So/f_tware Heritage

Roberto Di Cosmo
Directeur de Software Heritage

roberto@dicosmo.org

3 Avril 2017
Conférence de Presse
Software is everywhere
“Programs must be written for people to read, and only incidentally for machines to execute.” Harold Abelson

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

Hello World

<table>
<thead>
<tr>
<th>Program (excerpt of binary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6: 55</td>
</tr>
<tr>
<td>4004e7: 48 89 e5</td>
</tr>
<tr>
<td>4004ea: bf 84 05 40 00</td>
</tr>
<tr>
<td>4004ef: b8 00 00 00 00</td>
</tr>
<tr>
<td>4004f4: e8 c7 fe ff ff</td>
</tr>
<tr>
<td>4004f9: 90</td>
</tr>
<tr>
<td>4004fa: 5d</td>
</tr>
<tr>
<td>4004fb: c3</td>
</tr>
</tbody>
</table>
“Programs must be written for people to read, and only incidentally for machines to execute.” Harold Abelson

---

**Program (source code)**

```c
/* Hello World program */

#include<stdio.h>

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

**Program (excerpt of binary)**

<table>
<thead>
<tr>
<th>Address</th>
<th>Hex Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004e6</td>
<td>55</td>
</tr>
<tr>
<td>4004e7</td>
<td>48 89 e5</td>
</tr>
<tr>
<td>4004ea</td>
<td>bf 84 05 40 00</td>
</tr>
<tr>
<td>4004ef</td>
<td>b8 00 00 00 00</td>
</tr>
<tr>
<td>4004f4</td>
<td>e8 c7 fe ff ff</td>
</tr>
<tr>
<td>4004f9</td>
<td>90</td>
</tr>
<tr>
<td>4004fa</td>
<td>5d</td>
</tr>
<tr>
<td>4004fb</td>
<td>c3</td>
</tr>
</tbody>
</table>
“Programs must be written for people to read, and only incidentally for machines to execute.” Harold Abelson

**Program (source code)**

```c
/* Hello World program */
#include<stdio.h>
void main()
{
    printf("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
```

Software *source code* is *special, precious* knowledge
Free and Open Source software

Source code *first*

- everybody *can use, study, adapt, and distribute* the software and its *source code*

All industry is adopting the Open Source approach

- find, reuse, adapt, for faster development

Even old opponents are now becoming active adopters

- "*Microsoft loves Linux*"
  
  *Satya Nadella, October 2014*
Free and Open Source software

Source code *first*

ev[ery][b]ody can *use, study, adapt, and distribute* the software and its source code

All industry is adopting the Open Source approach

find, reuse, adapt, for faster development

Even old opponents are now becoming active adopters

*“Microsoft loves Linux”*

*Satya Nadella, October 2014*

Software source code is special, precious knowledge

are we taking care of it?
Our source code is spread all around.
Source code is fragile

- damage
- disaster
- malicious
- obsolete
- attack
- media
- aging
- dependencies
- reference
- deletion
- dangling
- wear
- corruption
- encryption
- format
Software is missing its own Research Infrastructure

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

Open approach
open source, transparency

In for the long haul
non profit, replication

Cultural Heritage | Industry | Research | Education

Software Heritage
Our principles

Open approach
open source, transparency

In for the long haul
non profit, replication

Software Heritage
A complex task

Software Heritage
A complex task

Software Heritage

Collect
- discover
  - sources
- harvest
  - protocols
- ingest
  - VCS
  - data models
A complex task

Software Heritage

Collect
- discover
  - sources
- harvest
  - protocols
- ingest
  - VCS
  - data models

Organise and Preserve
- enrich
  - metadata
- analyze
- replicate
  - locations
  - technologies
  - stakeholders
A complex task

Software Heritage

Collect
- discover
  - sources
- harvest
  - protocols
- ingest
  - VCS
  - data models

Organise and Preserve
- enrich
  - metadata
- analyze
- replicate
  - locations
  - technologies
  - stakeholders

Share
- download
- browse
- search
  - code
  - history
- watch
  - trends
June 30th 2016 (public announcement)

The archive

- 21 million repositories ingested
- 500 million commits
- 2.5 billion unique source files

Sponsors and supporters

Microsoft, DANS (Royal Academy of the Arts and Sciences), ACM, Creative Commons, Eclipse, FSF, OSI, GitHub, IEEE, OW2, SFC, SFLC, The Document Foundation, The Linux Foundation.
Today

The archive

- GitHub, Debian, GNU
- Gitorious, Google Code, ...
- first open API
Today

The archive

- GitHub, Debian, GNU
- Gitorious, Google Code, …
- first open API

Software Heritage for society as a whole

- preservation of knowledge embedded in software
- access to the knowledge embedded in software

Roberto Di Cosmo
The archive

- GitHub, Debian, GNU
- Gitorious, Google Code, …
- first open API

Software Heritage for society as a whole

- preservation of knowledge embedded in software
- access to the knowledge embedded in software

Sponsors and supporters