PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

11–13 June, Prague, Czechia

CC BY

Identifying Billions of Source Code Artifacts: the SWHID in Publication Workflows

Miguel Colom
Centre Borelli
ENS Paris-Saclay

Morane Gruenpeter
Software Heritage
Inria

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiativ

ARDC
Australian Research Data Commons

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

The importance of referencing source code

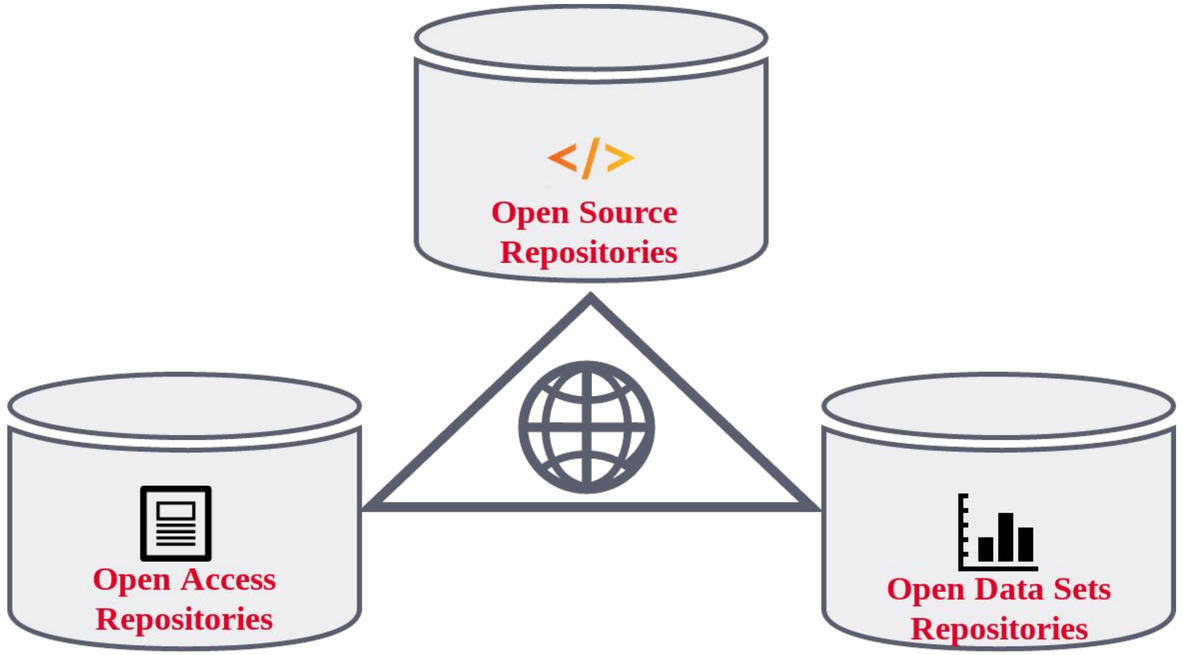
Open science: open access, open data, open source

The three **pillars** of **open science**:

- → **Open source repositories**
- Open datasets repositories
- Open access repositories

Software has **multiple facets:**

- a **tool**
- a research **outcome** or result
- the **object** of research



Three pillars of Open Science
Software Heritage CC-By 4.0 2019

The importance of archiving source code permanently

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

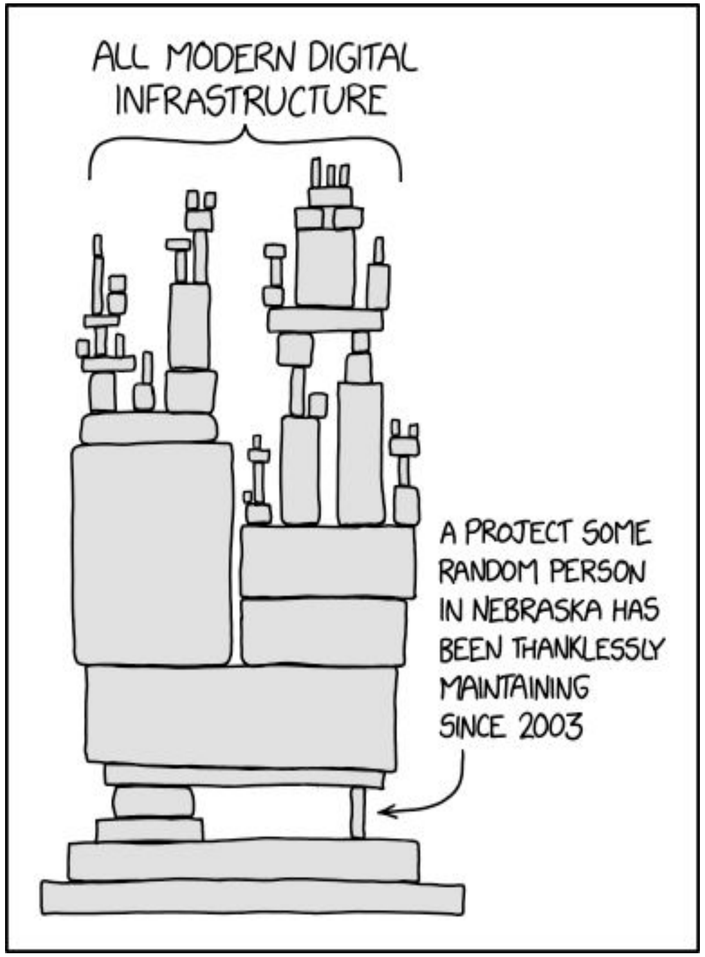
Software evolves over time

- projects **may last decades**
- the **development history** is key to its **understanding**

Contains part of the **knowledge** of **humanity**!

