Software Heritage
The universal source code archive

Roberto Di Cosmo, Morane Guenpeter
Director, Software Heritage
Computer Science full professor, Inria and IRIF
roberto@dicosmo.org

July 11th, 2018
Outline

1. Software Heritage

2. Relevance for research software publishing
Collect, preserve and share the source code of all the software

Preserving our heritage, enabling better software and better science for all

Reference catalog

find and reference all the source code

Universal archive

preserve all the source code

Research infrastructure

enable analysis of all the source code
Our principles

- Open approach
  - open source
  - transparency

- In for the long haul
  - non profit, mirrors
  - intrinsic identifiers

- Exhaustive
  - all software
  - open to all communities

Software Heritage

<table>
<thead>
<tr>
<th>Source files</th>
<th>Commits</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,536,067,027</td>
<td>1,024,675,748</td>
<td>83,801,775</td>
</tr>
</tbody>
</table>
Growing Support

Raising awareness: landmark agreement, 3/4/2017; grand opening, 7/6/2018

Sharing the vision

Sponsoring our work

Roberto Di Cosmo, Morane Gruenpeter
Outline

1. Software Heritage

2. Relevance for research software publishing
Zoom on the collection phase

Much more than an archive!

- GitHub
- Debian, GNU
- Gitorious, Google Code
- WIP: Bitbucket, FusionForge, GitLab.com
- *add your own plugins!*

Important properties

- **mission:** exhaustive and up to date collection of source code, specifically
- **strategy:** automatic harvesting + *deposit* from *selected* sources

The *richest* source code archive already, … and growing daily!
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.
The research software (deposit) use case

Deposit software in HAL

[Diagram showing the process of software deposit 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
  - ...

Accessible on the INRIA portal at http://hal.inria.fr
Identifying and retrieving source code


- provide integrity guarantees
- all software and VCS (not just git or GitHub)
- use for identifying a precise version of source code
- learn more in the forthcoming iPres 2018 paper

e.g: swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa

"Wayback-machine-style" identifiers

- point to software origins
- expose the SWH crawling history

use when no precise version is known
Demo Time!

- example deposits in HAL
- example use of https://archive.softwareheritage.org
"Our Parmap.parmap and Parmap.parfold functions may be used to seamlessly ..."
Selected unique benefits

All features of Software Heritage *for free*

- **intrinsic IDs** (integrity, not dependent on resolvers!), browse, download (now)
- metadata, licenses, provenance analysis (plagiarism detection), classification (wip)
- and many more (powerful connections with SE and Industry)

Coverage and uniformity

- **one** archive for **all** domains (industry included)
- you can reference **any** software, not just the deposited one
  
  *(thanks D. Katz for pointing this out)*

- **git-compatible** identifiers greatly simplify workflows

Sustainability

- **one** infrastructure  
  
  *independent* non profit foundation  
  
  *worldwide* mirrors  

*there are more!*
Breaking news

Operational adoption

**June 7th**  swMath.org points into SWH for the source code see "Code" link in, e.g. http://swmath.org/software/7116

**ongoing**  OpenAire detects source code links in articles, resolves them to SWH

**September**  HAL opens the software deposit doors on all portals

Institutional adoption

**July 4th**  Software Heritage is part of the french National Plan for Open Science

---

**Organiser**

- Soutenir la Research data alliance (RDA) et créer le chapitre français de l’alliance (RDA France).
- Soutenir Software heritage, la bibliothèque des codes sources
<table>
<thead>
<tr>
<th><strong>Come in, we’re open!</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>Adopt and reuse</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>deposit/reference research software in SWH</td>
</tr>
<tr>
<td>become a SWH mirror</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Standardise</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>metadata</td>
</tr>
<tr>
<td>identifiers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Support</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>sponsoring / partnership</td>
</tr>
<tr>
<td>donations</td>
</tr>
<tr>
<td>our own code</td>
</tr>
</tbody>
</table>
Outline

1. Intrinsic PID
2. Our role in the publication workflow
3. The Metadata challenge
4. Collection strategies
Our challenge in the PID arena

Long term
Identifiers must be there for the long term

No middle man
Identifiers must be meaningful even if resolvers go away

Integrity, not just naming
Identifier must ensure that the retrieved object is the intended one

Uniqueness by design
one name identifies a single object, and each object has only one name
Exploring the PID landscape

A lot of options out there…
URL, URI, PURL, URN, ARK, DOI, …

… some are widely used
- articles
- data
- even software artefacts!

We can get no satisfaction of all the key criteria
we adopted something radically different
Intrinsic identifiers in Software Heritage

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

Combination of
tree
hash function

Classical cryptographic construction
fast, parallel signature of large data structures, built-in deduplication
- satisfies all three criteria
- widely used in industry (e.g., Git, nix, blockchains, IPFS, …)
Back to basics: DIOs vs. IDOs

**DIO (digital identifier of an object)**
- digital identifiers for traditional (non digital) objects
  - epistemic complications (manifestations, versions, locations, etc.)
  - significant governance issues, …

**IDO (identifier of a digital object)**
- (digital) identifier for digital objects
  - much simpler to build/handle
  - can (and must) be intrinsic

**Separation of concerns**
- yes, we **need both** DIOs and IDOs
- no, we **must not mistake** DIOs for IDOs (and viceversa)
Example: links to software source code in an article

Leveraging the Software Heritage universal archive:

set of files  \texttt{swh:1:dir:06741c8c37c5a384083082b99f4c5ad94cd0cd1f}
  id of directory object listing all the files in a project (at a given time)

revision  \texttt{swh:1:rev:7598fb94d59178d65bd8d2892c19356290f5d4e3}
  id of commit object which a tree and (a pointer to) the history

Full specification available online:
\url{https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html}

metadata  this \emph{will} involve some form of DIO
  and we get all the complications back
Our role: handle all the software source code

At the end of the process

Explicit deposit, coordinated with the publisher

- store the final source code (no garbage)
- store only public source code
- N.B.: no embargo or access control (yet)

During the review

Access to the largest available source code base

- provenance, plagiarism detection (for new code)
- metrics (for long standing projects)

Later on

- Support embargo/access control
Outline

3 Intrinsic PID

4 Our role in the publication workflow

5 The Metadata challenge

6 Collection strategies
It’s the real world!

reconcile metadata from different origins, handle conflicts, synthesise missing information, classify (automatically) the projects, etc.
Outline

3. Intrinsic PID

4. Our role in the publication workflow

5. The Metadata challenge

6. Collection strategies
All the source code
All the source code, strategies
Online, open source code: automation overview

Roberto Di Cosmo, Morane Gruenpeter