Intrinsic identifiers and the SWHID

A digital fingerprint identifying software source code

Roberto Di Cosmo

January 29th, 2020

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30 years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20 years* of Free and Open Source Software
- *10 years* building and directing structures for the common good

1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*
150 members 40 projects 200Me

2008 *Mancoosi project* www.mancoosi.org

2010 *IRILL* www.irill.org

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

Outline

Extrinsic identifiers

The Identifier has no *per se* relation with the designated Object

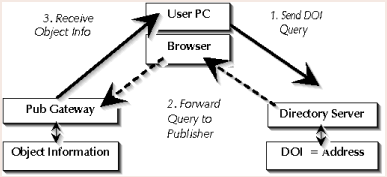
A *register* keeps the correspondence between the identifier and the object

pre-internet era passport number, social security number, ISBN, ISSN, etc.

internet era DOI, Handle, Ark, PURLs, RRID, etc.

A word about the *Persistent* adjective in Persistent Identifiers

- this technology *cannot guarantee* persistence by itself! Example:



- DOI resolution can change
- content at URL can change
- no way for the user to notice any of these changes from the outside

"persistence… is a function of *administrative* care"

RFC 3650 (Handle System Overview, 2003)

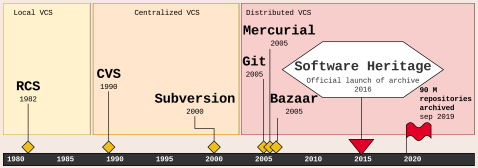
Intrinsic identifiers

The Identifier is derived from the designated Object

No register needed to keep the correspondence between the identifier and the object

pre-internet era musical notation, chemical notation (*NaCl* is table salt)

internet era cryptographic hashes for distributed software development, Bitcoin



- scientific breakthrough in the 1990's
- massive adoption in the 2010's
  - 150+M repositories (GitHub, BitBucket, GitLab)
  - 40.000.000 users

Persistence is built-in: nobody can change the designated object, and get away unnoticed!

Good news: now easily available for you via the Software Heritage Identifiers (SWHID)!

Outline

Software Heritage in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

Over 150 million repositories already, … and counting!

Addressing the ARDC key needs for (research) software source code…

Archive	ensure it is not lost	Describe	make it findable
Reference	identify the object	Cite	give credit to authors

… long term, non profit initiative with broad support

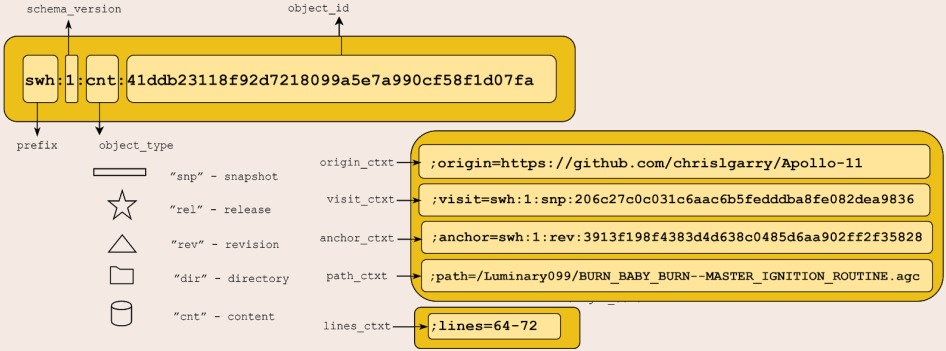
Academia and Government: Inria, UNESCO, CNRS, French National Open Science Fund, DANS, universities… Industry: Intel, Microsoft, GitHub, VMware, Societe Generale, …

SWHID: the source code fingerprint

Software Heritage Identifiers (SWHID)

link to full docs

20+B intrinsic, decentralised, cryptographically strong identifiers, SWHIDs



Emerging standard : Linux Foundation SPDX 2.2; IANA registered; WikiData P6138

Full fledged *source code references* for reproducibility

Examples: Apollo 11 AGC excerpt, Quake III rsqrt; Guidelines available, see ICMS 2020