eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

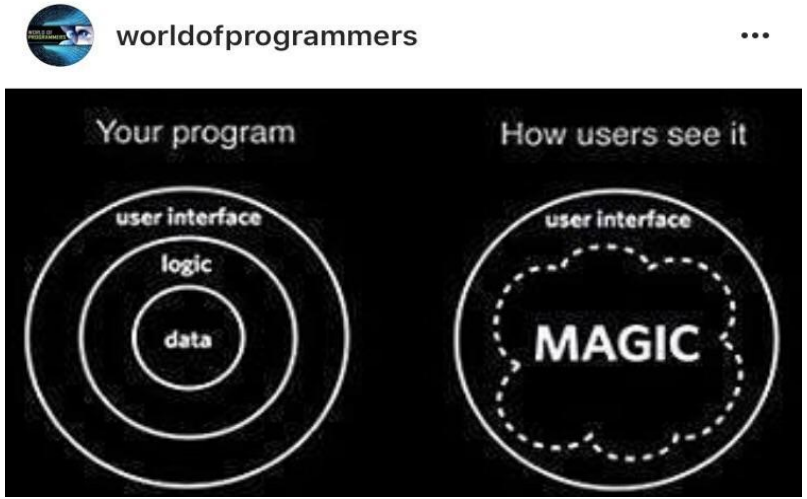
Archiving Source Code in Software Heritage

Morane Gruenpeter (Software Heritage, INRIA)

Work Package 6 Partners

Funded by
the European Union

Clarifying the magic - what is software?



https://www.reddit.com/r/ProgrammerHumor/comments/70fuamp/programming_is_magic/

Software as a concept

- **project** or entity
- the **community** around the project
- the software **idea** / algorithms / solutions

Not a digital artifact

Software artifacts

- Executables
- Source code

A very large collection of digital artifacts

Software Source Code is special

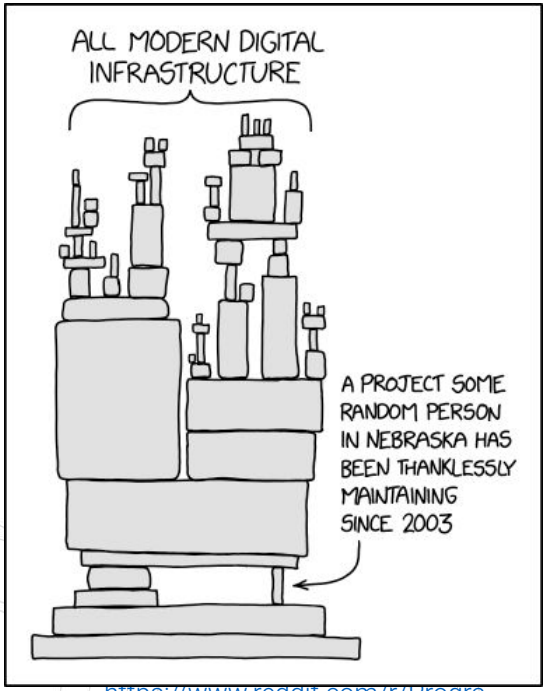
Software is not just another type of data

Software evolves over time

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

Complexity

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



https://www.reddit.com/r/ProgrammerHumor/comments/ic1zmc/dependency_xkcd/

Software is a pillar of Open Science

Research Software

➔ created

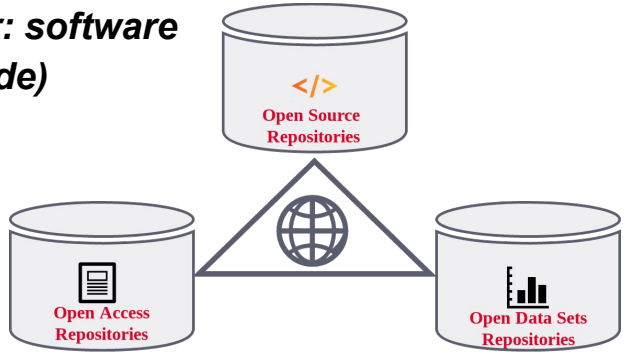
- **during the research process**
- **for a research purpose**

Software in research

➔ used for research

FAIR4RS output: Gruenpeter et al. Defining Research Software: a controversial discussion (Version 1). Zenodo. https://doi.org/10.5281/zenodo.5504016

A key pillar: software (source code)



← The links in the picture are **important**!

