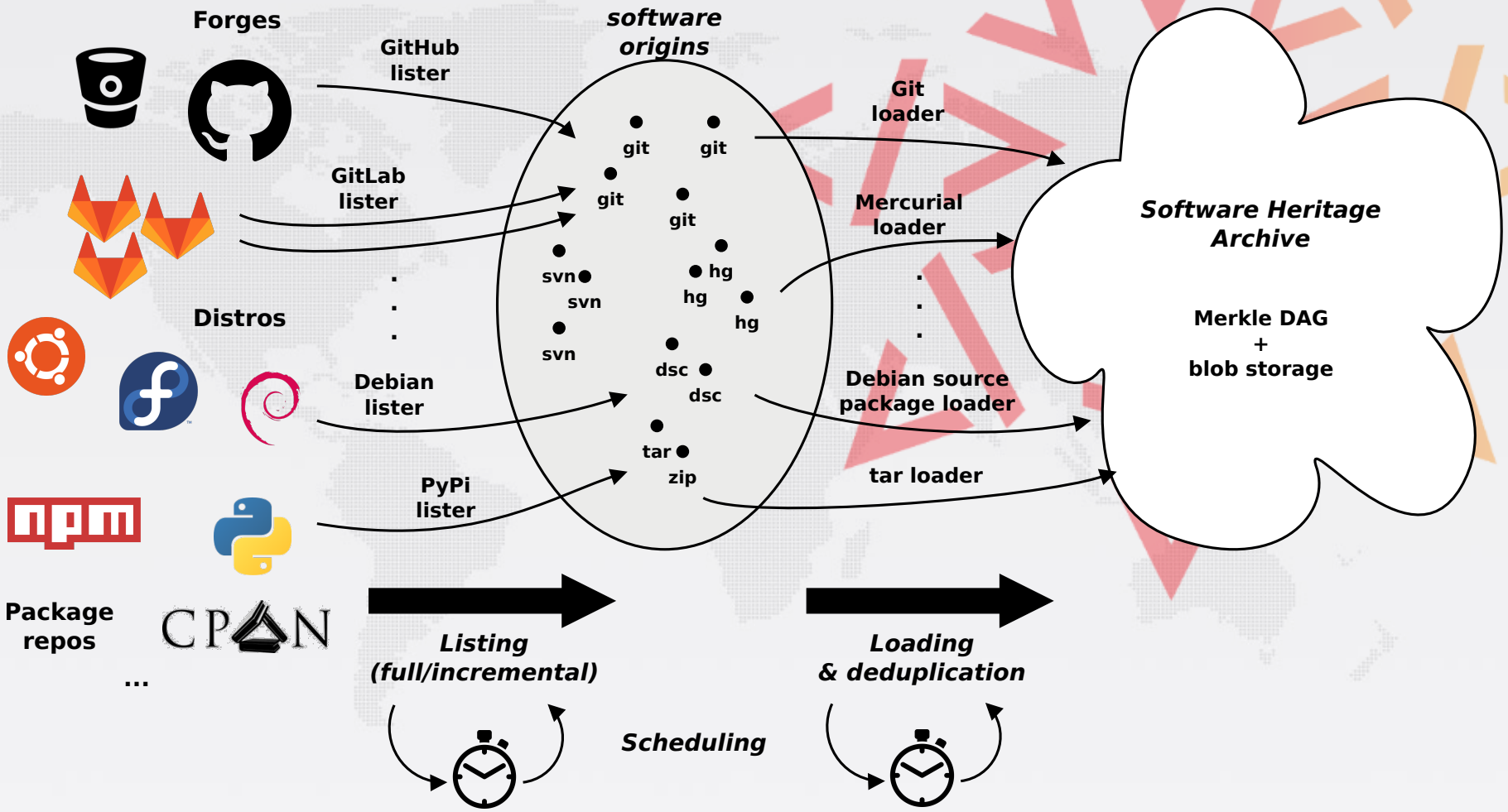
... in a VCS-/archive-agnostic canonical data model

