Software Heritage
Building the Universal Software Archive

Nicolas Dandrimont

July 4th, 2016
Outline

1. The need for Software Preservation
   - Software all around us
   - Software is Fragile

2. The Software Heritage project
   - Our mission
   - Our vision

3. Software Heritage in depth
   - Our current work
   - Our roadmap

4. How to contribute to Software Heritage
   - Developer information
   - Sponsoring opportunities

5. Conclusion
Software is Pervasive

At the heart of *our society*

- communication, entertainment
- administration, finance
- health, energy, transportation
- education, research, politics
- …
So/f_tware is Pervasive

At the heart of our society

- communication, entertainment
- administration, finance
- health, energy, transportation
- education, research, politics
- ...

At the heart of technology

- house appliances $\approx 10M$ SLOC
- phones $\approx 20M$ SLOC, cars $\approx 100M$ SLOC
- Internet of things, …
Information is a main pillar of our modern societies.

Absent an ability to correctly interpret digital information, we are left with [...] "rotting bits" [...] of no value.

Vinton G. Cerf IEEE 2011
Software is Knowledge

Key mediator for accessing all information (c) Banksy

Information is a main pillar of our modern societies.

Absent an ability to correctly interpret digital information, we are left with […] "rotting bits" […] of no value.

Vinton G. Cerf IEEE 2011

Software is an essential component of modern scientific research

[…] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)
Software is Knowledge

Key mediator for accessing all information (c) Banksy

Information is a main pillar of our modern societies. Absent an ability to correctly interpret digital information, we are left with [...] "rotting bits" [...] of no value.

Vinton G. Cerf IEEE 2011

Software is an essential component of modern scientific research

[...] the vast majority describe experimental methods or software that have become essential in their fields.

Top 100 papers (Nature, October 2014)

Bottomline: Software embodies our Knowledge and Cultural Heritage

It must be collected, preserved, referenced and made accessible!
So/ftware is Fragile

Bits rot, hosters shut down
- Have you tested your backups recently? How about git fsck?
- Gitorious
- Google Code

Software is scattered all around
GitHub, GitLab, BitBucket, SourceForge, alioth, …
… your personal home page, …

No uniformity or stability whatsoever
Software migrates from hosters to hosters, URIs aren’t perennial
Outline

1. The need for Software Preservation
   - Software all around us
   - Software is Fragile

2. The Software Heritage project
   - Our mission
   - Our vision

3. Software Heritage in depth
   - Our current work
   - Our roadmap

4. How to contribute to Software Heritage
   - Developer information
   - Sponsoring opportunities

5. Conclusion
Our mission

Collect, organise, preserve and share all the software source code that lies at the heart of our culture and our society.

https://www.softwareheritage.org/
“Programs must be written for people to read, and only incidentally for machines to execute.” Harold Abelson, *Structure and Interpretation of Computer Programs*

**Distinguishing features**

- executable and *human readable* knowledge (an *all time new*)
  - even hardware is… software! (VHDL, FPGA, …)
  - *text files are forever*
- naturally *evolves* over time
  - the *development history* is key to its *understanding*
- complex: large *web of dependencies*, millions of SLOCs

**In a word**

- software is *not just another* sequence of bits
- a software archive is *not just another* digital archive
We are working on the foundations

One infrastructure to build them all

- Mankind's memory
- Long term preservation
- Unique reference
- Software Wikipedia

Cultural Heritage

- Reference repository
- Provenance
- Certification
- Security

Industry

- Reproducibility
- Traceability
- Open Access
- Software studies

Research

- Universal SourceBook
- Reference examples
- Enriched source code
- Code documentation

Education

Software Heritage
Preserving the world’s software heritage

A structured archive of all of the world’s software

- preserve humanity’s technological and scientific knowledge
- enable continued access to all digital documents and information
- building block for thematic portals and collections
Better software for industry

A unique reference catalog of all industrial software components

- ensures long term preservation of critical software
- eases vulnerability tracking for more secure software
- simplifies traceability for better software integration
Supporting more accessible and reproducible science

A global library referencing all software used in all research fields

- completes the infrastructure for Open Access in science
- provides intrinsic persistent identifiers needed for scientific reproducibility
- enables large scale, verifiable software studies
Meet the team

- Roberto Di Cosmo, CEO
- Stefano Zacchirolì, CTO
- Antoine Dumont and Nicolas Dandrimont, Engineers
- Jordi Bertran de Balanda and Quentin Campos, Interns
- Guillaume Rousseau, Visiting Scientist
Our stack