Complexity

- **millions** of **lines** of code
- large web of **dependencies**
  - easy to **break**, difficult to **maintain**
- sophisticated developer communities



https://xkcd.com/2347/

The Software Heritage initiative

One common infrastructure
to **collect**, **preserve** and **share** *all source code*

Snapshot: May 30th 2024

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

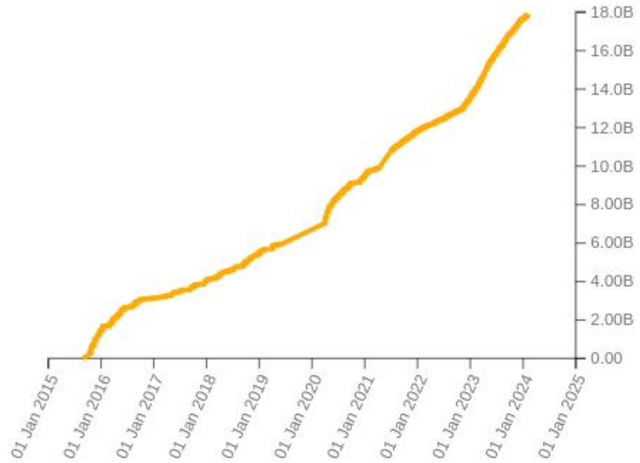
Are you familiar with Software Heritage?

Size

As of today the archive already contains and keeps safe for you the following amount of objects:

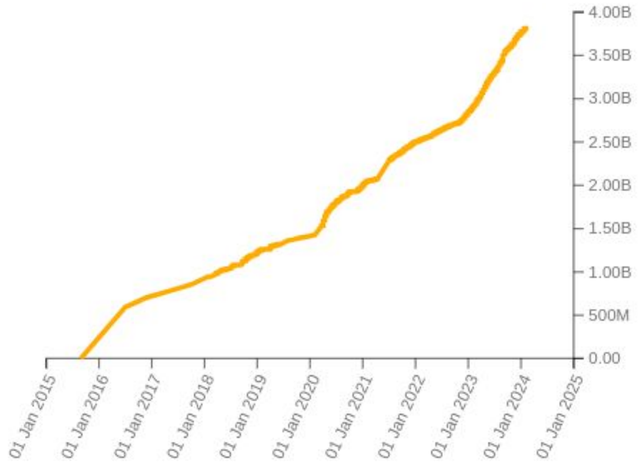
Source files

19,034,364,009



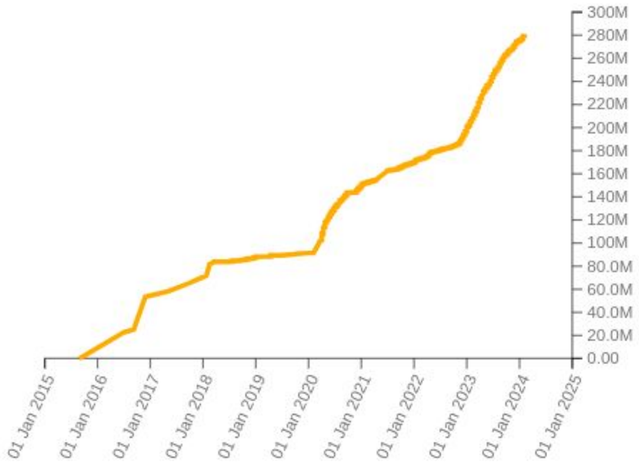
Commits

4,124,720,328



Projects

299,798,406



An international and non-profit initiative launched in 2016

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

built for the long term

Inria

Inria La Fondation

Sharing the vision

unesco

Donors, members, sponsors

cea

Microsoft HUAWEI SOCIETE GENERALE

cnrs

intel.

SORBONNE UNIVERSITÉ servicenow Hugging Face openinventionnetwork Université de Paris

AdaCore Google GitHub UNIVERSITÀ DI PISA

SCANOSS SCUOLA NORMALE SUPERIORE UNIVERSITÉ DE LORRAINE

Software GitLab ENGINEERING eclipse

CW2 fsfe nlnet FOUNDATION SIGSOFT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE INAPI

THE LINUX FOUNDATION

ADULLACT INFORMATICS EUROPE Computer History Museum software freedom conservancy

Software Freedom Law Center AdaCore gandi.net TDF

FREE SOFTWARE FOUNDATION

creative commons SIF open source initiative openinventionnetwork

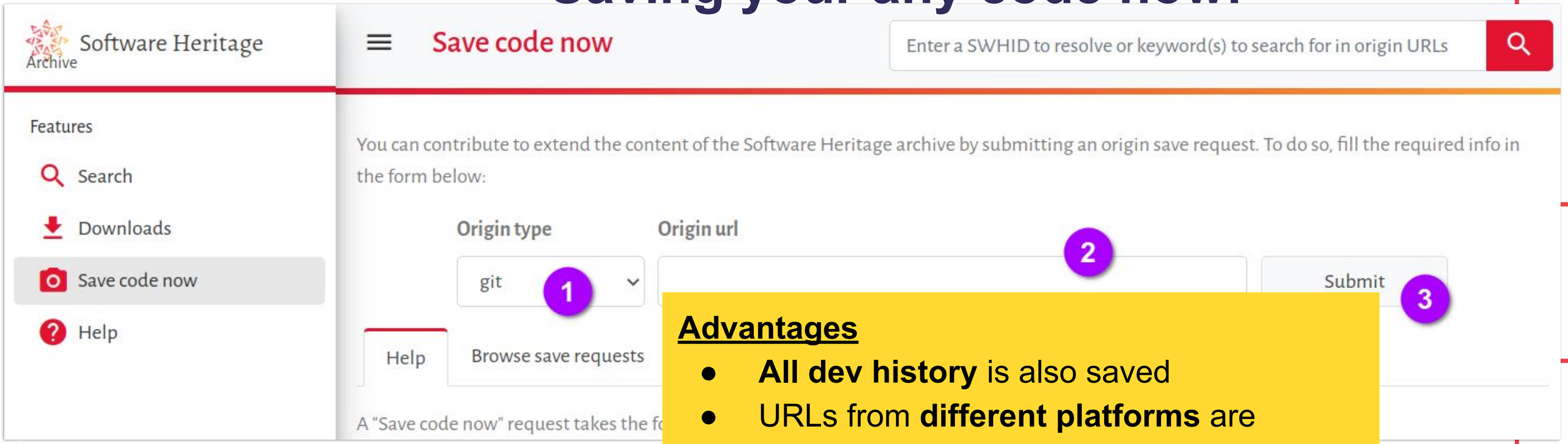
http://www.softwareheritage.org/support/testimonials

Save code feature

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

https://archive.softwareheritage.org/save/

Saving ~~your~~ any code now!