We DON'T archive

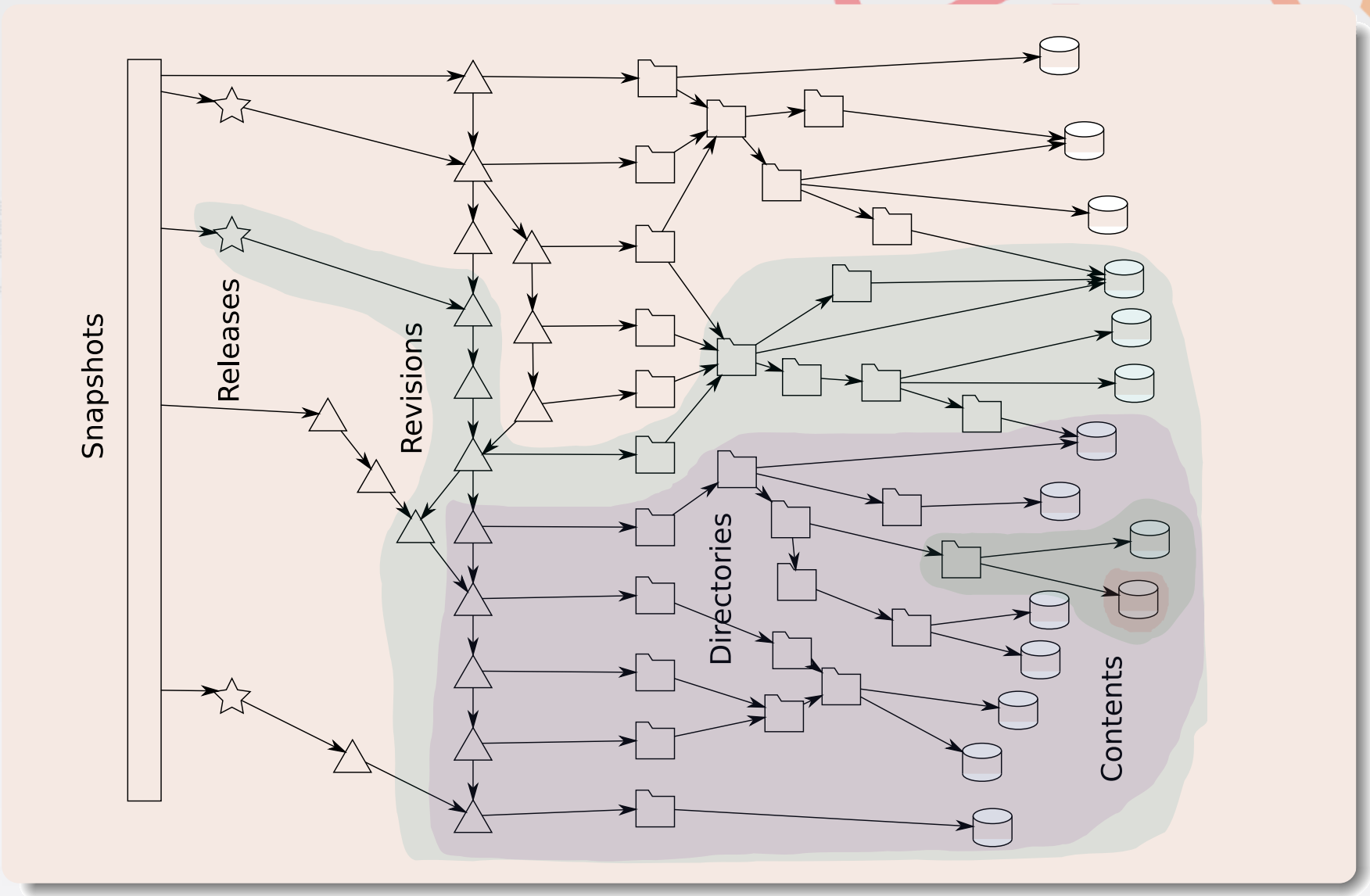
- homepages, wikis
- BTS/issues/code reviews/etc.
- mailing lists

