Software Heritage
l’archive universelle du logiciel libre

Nicolas Dandrimont & David Douard

Engineers, Software Heritage – Inria

16 March 2024
Assemblée Générale de l’April
Sorbonne Université
“Programs must be written for people to read, and only incidentally for machines to execute.”

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985
Software is built from *Source Code*

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)

```
P63SP0T3  CA  BIT6  # IS THE LR ANTENNA IN POSITION 1 YET
          EXTEND RAND  CHAN33  
          EXTEND BZF  P63SP0T4  # BRANCH IF ANTENNA ALREADY IN POSITION 1
          CAF  CODE508  # ASTRONAUT: PLEASE CRANK THE
          TC  BANKCALL  # SILLY THING AROUND
          CADR  GOPERF1
          TCF  GOTOPO0H  # TERMINATE
          TCF  P63SP0T3  # PROCEED SEE IF HE'S LYING
P63SP0T4  TC  BANKCALL  # ENTER INITIALIZE LANDING RADAR
          CADR  SETPOS1
          TC  POSTJUMP  # OFF TO SEE THE WIZARD ...
          CADR  BURNBABY
```
Software is built from **Source Code**

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)

<table>
<thead>
<tr>
<th>P635POT3</th>
<th>CA</th>
<th>BIT6</th>
<th># IS THE LR ANTENNA IN POSITION 1 YET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EXTEND</td>
<td>CHAN33</td>
<td># BRANCH IF ANTENNA ALREADY IN POSITION 1</td>
</tr>
<tr>
<td>BZF</td>
<td>P635POT4</td>
<td>CAF</td>
<td>CODE500 # ASTRONAUT: PLEASE CRANK THE</td>
</tr>
<tr>
<td>TC</td>
<td>BANCALL</td>
<td>#</td>
<td>SILLY THING AROUND</td>
</tr>
<tr>
<td>CADR</td>
<td>GOPERF1</td>
<td>TCF</td>
<td>GOTOPOOH # TERMINATE</td>
</tr>
<tr>
<td>TCF</td>
<td>P635POT3</td>
<td># PROCEED</td>
<td>SEE IF HE'S LYING</td>
</tr>
</tbody>
</table>

---

### Parcoursup source code (excerpt)

```java
public class AlgoOrdreAppel {

    /* la boucle principale de calcul des ordres d'appels.
     * Renvoie une exception en cas de problème. */
    public static AlgoOrdreAppel Sortie calculerOrdresAppels(AlgoOrdreAppelEntree data) throws VerificationException {
        VerificationEntree algoOrdreAppelVerifier(data);

        AlgoOrdreAppelSortie resultat = new AlgoOrdreAppelSortie();

        for (GroupeClassement gp : data.groupeClassements) {
            resultat.ordreAppel.put(gp.csClasse, gp.calculerOrdreAppel());
        }

        return resultat;
    }

    private AlgoOrdreAppel() {} 
}
```

---

Len Shustek, Computer History Museum 2006

“Source code provides a view into the mind of the designer.”

---

N. Dandrimont - D. Douard (CC-BY 4.0)
Software is built from *Source Code*

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

Apollo 11 source code (excerpt)

```
P63SP0T3       CA BIT6 # IS THE LR ANTENNA IN POSITION 1 YET
               EXTEND
               RAND CHA9N33
               EXTEND
               BZF P63SP0T4 # BRANCH IF ANTENNA ALREADY IN POSITION 1
               CAF CODE508 # ASTRONAUT: PLEASE CRANK THE
               TC BANKCALL # SILLY THING AROUND
               CADR GOPERF1
               TCF GOTOPO0M # TERMINATE
               TCF P63SP0T3 # PROCEED SEE IF HE'S LYING
P63SP0T4       TC BANKCALL # ENTER INITIALIZE LANDING RADAR
               CADR SETPOS1
               TC POSTJUMP # OFF TO SEE THE WIZARD ...
```