Advantages

- **All dev history** is also saved
- URLs from **different platforms** are accepted
- PID to **reference** specific pieces of code

Software Heritage

Meet the SWHID - The Software Hash Identifier

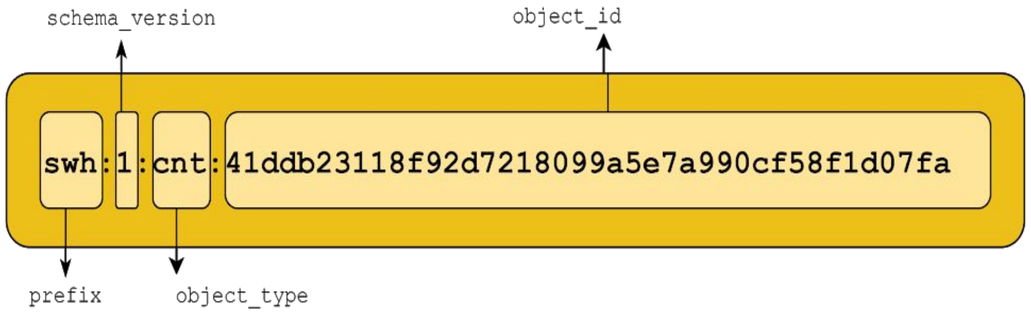
PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

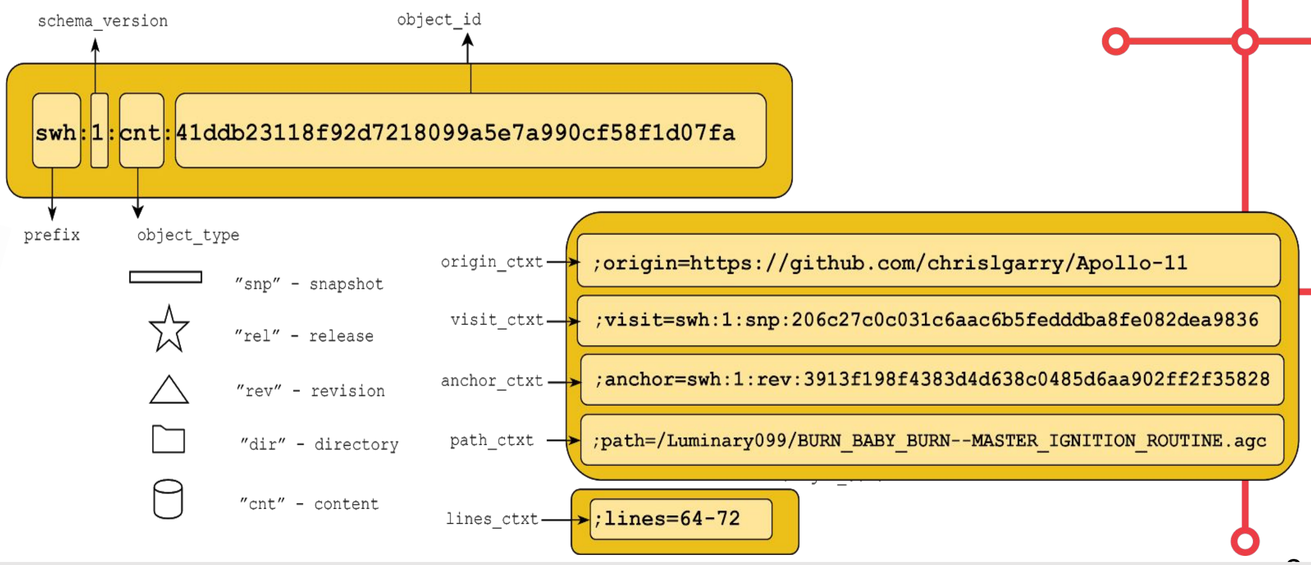
SWHIDs are **intrinsic identifiers** which are intimately **bound** to the **designated object**, they do **not need** a **register**, only agreement on a **standard**.

- **Intrinsic**: compute a unique **digital fingerprint**
- **decentralised**: do not need a registry, only agreement on a standard
- **cryptographically strong** identifiers

[Intrinsic vs. extrinsic blog post]