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

Software has multiple facets:

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

Use cases

Researchers

- **archive and reference** software used and created in articles
- **find** useful software
- **get credit** for developed software
- **verify/reproduce/improve** results

Laboratories/teams

- **track** software contributions
- **produce** reports
- **maintain** web page

Research Organization

know its **software assets** for:

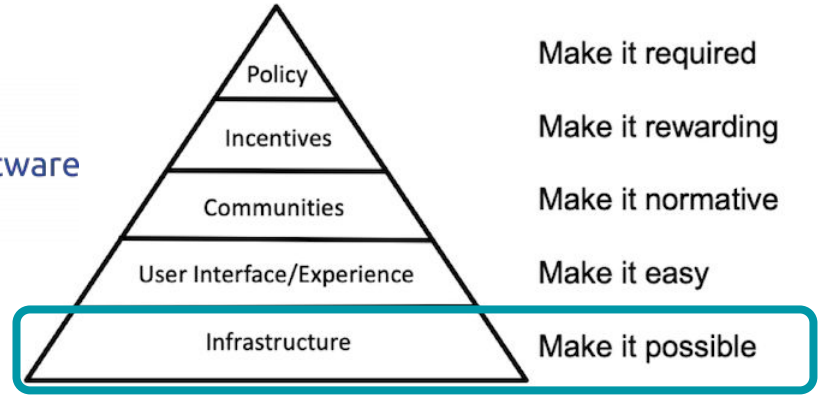
- technology transfer,
- impact metrics,
- Strategy

Curators

- **verify** and **curate** software metadata
- **provide** documentation on software curation
- **monitor** research teams' production

Culture change: Make software a first class output

RSAC
EOSC Research Software APIs & Connectors



Pyramid from Strategy for Culture Change: Brian Nosek (2019)
https://www.cos.io/blog/strategy-for-culture-change

swMATH
an information service for mathematical software

eLife

IPOL
IMAGE PROCESSING ON LINE

DataCite

INVENIO RDM

zenodo

EPIsciences
overlay journals

grnet

OpenAIRE

Software Heritage

CERN

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Data Archiving and Networked Services
DANS

What is at stake?

Archive

➔ make sure we can access to retrieve the software (reproducibility)

Reference

➔ make sure we can identify the software artifacts (reproducibility)

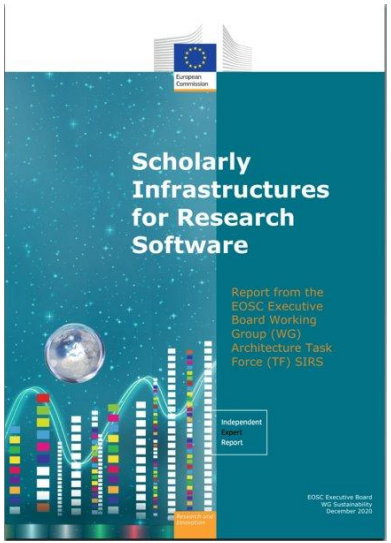
Cite (for credit)

➔ make it rewarding to create software by giving credit to authors (evaluation!)