Parcoursup source code (excerpt)

```
public class AlgoOrdreAppel {
    /* la boucle principale de calcul des ordres d'appels.
       Renvoie une exception en cas de probleme. */
    public static AlgoOrdreAppelSortie calculerOrdresAppels(AlgoOrdreAppelEntree data) throws VerificationException {
        VerificationEntreeAlgoOrdreAppel. verifier(data);
        AlgoOrdreAppelSortie resultat = new AlgoOrdreAppelSortie();
        /* calcul de l'ordre d'appel de chaque groupe de classement */
        for (GroupeClassement ga : data.groupesClassements) {
            resultat.ordresAppel.put(ga.getIdCode, ga.calculerOrdreAppel());
        }
        /* verification avant retour des resultats */
        new VerificationsResultatsAlgoOrdreAppel(). verifier(data, resultat);
        return resultat;
    }

    private AlgoOrdreAppel() {
    }
}
```

Len Shustek, Computer History Museum

“Source code provides a view into the mind of the designer.”
Experts call for greater recognition of software source code as heritage for sustainable development

6 November 2018

UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …
Software source code as a key asset of Humankind

UNESCO, Inria, Software Heritage invite
40 international experts meet in Paris …

The call is published on February 2019
UNESCO, Inria, Software Heritage invite 40 international experts meet in Paris …

Experts call for greater recognition of software source code as heritage for sustainable development

The call is published on February 2019

“Recognise software source code as a fundamental enabler in all aspects of human endeavour"
Software source code is fragile

Endangered source code ...

- **link rot**: projects are created, moved around, removed
- **data rot**: physical media with legacy software decay
- **platform consolidation** endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250,000
  - 2022 GitLab.com: remove inactive projects?

Bottomline: we need a global, long term effort to build a universal archive of all software source code make it resilient and make it sustainable

N. Dandrimont - D. Douard (CC-BY 4.0)
Software source code is fragile

Endangered source code …

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250,000
  - 2022 GitLab.com: remove inactive projects?

… is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind
Software source code is fragile

Endangered source code …

- *link rot*: projects are created, moved around, removed
- *data rot*: physical media with legacy software decay
- *platform consolidation* endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: remove inactive projects?

… is endangered knowledge!

broken links and missing pieces in the *web of knowledge* of humankind

Bottomline: we need a global, long term effort

  to build a *universal archive of all software source code*
  
  make it *resilient*
  
  and make it *sustainable*
Software Heritage in a nutshell

Unveiled in 2016

Collect, preserve and share all software source code

Preserving our heritage, enabling better software and better science for all
Software Heritage in a nutshell

Unveiled in 2016

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
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
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
Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared
Today: a *universal* software archive, as a shared infrastructure

One infrastructure
open and shared

Software Heritage

The largest archive ever built

<table>
<thead>
<tr>
<th>Category</th>
<th>Figures as of January 25 2024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source files</td>
<td>17,798,218,376</td>
</tr>
<tr>
<td>Commits</td>
<td>3,802,143,973</td>
</tr>
<tr>
<td>Projects</td>
<td>278,187,495</td>
</tr>
<tr>
<td>Directories</td>
<td>14,364,868,206</td>
</tr>
<tr>
<td>Authors</td>
<td>69,923,710</td>
</tr>
<tr>
<td>Releases</td>
<td>82,196,102</td>
</tr>
</tbody>
</table>
Today: a *universal* software archive, as a shared infrastructure

One infrastructure open and shared

The largest archive ever built

Source files: 17,798,218,376
Commits: 3,802,143,973
Projects: 278,187,495
Directories: 14,364,868,206
Authors: 69,923,710
Releases: 82,196,102

Software Heritage

N. Dandrimont - D. Douard (CC-BY 4.0)

Software Heritage softwareheritage.org 16/03/2024 6 / 11
An operational, evolving infrastructure

Harvest and archive

[Diagram showing software heritage infrastructure with various package repos and forges leading to software origins and listing (full/incremental).]