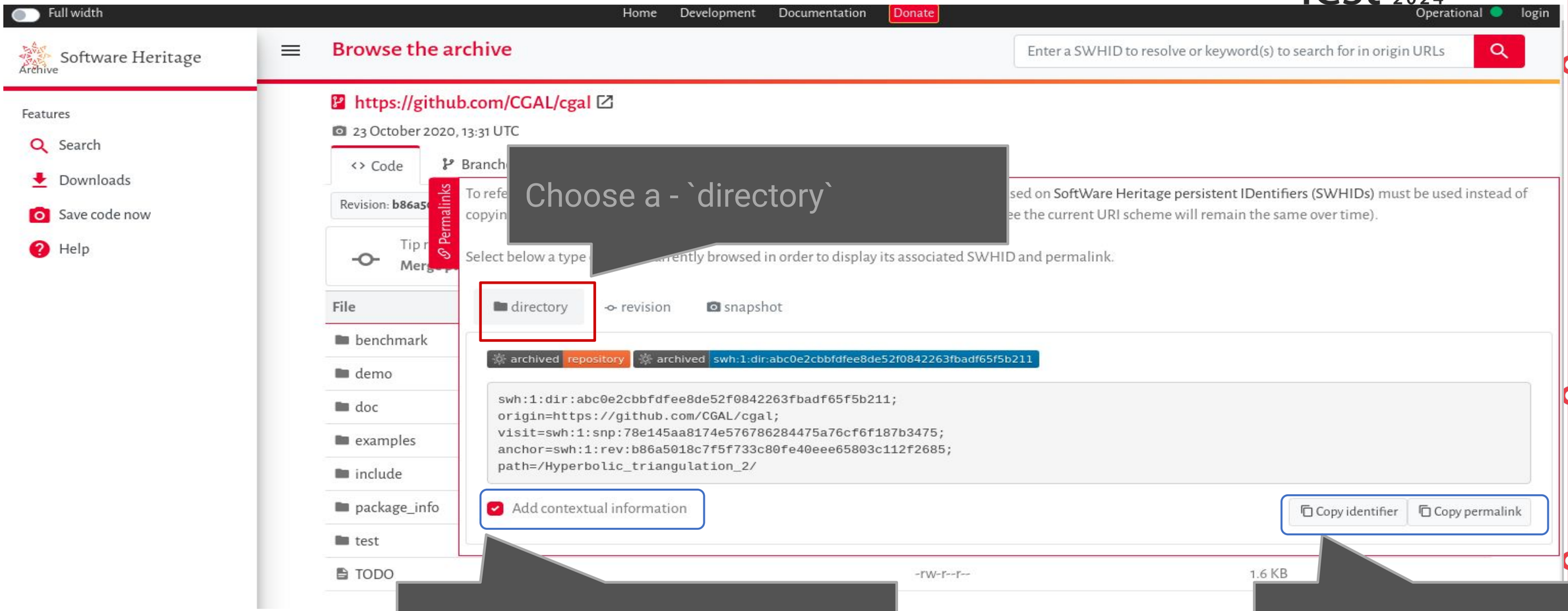
Go to API endpoint





Software Heritage

Choose a SWHID on Software Heritage



The need of a standard PID to reference and fully describe SW

PIDfest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Why do we *need* a **standard**?

- In **Open Science** there are plenty of different journals based on **different formats**
- To **define** an **agreed upon format**, we need the **standardisation** of mechanisms to **reference** software ⇒ **SWHID** is **recommended** in the **EOSC SRIA** and in the **EOSC SIRS** reports
- A **link** to Github or another **forge** is **not enough** to ensure **access** and **interoperability**

→ An open process has been put in place in order to produce a publicly available specification

The SWHID Working Group

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Working group and governance

- In **2023** the WG has been **established** https://swhid.org/
- **Governance**: https://www.swhid.org/governance/
- **Specification completed**: v1.0 (21/06/2023), v1.1 (6/11/2023)

Current status of SWHID

- 30+B SWHIDs in the archive
- Mention in Linux Foundation's SPDX 2.2
- IANA registered
- WikiData P6138

Go to API endpoint

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

Identification vs. location

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Identification of a book

Goal: attribute the same code to the same type of objects. *What is it?*

- One ISBN number per published book
- ISO 2108 Standard specification

ISBN.JPG, https://commons.wikimedia.org/w/index.php?title=File:ISBN.JPG&oldid=456239269 (last visited January 26, 2024).

Location of a copy of a book

Goal: find (a copy of) a book. *Where is it?*

- Many locations (locations can change!)
- Many approaches

https://commons.wikimedia.org/wiki/File:Art_Books_on_Library_Shelf.JPG

→We are mainly interested in **identification** and not that much with *location*

Extrinsic vs. Intrinsic identifiers

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Main difference:

- how the **relation** between **identifier** and designated **object** is **created** and **maintained**.

Persistence is a key **desired property**.

	Extrinsic	**Intrinsic**
Relation	Register	Convention - agreed protocol/naming
Pre-internet	Passport number, ISBN, SSN, etc.	Music / Chemistry notation
Internet era	DOI, Handle, ARK, etc.	Cryptographic hashes, e.g git, bitcoin, SWHID

[Intrinsic vs. extrinsic blog post]

Which part of the software do we want to *identify*? (granularity)