Hardware

- Hosted by Inria
- One big hypervisor with a dozen virtual machines
- One high density storage array (60 * 6TB hard drives => 300TB usable)
- Another copy of the data in another server room; logical leader/follower mirroring
- Soon to enable a mirror network to duplicate our contents

Software

- Debian for all our machines
- PostgreSQL for metadata storage
- Python3 and psycopg2 for the backend
- Flask for the web apps
- RabbitMQ for task scheduling
Our values are those of Debian

100% FOSS licenses
- GPLv3 for the backend code
- AGPLv3 for the frontend
- Apache2 for the Puppet manifests

Community-minded
We encourage bug reports and code contributions from everyone interested in pursuing our software preservation mission.
Our forge opens today

https://forge.softwareheritage.org/

We’ve timed the opening of our forge for DebConf, as a thank you for what the community has given to us.
Current data sources

Ingest all the software

- all the "non-fork" GitHub repositories
- all the Debian packages from snapshot.debian.org
- the GNU project FTP archive

Preserve all the software

- Google Code
- Gitorious
The structure of the archive

On-disk storage

- flat file storage for contents
- postgres database for the metadata

Data model: one big Merkle DAG, inspired by the git model

- Origins (= repositories)
- Occurrences (= branches)
- Releases (= tags)
- Revisions (= commits)
- Directories (= trees)
- Contents (= blobs)
So/f_tware Heritage in numbers

Volume
- 120TB used by (gzipped) files on disk
- 3.1TB PostgreSQL database for the metadata

Counts
- 2.7 billion files
- 2.2 billion directories
- 600 million revisions
- 12 million people
- 5 million releases
Software Heritage in numbers

Volume
- 120TB used by (gzipped) files on disk
- 3.1TB PostgreSQL database for the metadata

Counts
- 2.7 billion files
- 2.2 billion directories
- 600 million revisions
- 12 million people
- 5 million releases

By far, the biggest DVCS tree in existence
Planned features...

- **lookup** by hashes for contents (done)
- **provenance information** for all the content
- **browsing**: wayback machine for software source code
- **full text search**: dive into the Software Heritage archive
- **download**: git clone from Software Heritage
The road ahead

Planned features…

- **lookup** by hashes for contents (done)
- **provenance information** for all the content
- **browsing**: wayback machine for software source code
- **full text search**: dive into the Software Heritage archive
- **download**: git clone from Software Heritage

… and many more one could imagine

all the world’s software development history in a single graph!  
*that makes a 3.1TB database already…*
Technical Challenges for data sources

Handling the backlog
- ingesting all the data saved one-shot

Staying current
- new repositories and commits on GitHub
- sync with snapshot.debian.org
- ...

*We need reliable, standardised event feeds*

Expanding the archive
- discover and classify *all* the software sources
- importers for other VCSs (SVN, Hg, …)

*Wonderful playground if you have time to help!*

Nicolas Dandrimont
1. The need for Software Preservation
   - Software all around us
   - Software is Fragile

2. The Software Heritage project
   - Our mission
   - Our vision

3. Software Heritage in depth
   - Our current work
   - Our roadmap

4. How to contribute to Software Heritage
   - Developer information
   - Sponsoring opportunities

5. Conclusion
Join our community of developers

Our forge is open at
https://forge.softwareheritage.org/

Subscribe to our mailing list
https://deb.li/swhdevelml

Have a look at our wiki
https://wiki.softwareheritage.org/
Inria as initiator

French national institute for research in Computer Science.
- Contributed to the birth of W3C
- 4500 people, many prestigious scientists
- In the news: Freak and Logjam TLS vulnerabilities

Inria is *fully supporting* the bootstrap phase of Software Heritage.

More info for sponsors on our website:

https://www.softwareheritage.org/support/sponsors/
Outline

1. The need for Software Preservation
   - Software all around us
   - Software is Fragile

2. The Software Heritage project
   - Our mission
   - Our vision

3. Software Heritage in depth
   - Our current work
   - Our roadmap

4. How to contribute to Software Heritage
   - Developer information
   - Sponsoring opportunities

5. Conclusion
Conclusion

Software Heritage is

- a revolutionary *reference archive* of all software ever written
- a unique *complement* for *development platforms* like GitHub
- an international, open, nonprofit, *mutualized infrastructure*

*ready to work with you!*
Conclusion

Software Heritage is

- a revolutionary *reference archive* of all software ever written
- a unique *complement* for *development platforms* like GitHub
- an international, open, nonprofit, *mutualized infrastructure*

*ready to work with you!*

Questions?

Contact us

*email:* olasd@softwareheritage.org / info@softwareheritage.org
*website:* https://www.softwareheritage.org