Reference (35 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers

Now in SPDX 2.2, Wikidata, ISO is coming

Global development history permanently archived in a uniform data model

Over 17 billion unique source files from over 270 million software projects

~1.5PB (compressed) blobs, ~35 B nodes, ~500 B edges

Significant research challenges to explore it efficiently

N. Dandrimont - D. Douard (CC-BY 4.0)
An operational, evolving infrastructure

Harvest and archive

Software Heritage Archive
Merkle DAG + blob storage

Loading & deduplication

Origins
Snapshots
Revisions
Revisions directories
Contents

Forges
GitHub
Git
GitLab
Git

Distros
Debian
Git

Lister
PyPi lister

Package repos
CPAN
Debian lister

Scheduling

N. Dandrimont - D. Douard (CC-BY 4.0)
An operational, evolving infrastructure

Harvest and archive

- save.softwareheritage.org
- deposit.softwareheritage.org
An operational, evolving infrastructure

Harvest and archive

- save.softwareheritage.org
- deposit.softwareheritage.org

Reference (35 billion SWHIDs)

Intrinsic, decentralised, cryptographically strong identifiers

Now in SPDX 2.2, Wikidata, ISO is coming
An operational, evolving infrastructure

**Harvest and archive**

- Git loader
- Mercurial loader
- Debian source package loader
- PyPi source package loader
- Software Heritage Archive Merkle DAG + blob storage
- Loading & deduplication

- Forgés
- GitLab lister
- Git lister
- Mercurial lister
- Debian lister
- PyPi lister
- Distros

**Package repos**

- GitHub lister
- GitLab lister
- Debian lister
- PyPi lister
- Distros
- Forge lister
- PyPi lister
- Debian lister

**Scheduling**

- Listing (full/incremental)
- Tar origins
- Snapshots
- Releases
- Revisions
- Directories
- Contents

**Reference (35 billion SWHIDs)**

Intrinsic, decentralised, cryptographically strong identifiers

- **save.softwareheritage.org**
- **deposit.softwareheritage.org**

Global development history permanently archived in a uniform data model

- **over 17 billion** unique source files from over **270 million** software projects
- **~1.5PB** (compressed) blobs, **~35 B nodes, ~500 B edges**

Significant research challenges to explore it efficiently

**Now in SPDX 2.2, Wikidata, ISO is coming**
A walkthrough

**General**
- Browse the archive, get and use SWHIDs, e.g. Apollo 11 excerpt, Parcoursup excerpt
- Trigger archival with the browser extension or webhook forge integration

**Open Science**
- Curated deposit via HAL, e.g.: LinBox, SLALOM, Givaro, SumGra, Coq proof, …
- Cite software with the biblatex-software style, e.g.: article from IPOL

**History of software: rescuing landmark legacy software**

see SWHAP process, Software Stories, and SWHAP Days 2022

**Public code**

Archived source code from code.gouv.fr
An international, non profit initiative for the long term

Sharing the vision

United Nations Educational, Scientific and Cultural Organization

And many more ...

www.softwareheritage.org/support/testimonials

N. Dandrimont - D. Douard (CC-BY 4.0)
An international, non profit initiative for the long term

Sharing the vision

And many more ...

www.softwareheritage.org/support/testimonials

Donors, members, sponsors

Diamond sponsor
Platinum sponsors
Gold sponsors
Silver sponsors
Bronze sponsors

we are all concerned, anyone can join and help

N. Dandrimont - D. Douard (CC-BY 4.0)
A growing and active community

Core Team

All together, 2024 Summit

Ambassadors

N. Dandrimont - D. Douard (CC-BY 4.0)
A call to realize a grand vision

Bring together academia, industry, civil society and governments to build

"a global infrastructure for open and better software at the service of humankind"

Software Heritage

www.softwareheritage.org
@swheritage@mstdn.social

We’re hiring!
- sysadmin
- big data engineer
- backend developer

Spread the word
- become an ambassador
- advocate for SWH in your communities