PIDfest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

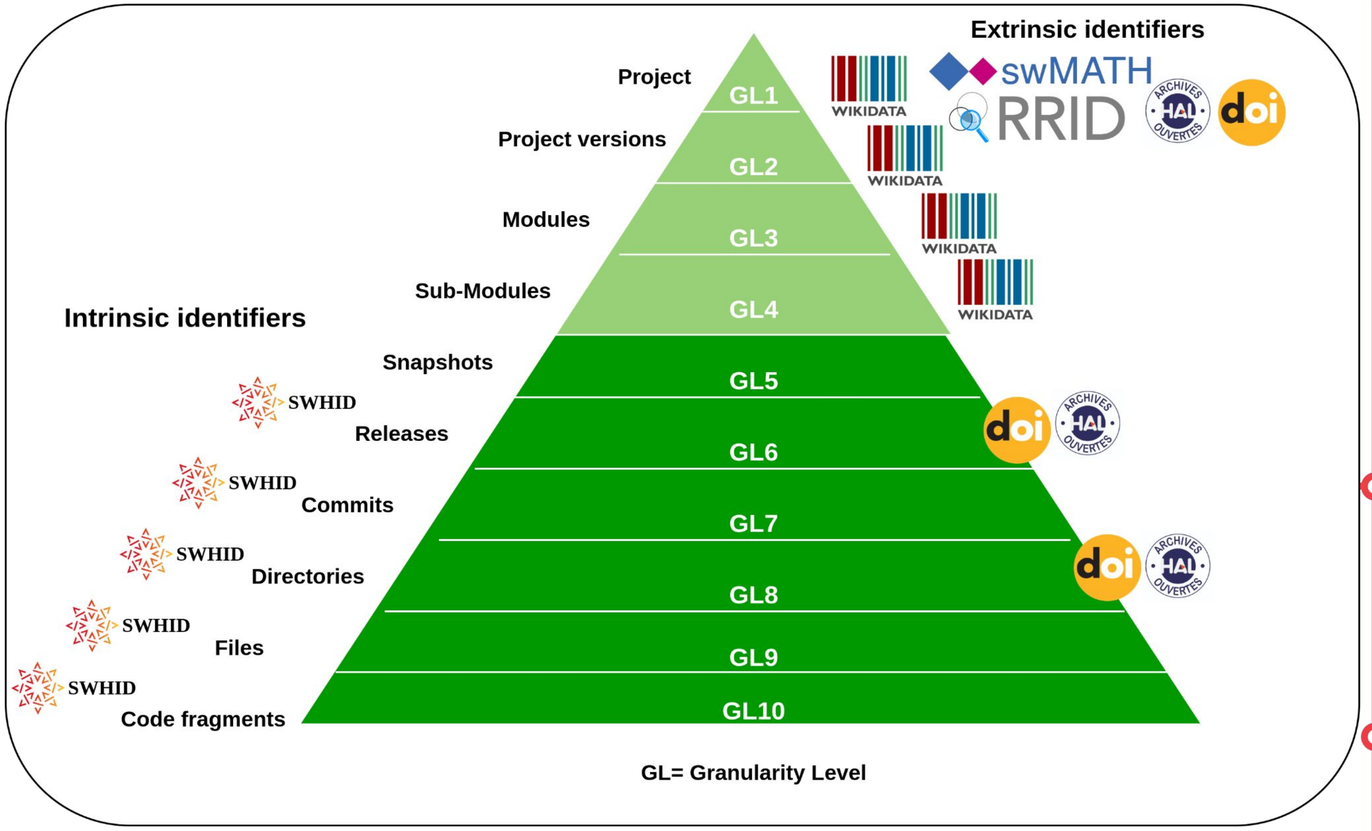
Software concept / project / collection

Description in registry, a homepage or any other form of metadata record

- ➔ Project versions (for example Python2 and Python3)
- ➔ Modules
- ➔ Sub-modules

Software artifact

- ➔ Executable (download link)
- ➔ Software source code
  - ◆ Dynamic artifact - current development code
  - ◆ Archived copy
    - Snapshot (all branches, all dev history)
    - Release / Package
    - Commit- a specific point in development history
    - Directory
    - File
    - Algorithm



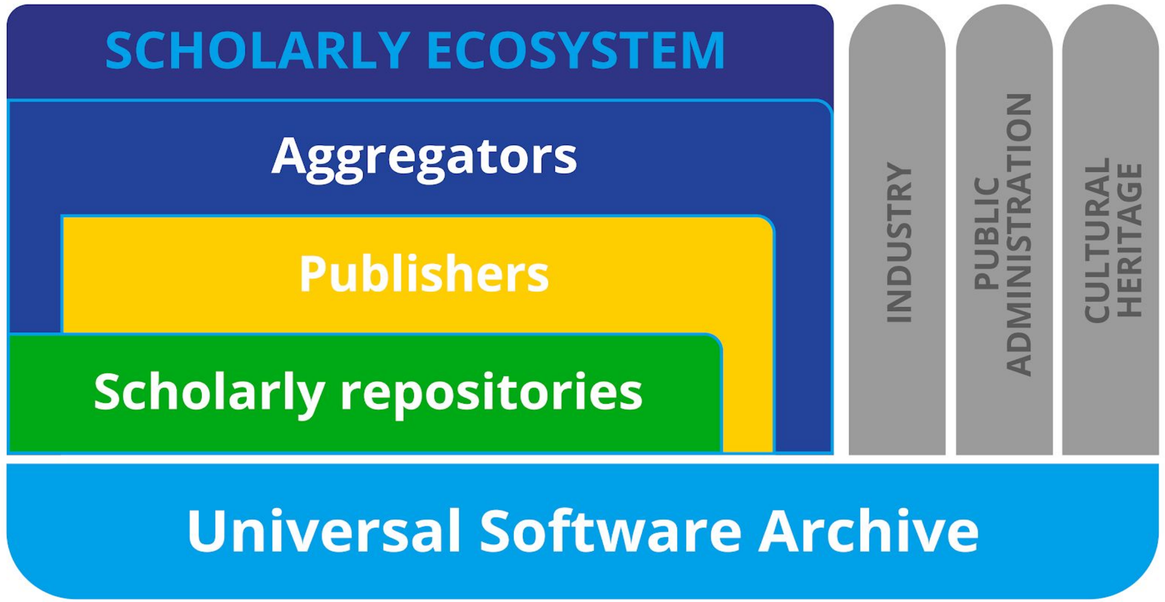
Research Data Alliance/FORCE11 Software Source Code Identification WG et al. (2020). Use cases and identifier schemes for persistent software source code identification (V1.1). *Research Data Alliance*. https://doi.org/10.15497/RDA00053