Describe

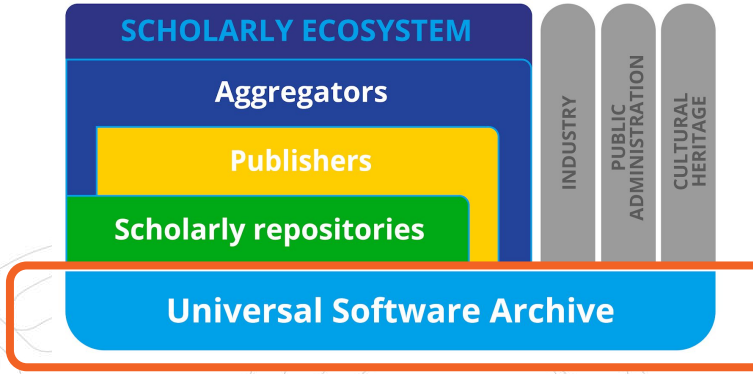
➔ make it easy to discover the software projects (visibility)

Archive, **Reference**, **Describe** and **Cite** Software



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**

- Focus on Software Source Code
- State of the Art
  - Best Practices & Open Problems
  - Cross Cutting Concerns
- The Road ahead
  - Requirements & Criteria
  - 13 Workflows /
    - Use Cases examples
- Recommendations
  - Standards & Tools
  - Policy recommendations
  - Long term perspectives



Software Heritage

Why archiving (research) software is important?

Source code is **fragile**!

Software Heritage

An international and non-profit infrastructure launched in **2016**

Inria — Inria La Fondation

Sharing the vision

unesco

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 — FREE SOFTWARE FOUNDATION — creative commons — SIF — open source initiative — openinventionnetwork

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

built for the long term

Donors, members, sponsors

Diamond sponsor

cea

Platinum sponsors

cnrs — SOCIETE GENERALE — intel — Microsoft — MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE — HUAWEI

Gold sponsors

openinventionnetwork — SORBONNE UNIVERSITÉ — Université de Paris

Silver sponsors

AdaCore — MINISTÈRE DES ARMÉES — DGA — Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE — GitHub — Google — UNIVERSITÀ DI PISA

Bronze sponsors

Red Hat — SCANOSS — SCUOLA NORMALE SUPERIORE

Software Heritage

The mission to **collect**, **preserve** and **share** source code

statistics on June 19th 2023 on the Software Heritage archive

Source files	Commits	Projects
15,769,273,132	3,276,619,808	240,995,441

DEPOSIT
to submit software source archives and its associated metadata

IPOL Journal

eLife

HAL science ouverte

CRAWLING

Bitbucket 2,036,916 origins	git 7,300 origins	21,620 origins
debian 129,507 origins	6,463 origins	GitHub 156,095,789 origins
GitLab 3,990,594 origins	Guix 12,466 origins	GNU 354 origins
heptapod 1,098 origins	launchpad 368,181 origins	Maven 93,710 origins
NixOS 12,466 origins	npm 1,799,296 origins	4,083 origins
Phabricator 184 origins	python 432,092 origins	SOURCEFORGE 308,965 origins

SAVE CODE NOW SERVICE

At any time, for free:

- For a project
- For a forge

And also …

🚒 **Rescue operations** 🔥

GITORIOUS

Google code

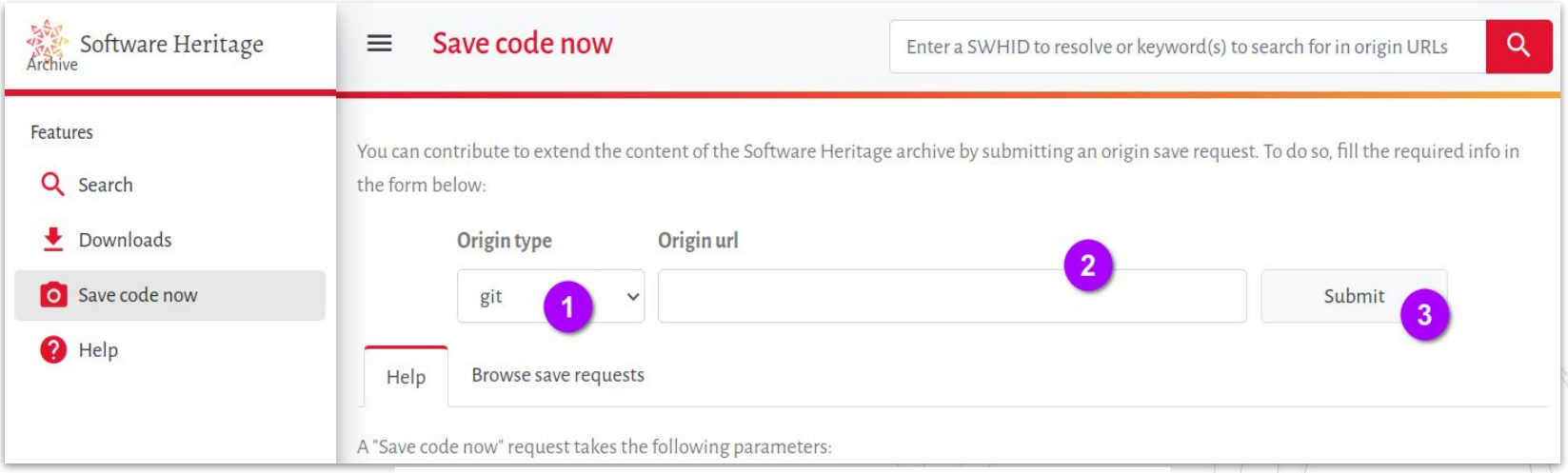
{P}

Bitbucket

Save any ~~your~~ code now!

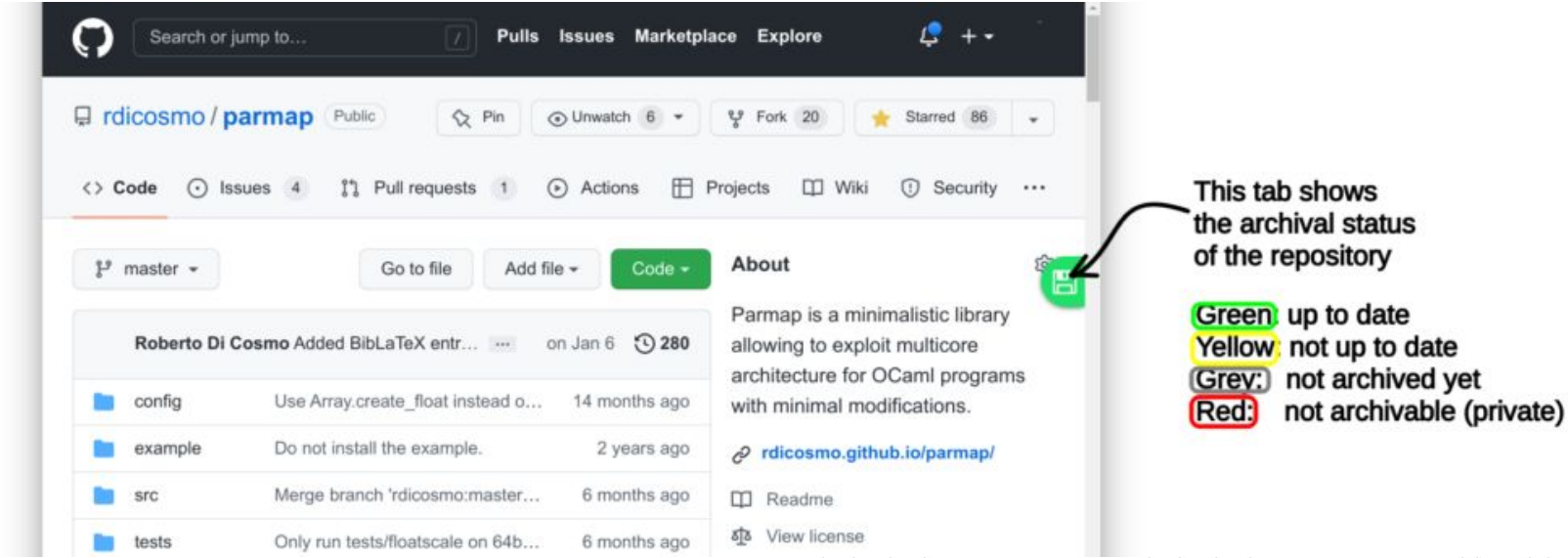
https://save.softwareheritage.org/

Save code now



- also via the webhook endpoint (doc) or browser extension

Browser extension - making it easy



Where is the metadata available ?