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

Data flow

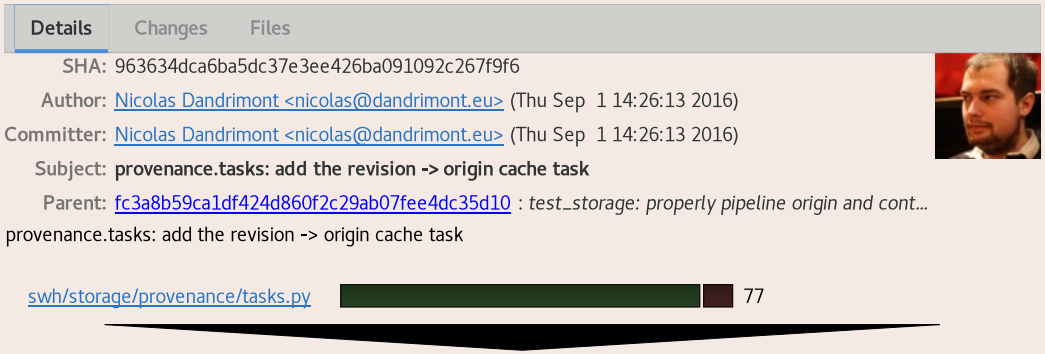

Forges
GitHub lister
GitLab lister
Distros
Debian lister
PyPi lister
Package repos
...
software origins
git
git
git
git
svn
svn
svn
hg
hg
hg
dsc
dsc
tar
zip
Git loader
Mercurial loader
Debian source package loader
tar loader
Software Heritage Archive
Merkle DAG
+
blob storage
Listing (full/incremental)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

The archive: a (giant) Merkle DAG



Example: a Software Heritage revision

Revisions



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

Note: most object kinds currently have Git-compatible identifiers

Archive coverage

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — processing (Archive Team & Google)
- Bitbucket — WIP

Archive coverage

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — processing (Archive Team & Google)
- Bitbucket — WIP

Some numbers

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)

Archive coverage

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — processing (Archive Team & Google)
- Bitbucket — WIP

Some numbers

150 TB blobs, 5 TB database (as a graph: 7 B nodes + 60 B edges)

The *richest* source code archive already, … and growing daily!

Software stack

3rd party

- Debian, Puppet
- PostgreSQL for metadata storage, with barman & pglogical
- Celery (RabbitMQ backend) for task scheduling
- Python3 and psycopg2 for the backend
- Flask and Bootstrap for Web stuff
- Phabricator

in house

- *ad hoc* object storage (to avoid imposing tech to mirrors)
- data model implementation, listers, loaders, scheduler
- ~50 Git repositories (~20 Python packages, ~10 Puppet modules)
- ~30 kSLOC Python / ~12 kSLOC SQL / ~4 kSLOC Puppet
- licence choice: GPLv3 (backend) / AGPLv3 (frontend)

Hardware stack

in house

- 2x hypervisors with ~20 VMs
- 2x high density storage array (60 * 6TB => 300TB usable)

on Azure

- full object storage mirror
- workers for content indexing

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/`

Roadmap

Features...

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) via Web API
  - (todo) via Web UI
- (todo) download: `wget` / `git clone` from the archive
- (todo) provenance information for all archived content
- (todo) full-text search on all archived source code files

Roadmap

Features...

- (done) lookup by content hash
- browsing: "wayback machine" for archived code
  - (done) via Web API
  - (todo) via Web UI
- (todo) download: `wget` / `git clone` from the archive
- (todo) provenance information for all archived content
- (todo) full-text search on all archived source code files

... and much more than one could possibly imagine

all the world's software development history in a single graph!

You can help!

Coding

- `forge.softwareheritage.org` — our own code

★★★	listers for unsupported forges, distros, pkg. managers
★★★	loaders for unsupported VCS, source package formats
★★	Web UI: eye candy wrapper around the Web API

You can help!

Coding

- `forge.softwareheritage.org` — our own code

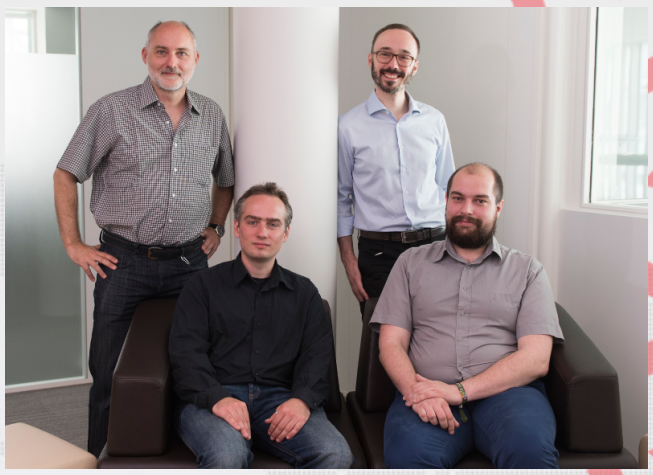
★★★	listers for unsupported forges, distros, pkg. managers
★★★	loaders for unsupported VCS, source package formats
★★	Web UI: eye candy wrapper around the Web API