Use cases - connecting an ecosystem with SWHIDs

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

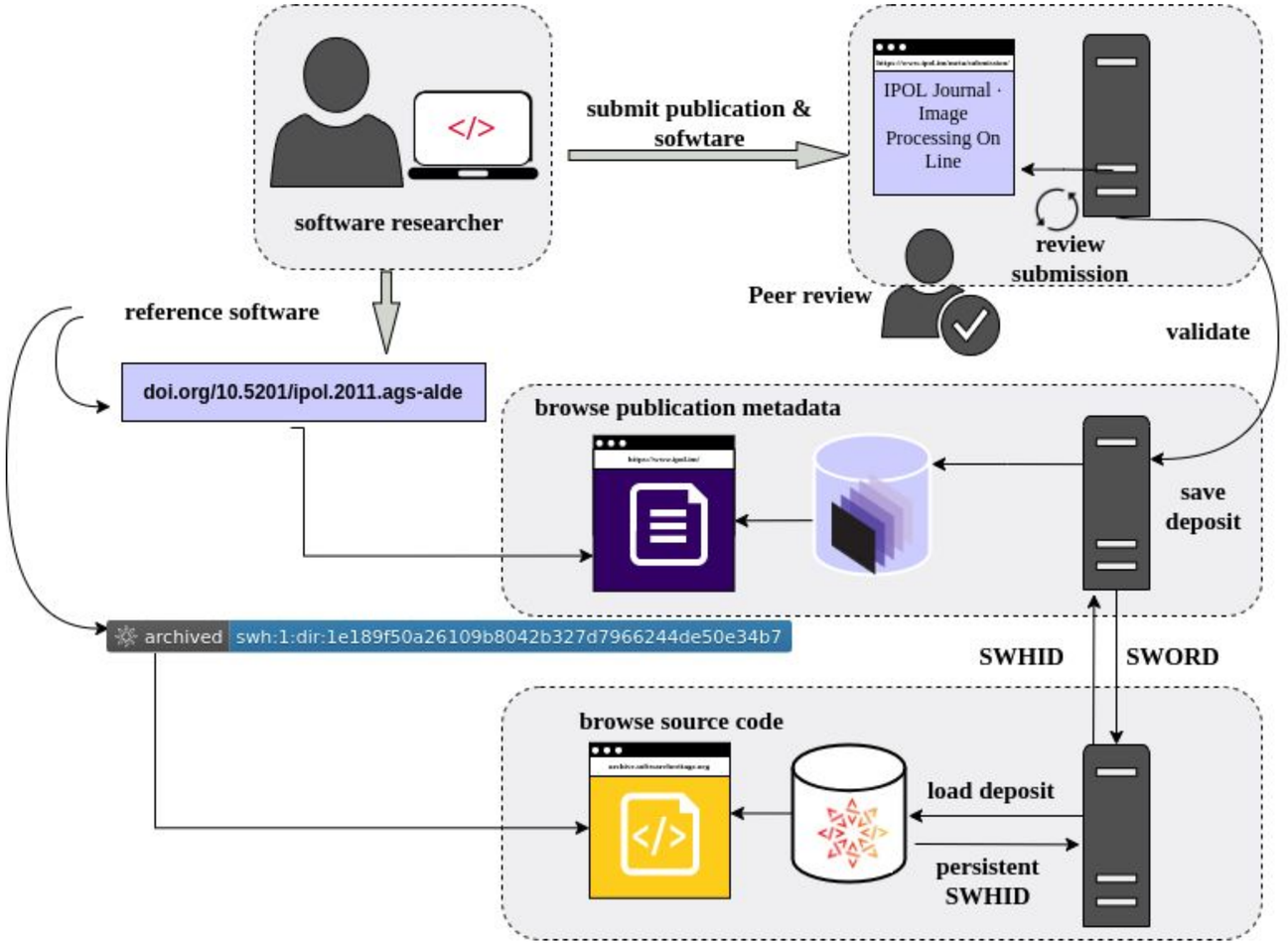
SIRS report: European Commission, Directorate-General for Research and Innovation, *Scholarly infrastructures for research software : report from the EOSC Executive Board Working Group (WG) Architecture Task Force (TF) SIRS*, Publications Office, 2020, **https://data.europa.eu/doi/10.2777/28598**

Video: EOSC Software Infrastructures for Research Software: J. B. Gonzalez Lopez (CERN)



eosc | FAIRCORE4EOSC Core Components Supporting a FAIR EOSC is implementing the SIRS report with the **SWHID**

IPOL publishing use case



IPOL journal publishing use case

- **Journal** Image Processing On Line (**IPOL**, https://www.ipol.im/)

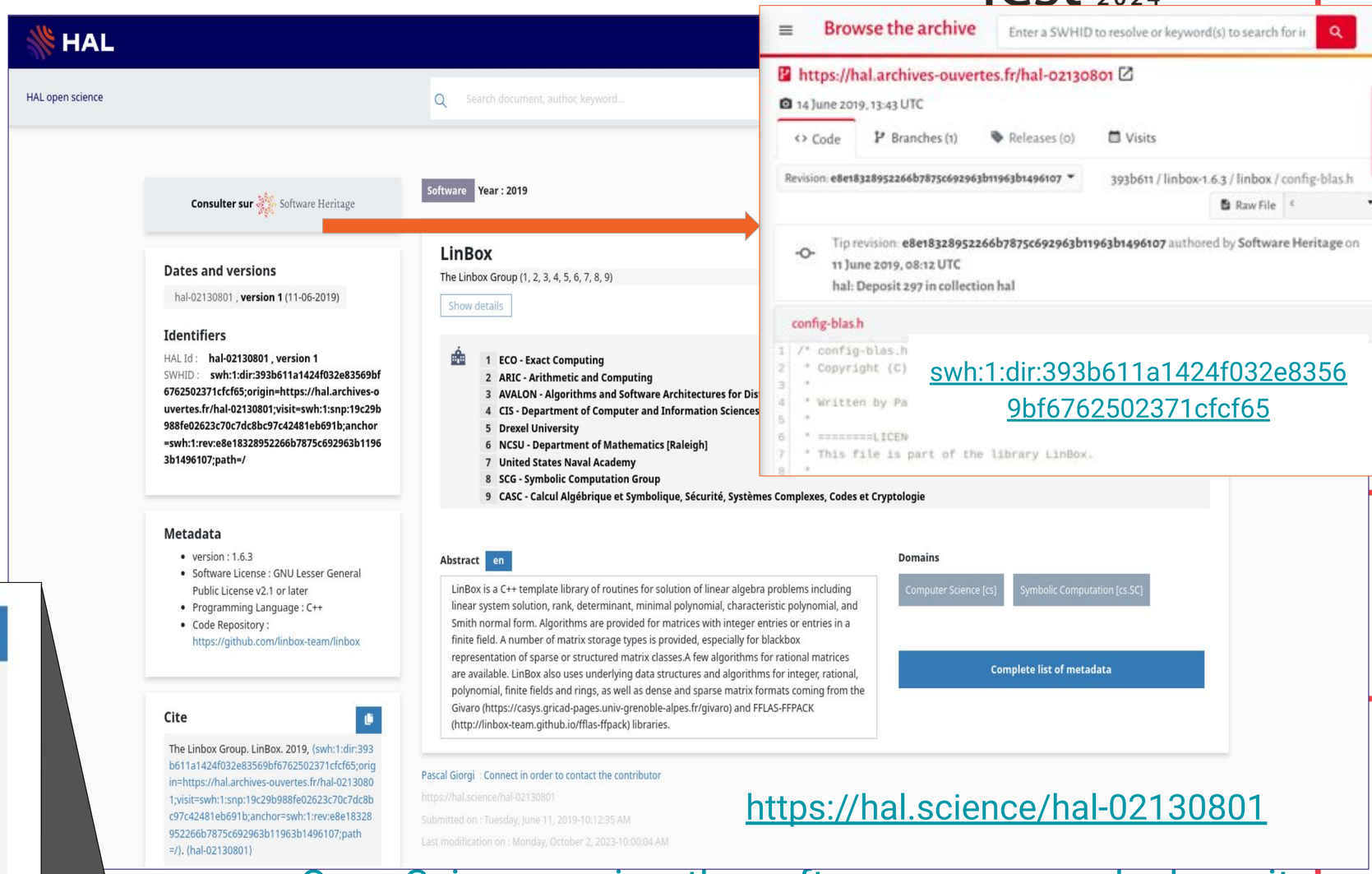
- **Research software** packages are **identified** with:
  - The article **DOI**: (https://doi.org/10.5201/ipol.2021.286)
  - The **software SWHID**: the publisher deposits the software in Software Heritage with the DOI as an origin (https://archive.softwareheritage.org/swh:1:dir:2cb75d8c95eb61d047d89428d0ec40a2286c0311;origin=https://doi.org/10.5201/ipol.2021.286;visit=swh:1:snp:23a5f7ee209b593e9b3e60ebe2bc42f1e6b76ff3;anchor=swh:1:rel:2de235c8fc3dd527cfaaba5cbf1d8144fee14f40)