Extrinsic metadata lives alongside the software

Catalogs, registries & Aggregators

- ASCL
- SwMath
- OpenAire
- libraries.io
- Research Software Directory - escience center
- WikiData
- …

Scholarly repositories

- Zenodo (InvenioRDM)
- HAL
- DANS (DataVerse)
- …

Scholarly publishers

- IPOL
- eLife
- Dagstuhl
- Episciences
- …

Software development platforms (on platform page)

- GitHub
- Bitbucket
- SourceForge
- …

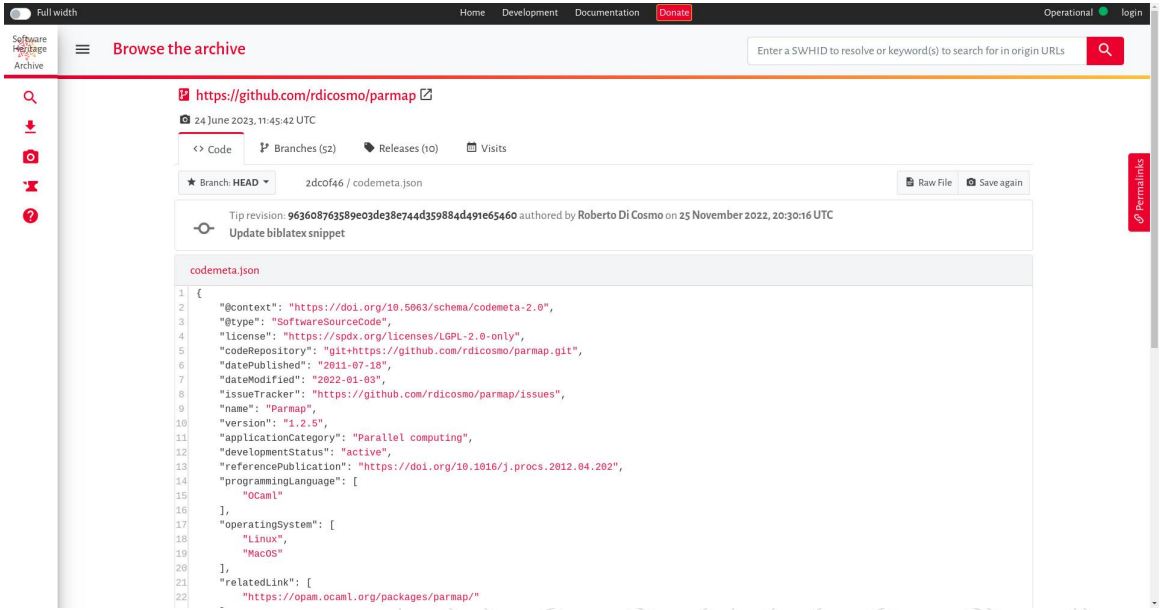
Package manager platform (not intrinsic file)

- NPM
- PyPI
- …

The case of intrinsic metadata

In the *software source code* itself

- README
- LICENSE
- AUTHORS
- **codemeta.json**
- package management
  - pom.xml
  - package.json
  - …
- CITATION.cff
- .About
- …

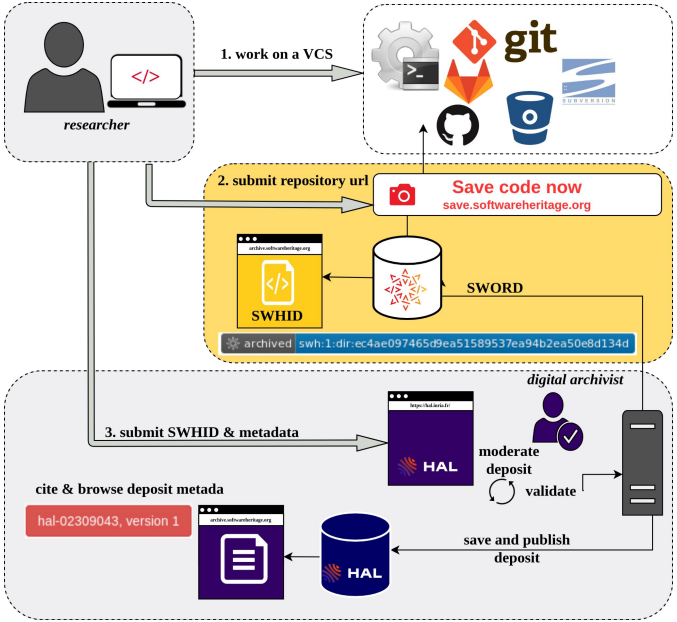


Human readable (e.g README)