Community

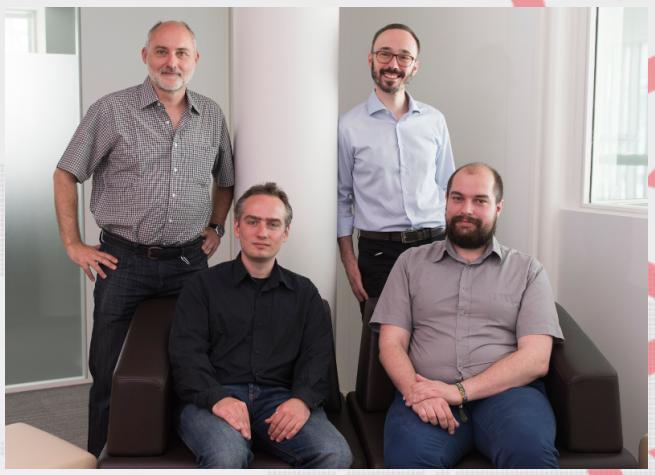
★★	spread the news, help us with long-term sustainability
★★★	document endangered source code

`wiki.softwareheritage.org/index.php?title=Suggestion_box`

The Software Heritage community



The Software Heritage community

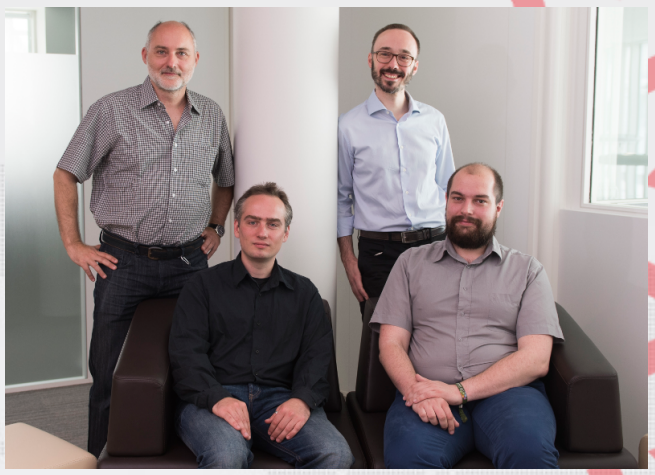


Inria as initiator

Inria
INVENTEURS DU MONDE NUMÉRIQUE

- .fr national computer science research entity
- strong Free Software culture

The Software Heritage community



Inria as initiator

Inria
INVENTEURS DU MONDE NUMÉRIQUE

- .fr national computer science research entity
- strong Free Software culture

Early Sponsors and Supporters

Société Générale, Microsoft, Huawei, Nokia, DANS, Univ. Bologna, ACM, Creative Commons, Eclipse, Engineering, FSF, Gandi, GitHub, IEEE, OIN, OSI, OW2, Software Freedom Conservancy, SFLC, The Document Foundation, …

Conclusion

Software Heritage is

- a *reference archive* of *all* Free Software ever written
- a unique *complement* for *development platforms*
- an international, open, nonprofit, *mutualized infrastructure*
- at the service of our community, at the service of society

Come in, we're open!

`wiki.softwareheritage.org` — *leads*
`forge.softwareheritage.org` — *our own code*
`www.softwareheritage.org` — *sponsoring*

Questions?

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

Q: how about SHA1 collisions?

```
create domain sha1 as bytea
  check (length(value) = 20);
create domain sha1_git as bytea
  check (length(value) = 20);
create domain sha256 as bytea
  check (length(value) = 32);

create table content (
  sha1       sha1 primary key,
  sha1_git   sha1_git not null,
  sha256     sha256 not null,
  length     bigint not null,
  ctime      timestamptz not null default now(),
  status     content_status not null default 'visible',
  object_id  bigserial
);

create unique index on content(sha1_git);
create unique index on content(sha256);
```