- **Links from** the **paper** and **metadata DOI to**:
  - the **software deposit** and its **SWHID**,
  - the **live demo** of the **software** (in the *demo* tab)

HAL (CCSD) x SWH = combined solution for researchers

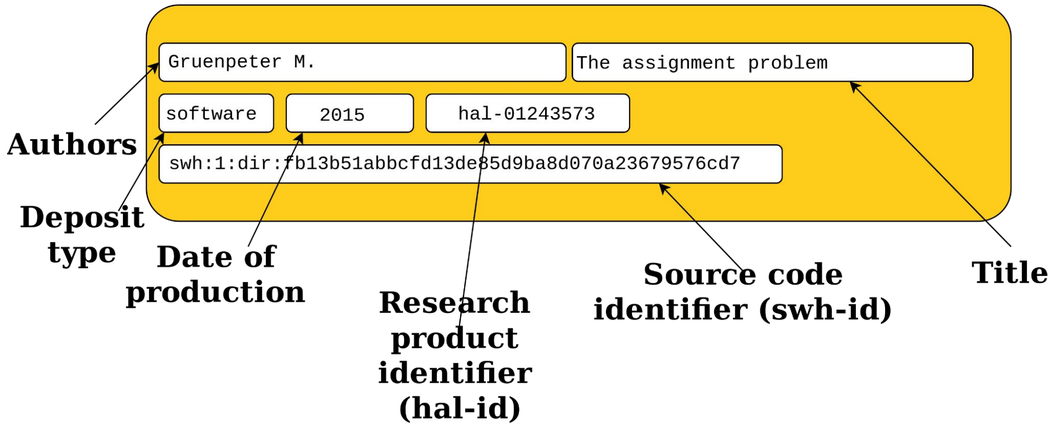
The **institutional** Research Software **deposit** in the French National Archive - **HAL**

- ★ Two options for researchers:
  - 🚀 Copy-pasting a PID
  - 📦 Adding files
- ★ Moderation of the metadata



[Open Science series: the software source code deposit]

For which use case: *reference* vs. *citation*



Archive & Index

- metadata record (extrinsic)
- artifact itself (intrinsic)

Credit & Attribution

- a metadata record
- all authors & contributors

Reuse & Reproducibility

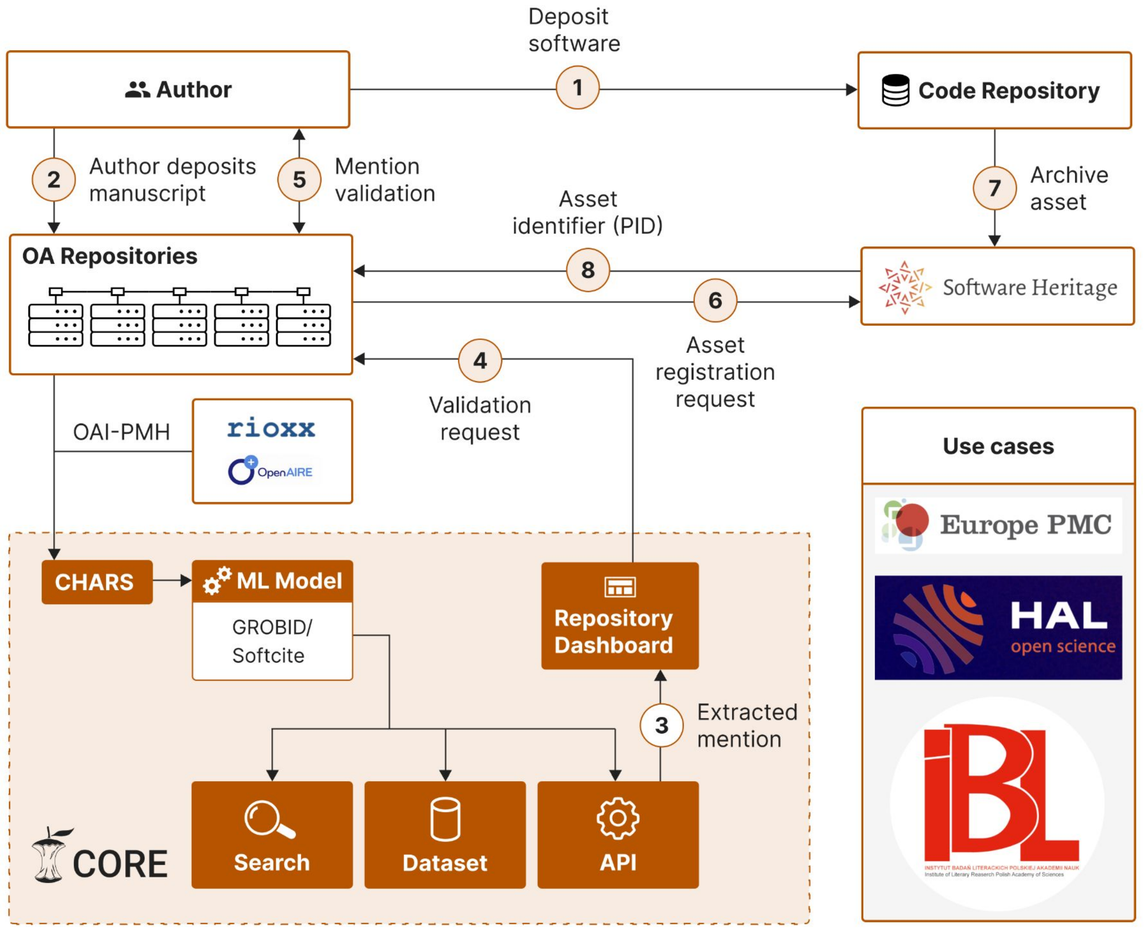
- a specific artifact
- appropriate metadata/docs