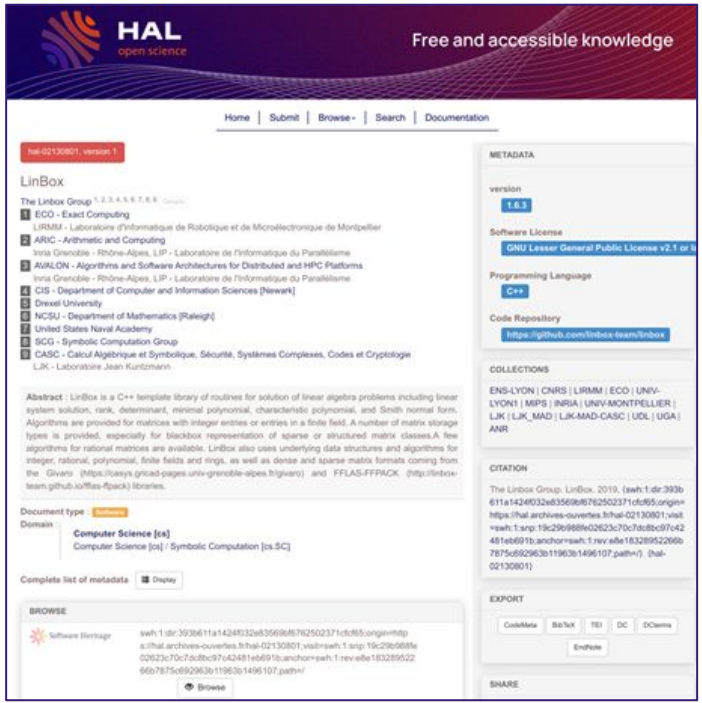
Machine actionable (e.g codemeta.json)

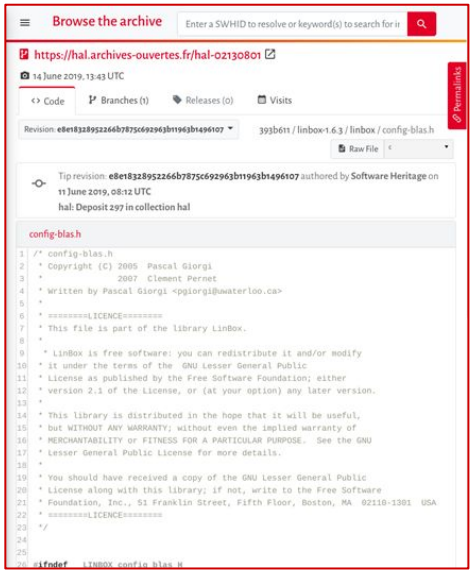
swh:1:cnt:0be13d6d1d0c0b377cd6e9e152d24c4e0b08d37d

HAL (CCSD) x SWH = Connecting the scholarly ecosystem



https://hal.science/hal-02130801





swh:1:dir:393b611a1424f032e83569bf6762502371cfcf65

In HAL, 2 options for researchers:

- 🚀 Copy-pasting a PID
- 📦 Adding files

Open Science series: the software source code deposit

RSAC
EOSC Research Software APIs & Connectors

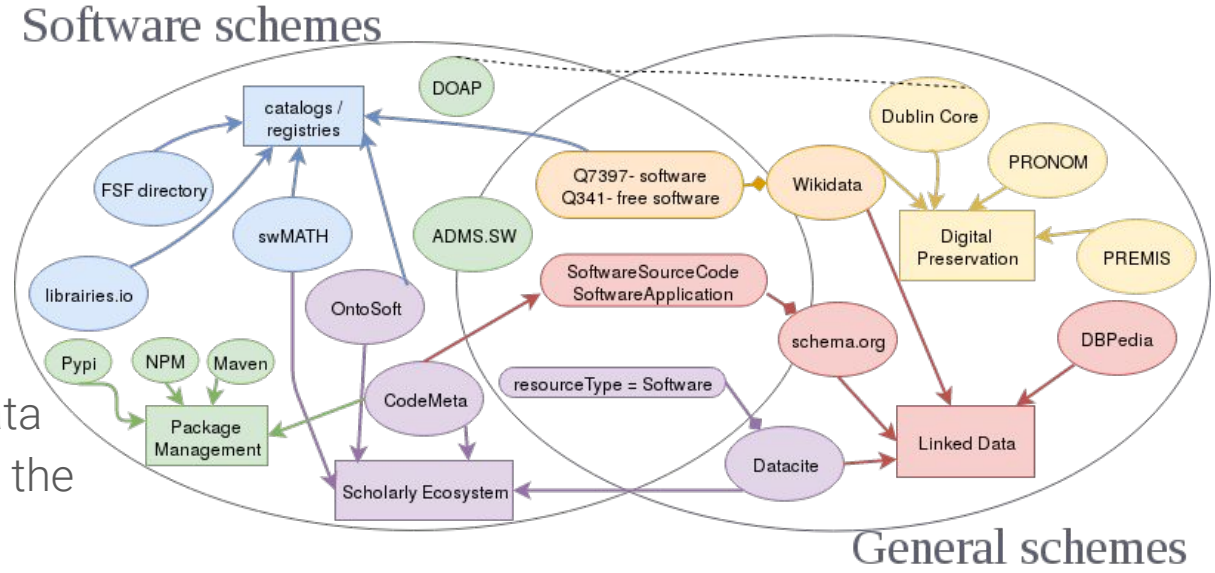
SWHM
EOSC Software Heritage Mirror

Work Package 6 Objectives

- ★ Develop tools and services for ***archival***, ***reference***, ***description*** and ***citation*** of research software artifacts by
  - implementing the key recommendations of the **EOSC SIRS report**
  - using the CodeMeta standard, and the SWHID identifier.
- ★ 2 new **EOSC-Core components** supporting a FAIR EOSC:
  - **EOSC RSAC**: Research Software APIs and Connectors to ensure the long-term preservation of research software in different disciplines
  - **EOSC SWHM**: Software Heritage Mirror to equip EOSC with a mirror of the Software Heritage universal source code archive
- ★ Design of **curation mechanisms** to support quality metadata, and generation of appropriate citations from the metadata.
- ★ Contribute to **standardisation** for software metadata and identifiers
- ★ Set up of regular archival in Software Heritage for core EOSC code hosting platforms

Why CodeMeta?

- A subset of schema.org
- An academic community discussing software metadata
- A crosswalk table - mapping the metadata landscape



Gruenpeter M. and Thornton K. (2018) Pathways for Discovery of Free Software (slide deck from LibrePlanet 2018). https://en.wikipedia.org/wiki/File:Pathways-discovery-free.pdf accessed on 6.11.2020.

Mappings => Translation

eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

Thank you for your attention!

- 📦 Archive your code!
  https://save.softwareheritage.org/
- 👓 Describe your code with metadata
  README, LICENSE, AUTHORS, codemeta.json

👉 Hot from the press: **Research Software MetaData**
The FAIR-IMPACT Deliverable - Guidelines for recommended metadata standard for research software within EOSC (V1.0 DRAFT NOT YET APPROVED BY EUROPEAN COMMISSION). Zenodo.
10.5281/zenodo.8097536
#RSMD_guidelines

Keep in touch

- morane@softwareheritage.org
- @moraneottilia @FAIRCORE4EOSC @SWHeritage

- Software Heritage newsletter
- FAIRCORE4EOSC newsletter

Funded by the European Union