Building bridges between infrastructures

SoFAIR project, launched in January 2024

- **Goal**: establish a machine-assisted workflow embedded into widely used open scholarly infrastructures to **assist researchers** in
  - **identifying**,
  - **describing**,
  - **registering**,
  - **linking**,
  - and **archiving research software**.
- Funded by **CHIST-ERA**

- https://sofair.org/



Stay tuned…

Conclusions (1 / 2)

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- **SW** is **valuable** by itself
  - Not just a *supplementary material* of a scientific article, for example

- **SW** needs to be **preserved** for the long term
  - It's part of the knowledge heritage of humanity

- Preservation of SW and proper description ⇒ allow **reproducible research**

- **Good news**: **SWH** is **preserving all** publicly available **source codes** :-)

Conclusions (2 / 2)

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

- **SW** needs to be **properly referenced**
  - A DOI is not enough. <u>We need a **complete PID**</u>
  - We propose the **SWHID** to point to **permanently-stored source code** and properly **describe** it

  - Our goal is that the **SWHID** is a **general standard** to **all kind of software** for:
    - **Preservation**
    - **Proper citation**
    - **Interoperability**

PiD fest CONFERENCE ON PERSISTENT IDENTIFIERS 2024

Thank you for your attention

https://www.softwareheritage.org/

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

IPOL
IMAGE PROCESSING ON LINE

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

SD Times

Google begins shutdown of its code repository

Latest News | Published: March 12th, 2015 - Michael Pehel

After nine years, Google's open-source code repository, Google Code, started closing shop today by disabling new projects and announcing the permanent shut down of the service by January 26, 2016.

Google Code started as Google's answer to SourceForge, the predominant code repository back in 2006. The reliability of SourceForge was brought into question that year when SourceForge.net's database was hacked and user data was compromised. Problems continued the following summer in 2007 with a temporary service outage in August. The appeal of a repository backed by Google only increased as a result.

TNW News

Latest | Deep tech | Sustainability | Ecosystems | Data and security | Fintech and ecommerce | Future of work | More

This article was published on March 3, 2015

INSIDER

Code collaboration platform GitLab acquires rival Gitorious, will shut it down on June 1

March 3, 2015 - 4:11 pm

Gabriel Altay
@gabrielaltay

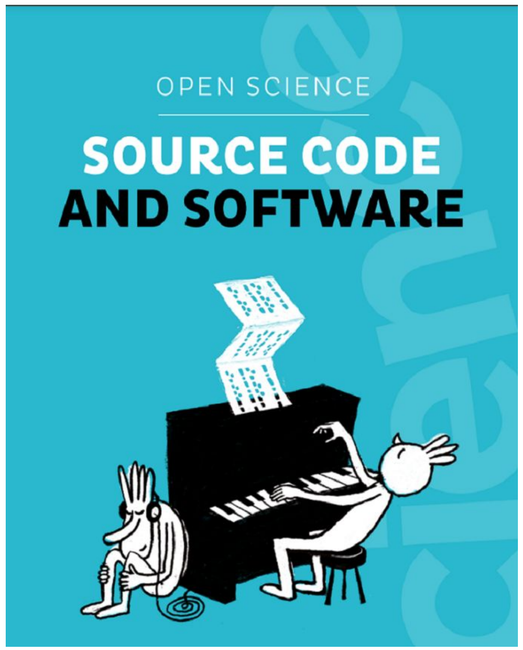
Just realized @Bitbucket disabled all mercurial repositories when the @asclnet informed me that a link associated with an old paper of mine was down. Thought all was lost, but someone archived all the repos! very classy move by @octobus_net and @SWHeritage.

Traduire le Tweet

1:48 AM · 31 août 2020 · Twitter Web App

Adoption in Academia: a few indicators

Policy



- Funding agencies recommendations ANR 2023 guidelines (p. 17)
- HAL+SWH in the Open Science software booklet from the French Ministry of Higher Education and Research

Citation & ACM

- ➔ November 2017: first meeting with ACM Digital Library
- ➔ May 2020 BibLaTeX released
- ➔ April 2022: inclusion in ACMART Class for typesetting publications of ACM

Replicability stamp

the Computer Graphic Replicability Stamp Initiative (GRSI)

Guix

The first free software distribution backed by a stable archive - making reproducible research papers possible